frontiers | Research Topics

Applications of statistical methods and machine learning in the space sciences

Edited by
Bala Poduval, Karly Pitman, Olga Verkhoglyadova and Peter Wintoft

Published in
Frontiers in Astronomy and Space Sciences
Frontiers in Physics

ISSN 1664-8714
ISBN 978-2-83252-058-1
DOI 10.3389/978-2-83252-058-1

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Applications of statistical methods and machine learning in the space sciences

Topic editors

Bala Poduval — University of New Hampshire, United States
Karly Pitman — Space Science Institute, United States
Olga Verkhoglyadova — NASA Jet Propulsion Laboratory (JPL), United States
Peter Wintoft — Swedish Institute of Space Physics, Sweden

Citation

Poduval, B., Pitman, K., Verkhoglyadova, O., Wintoft, P., eds. (2023). *Applications of statistical methods and machine learning in the space sciences*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83252-058-1

Table of contents

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Editorial
PUBLISHED 16 March 2023
DOI 10.3389/fspas.2023.1163530

OPEN ACCESS

EDITED AND REVIEWED BY
Didier Fraix-Burnet,
UMR5274 Institut de Planétologie et d'Astrophysique de Grenoble (IPAG), France

*CORRESPONDENCE
Bala Poduval,
bala.poduval@unh.edu

SPECIALTY SECTION
This article was submitted to Astrostatistics, a section of the journal Frontiers in Astronomy and Space Sciences

RECEIVED 10 February 2023
ACCEPTED 16 February 2023
PUBLISHED 16 March 2023

CITATION
Poduval B, Pitman KM, Verkhoglyadova O, and Wintoft P (2023), Editorial: Applications of statistical methods and machine learning in the space sciences. *Front. Astron. Space Sci.* 10:1163530. doi: 0.3389/fspas.2023.1163530

Editorial: Applications of statistical methods and machine learning in the space sciences

Bala Poduval[1,2]*, Karly M. Pitman[2], Olga Verkhoglyadova[3] and Peter Wintoft[4]

[1]Space Science Center, Institute for the Study of Earth, Oceans, and Space, University of New Hampshire, Durham, NH, United States, [2]Space Science Institute, Boulder, CO, United States, [3]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, United States, [4]Swedish Institute of Space Physics, Lund, Sweden

KEYWORDS

machine learning, statistical methods, virtual conference, space science, astrophysics, space weather, heliophysics, artificial intelligence

Editorial on the Research Topic
Applications of statistical methods and machine learning in the space sciences

The fully virtual conference, *Applications of Statistical Methods and Machine Learning in the Space Sciences*, hosted by Space Science Institute's (SSI) Center for Data Science (CDS) and sponsored by the National Science Foundation (NSF), was held during 17–21 May 2021 (http://spacescience.org/workshops/mlconference2021.php). This event brought together experts in various disciplines of the space sciences (such as solar physics and aeronomy, planetary and exoplanetary sciences, geology, astrobiology, and astronomy) and industry to leverage the advancements in statistics, data science, methods of artificial intelligence (AI), and information theory with the aim of improving the analytic models and their predictive capabilities utilizing the enormous volume of data in these fields.

This multidisciplinary conference provided a vibrant forum for industry professionals, senior scientists, early career researchers, and students to present their latest results using a wide variety of techniques and methods in advanced statistics, to enhance their knowledge on the recent trends in AI and to participate in a platform for future collaborations. The conference covered a wide range of Research Topics, such as advanced statistical methods, deep learning and neural networks, time series analysis, Bayesian methods, feature identification and feature extraction, physics-based models combined with machine learning (ML) techniques and surrogate models, space weather prediction and other domain Research Topics where AI is applied, model validation and uncertainty quantification, turbulence and non-linear dynamics in space plasma, physics informed neural networks, information theory, and data reconstruction and data assimilation.

AI methods have already been applied to various problems in the field of solar-terrestrial physics since the 1990s (Newell et al., 1991; Lundstedt, 1992; Lundstedt, 1996; Wintoft and Lundstedt, 1997; Wing et al., 2005; Lundstedt, 2006). These included classifications of auroral particle precipitation, predictions of solar wind velocity, geomagnetic disturbances, and the planetary K-index K_p, used to characterize the

magnitude of geomagnetic storms (https://www.gfz-potsdam.de/en/section/geomagnetism/data-products-services/geomagnetic-kp-index). Information theory has proved useful in establishing linear and non-linear relationships and causalities in the studies of solar and space physics (Wing et al., 2016; Wing et al., 2018). Early attempts to apply ML techniques involve the forecasting of geomagnetic indices (e.g., Wu and Lundstedt, 1996; Wu and Lundstedt, 1997), the relativistic electrons at geosynchronous orbits (e.g., Stringer et al., 1996), and solar eruptions (Fozzard et al., 1988; Camporeale et al., 2019). A summary of current efforts on applying ML methods in the field of space sciences in comparison with those efforts in other fields of natural sciences and recommendations for ML in planetary science to funding agencies and the planetary community can be found in Azari et al. (2021). Figure 1 of Azari et al. (2021) illustrated that heliophysics and space physics had the highest percentage of published works discussing ML in 2020, followed by astrophysics and Earth science, and they concluded with recommendations for the next decade for supporting a data-rich future for planetary science.

The "International Workshop on Artificial Intelligence Applications in Solar-Terrestrial Physics," held in 1993, was one of the first of its kind which focused on "neural network applications of Multi-Layer-Error-Back-Propagation (MLBP) and Self-Organizing Map (SOM) neural nets and traditional expert systems and fuzzy expert systems" (Joselyn et al., 1993). Unlike this and other conferences on ML (Camporeale and SOC-ML-Helio, 2020), the SSI virtual conference had an emphasis on understanding the physics and dynamics of systems while seeking accurate solutions using ML methods ("black box" versus "interpretable" models). Furthermore, this virtual conference highlighted the interdisciplinary nature of ML applications in space sciences, the main theme of the conference. The research works presented revealed close collaborations among researchers in space science, statistics, computer science, and AI, showcasing how these experts can collaborate to soundly improve their models and predictions.

The virtual conference served as an initiative of SSI/CDS to bring together domain experts in space sciences and highly skilled corporate talents sharing a common interest in data science and ML. The CDS aims to inspire the scientific community to utilize key insights on emerging technologies, transforming this possibility into reality. SSI hosted 219 registered participants from more than 25 countries over Zoom for this event. Though participants were not asked to provide their demographic information, based on 103 of the conference registrants for whom the conference organizers could reasonably determine their backgrounds, we understand that there were 32 female participants, 43 from underrepresented minorities, and 45 early career (within 5 years after earning their Ph.D.s) scientists. We had 79 oral and 28 e-poster presentations in addition to interactive sessions demonstrating data processing and ML methods. The virtual conference featured 14 keynote speakers, 50% of whom were female scientists and 5 early career scientists. Links to these presentation slides and the recordings are available at the conference website (http://spacescience.org/workshops/mlconference2021.php).

The highlight of the conference was the lively discussion sessions. The virtual conference designated 45 min each day for live discussion sessions to discuss AI and ML trends in specific domains of space science and to encourage cross-disciplinary approaches to problems in different fields. Discussions were distributed among different Research Topics and centered around the applicability of Statistical Methods and ML in Astronomy, Aeronomy, Heliophysics, Magnetospheric Studies, Planetary Sciences and Exoplanets, and Turbulence and Non-linear Dynamics. Moreover, these sessions highlighted the importance and the impact of a few fundamental aspects in all the space science domains, such as the interpretability and explainability of ML models, reproducibility, and the need and availability of *AI-ready* data. These designated sessions addressed: the challenges of big data and small data sets; how to handle overfitting; uncertainties and gaps in the data sets and how they are incorporated into the models; supervised and unsupervised ML; and how to compare models. These discussions defined and emphasized the necessity of AI-ready data in all the disciplines of space sciences, and the participants shared information on the various data sets currently available and what are the steps to be taken to create better and more concise AI-ready data. We believe that these discussion sessions were particularly helpful for the students, early career researchers, and early ML practitioners who constituted a substantial fraction of the conference attendees, because these sessions covered links and access to a number of educational, software, and data resources. These discussions revealed the interdisciplinary nature of ML applications in the space sciences and how this virtual conference presented itself as a platform for connecting the various components of this fast emerging, dynamical trend of AI applications.

This topical collection compiles the works presented at the above virtual conference, along with new contributions from the broader scientific community in the form of original research articles, reviews/mini-reviews, brief reports and commentaries on the present scenario of AI applications in the space sciences, and scope of statistical and ML methods in the various fields of space sciences.

Active galactic nuclei (AGNs) are very bright, compact regions at the center of certain galaxies, the brightness of which arise from the accretion disks around supermassive black holes. Implementation of ML techniques in the redshift estimation of AGNs is becoming a common practice in astrophysics, but the data gaps in large-scale galactic surveys are often a hindrance to the smooth and reliable application of ML —a common problem in ML applications in general. Gibson et al. presents a technique for rectifying the missing data problem called Multivariate Imputation by Chained Equations (MICE) following Dainotti et al. (2021).

Outliers, observations that appear to differ considerably from others in the sample, are of great significance, especially in scientific data, for at least two reasons: 1) they may imply bad data, or a mistake in the experiment, code, or observation which, if detected, needs to be eliminated from the analysis, and 2) they may instead be scientifically interesting, indicating, for example, a random variation, and thereby, need to be detected and analyzed separately. In either case, detection of outliers is not an easy task, especially if the data set is enormously huge. Kerner et al. present a technique, Domain-agnostic Outlier Ranking Algorithms (DORA), for the automatic detection of outliers. DORA is a configurable pipeline for evaluation of outlier detection methods in different domains, supporting different data types such as image, raster, time

series, or feature vector and outlier detection methods including Isolation Forest, DEMUD, PCA, RX detector, Local RX, negative sampling, and probabilistic autoencoder. They experimented with various data sets and algorithms, and report their findings in Kerner et al.

In a Perspective article, Delzanno and Borovsky brings out the need for and the importance of a combined system science approach to global magnetospheric models and to spacecraft magnetospheric data. They opine that this approach provides statistical validation of global magnetospheric models without directly comparing with spacecraft data in addition to revealing the drawbacks of the model while providing the physics support to system analysis performed on the magnetospheric system. They emphasize that the question in this context is in fact, "Do simulations behave in the same manner as the magnetosphere does?", instead of the standard question, "How well do simulations reproduce spacecraft data?". The authors consider that this approach will provide statistical validation of global magnetospheric models without a direct comparison with spacecraft data and expose the deficiencies of the models, while providing physics support to the system analysis conducted on the magnetospheric system.

Blandin et al. compares the predictions of the magnitude of the north-south component of the geomagnetic field $|BN|$ using a multivariate Long Short Term Memory (LSTM) neural networks with the predictions of multivariate linear regression models. Both models use the same input, namely, a 15-year solar wind and heliospheric magnetic field from the NASA/GSFC's OMNI database accessible through https://omniweb.gsfc.nasa.gov.

For a direct comparison with the Geospace Environment Modeling (GEM) challenge of ground magnetic field perturbations for evaluating the predictive capabilities of empirical and first principle models and to select a model for operational purposes (Pulkkinen et al., 2013), Pinto et al. carried out a prediction of the horizontal component of the ground magnetic field rate of change (dB_H/dt) over six different ground magnetometer stations utilizing ML models based on feed-forward neural network, LSTM recurrent network, and CNN to forecast, and present the results.

Yeakel et al. utilized particle and magnetic field instrument data from the *Cassini* spacecraft mission to classify orbit segments as magnetosphere, magnetosheath, or solar wind. They trained and tested ML algorithms for classification, such as random forest, support vector machine, logistic regression, and LSTM, using a list of manually detected magnetopause and bow shock crossings by *Cassini* mission scientists, and present the results of this classification and a detailed error analysis.

Zhu et al. presents a new empirical reconstruction model of the three-dimensional magnetic field and the associated plasma currents, combining observations made by a constellation of satellites and a set of physics-based equations as physical constraints to build spatially smooth distributions. Here, the authors implement a stochastic optimization method to minimize the loss function characterizing the model-measurement differences and the model departures from linear or non-linear physical constraints. They further detail their discovery when applied to NASA's Magnetospheric Multiscale mission data.

Prediction of solar flares has been one of the greatest challenges in the domain of space weather, both operationally and from the perspective of scientific research. Pandey et al. present new heuristics in the training and deployment of the operational solar flare prediction method. They present two models, one based on full-disk and the other based on active regions (AR), for the prediction of flares belonging to classes $\geq M1.0$. They show that their model could predict a full-disk flare probability for the next 24 h and their proposed logistic regression, an ensemble model, improves on the full-disk and AR-based models (both base learners). They also discuss the model performances based on various metrics such as True Skill Statistic and Heidke Skill Score.

Bayesian inference is one of the ML applications that has been widely used in the field of space sciences in recent years and Arregui presents an example where it has successfully applied in coronal seismology and shows how the method can be applied to related areas of coronal loops, prominences, and other extended coronal regions. They point out that the Bayesian method becomes successful in these regions mainly because information about these regions is already incomplete and uncertain due to lack of direct access and most of the studies involve comparison of model predictions and remote observations, leading to the results being interpreted in terms of probabilities.

Narock et al. explores the utility of CNN in the prediction of the orientation of the embedded magnetic flux rope that are identified in the *in-situ* solar wind. They used magnetic field vectors from simulated flux rope data, that includes a number of possibilities in the spacecraft trajectories and flux rope orientations, to train the CNN. They explore different neural network topologies, the various factors that influence the prediction accuracy, and compares with an Interplanetary Coronal Mass Ejection (ICME) observed by Wind spacecraft.

The mini review by Telloni highlights the author's previous works based on statistical analyses of interplanetary and geomagnetic data in the context of space weather prediction. The first two of the three papers reviewed here were on what triggers the space weather effects, such as the geomagnetic storm; the first paper focuses on the detection, characterization, and geo-effectiveness of ICMEs and the second one considers other solar events, during the same period of study as in the first paper, and focuses on the connection between solar wind energy and geomagnetic activity. The third paper addresses the recovery phase and explores the reasons for the slow restoration of equilibrium conditions of the Earth's magnetosphere.

Verkhoglyadova et al. discuss their perspectives on implementing a mixture method approach and a computer vision approach in quantitatively addressing the anomalies and high density regions (HDRs) that are present in a global ionospheric map, and how the number of the HDRs and their intensities depend on solar and geomagnetic activities. The article finds that they are complementary and helpful in understanding the properties of the global ionosphere and emphasize the importance of a consistent definition of large-scale ionospheric structures.

One of the mechanisms of radiation belt loss is the electron precipitation (EP) through two known processes, wave-particle interactions (relativistic electron precipitation, REP) or current sheet scattering (CSS), and which of these processes dominates is still not fully understood (e.g., Schulz and Lanzerotti, 1974). It is well-known that EP drives atmospheric effects that are related to space weather adversities. Capannolo et al. developed a model based on LSTM to identify relativistic precipitation events and, their associated driver

(REPs or CSSs) and classify them as REPs or CSSs. They find that this large data set of REP and CSS events is useful in obtaining the location and properties of the precipitation driven by these two processes at all L-shells and magnetic local time sectors, thereby improving the radiation belt models.

Solar granulation, the dark and bright granular structure visible on the photosphere, depicts the overturning convective transport of magnetized plasma and energy in the region right below the photosphere (see Stix, 2002, for details). There exist specific and systematic morphological patterns including the exploding granules and bright points that have been extensively studied. U-net, a CNN used for biomedical image segmentation, has been found to be promising in the classification of solar granulation structures as shown by Díaz Castillo making use of the continuum intensity maps of the IMaX instrument on board Sunrise I and corresponding segmented maps as a training set. The authors find that U-net architecture is quite promising in identifying cellular patterns in solar granulation images with an average accuracy above 80%.

Song et al. presents their automatic identification algorithm to detect the magnetopause crossing events in THEMIS data from 2007 to 2021 in a study of overshoot structure in the magnetospheric magnetic field. They found that about half of the identified magnetopause crossing events near the subsolar region "appear [to have] an overshoot structure." The rate of change of a magnetospheric magnetic field near the magnetopause bears a linear relation to the magnetopause velocity, implying that the cause of the overshoot structure can be considered as the magnetospheric magnetic field redistribution caused by the rapid motion of the magnetopause.

El Mir and Perinpanayagam reviews the current certification of landing gear available for use in the aerospace industry. The authors discuss the role of ML techniques in structural health monitoring and points out that the non-deterministic nature of deep learning algorithms could be a hurdle for certification and verification in the industry. For implementing ML methods successfully, the safe-life fatigue assessment needs to be certified so that the remaining useful life may be accurately predicted and trusted. They further discuss the risk management and explainability for different end user categories involved in the certification process.

In addition to this topical collection that reveals the interdisciplinary nature of the applications of AI and statistical methods, as the virtual conference aimed at, the most significant outcome is the multi-authored white paper on the AI-readiness (Poduval et al., 2022) of the numerous space science data for AI/ML applications that was submitted to The National Academies of Science, Engineering, and Medicine's Decadal Survey for Solar and Space Physics (Heliophysics) 2024–2033. There is a strong urgency in the space sciences to make all existing data AI-ready within a decade, which is ambitious, not only because of the timescale and enormity of the data sets involved, but also because AI-readiness lacks a concrete definition within and across all fields in space science. Poduval et al. (2022) provides a definition of AI-readiness that conveys the widely accepted norms and concepts in the space sciences community and recommend mitigation strategies such as unambiguously defining AI-readiness; prioritizing certain data sets, their storage and accessibility; and identifying the agencies, private sector partners, or funded individuals who will be responsible. We hope this topical collection will help the scientific community to further advance the initiative to get the space science data AI-ready in a timely fashion.

Author contributions

BP was in lead in organizing the virtual conference with KP, OV, and PW as part of the scientific organizing committee and co-editors of the topical collection. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

The material presented here is based upon work supported by the National Science Foundation under Award No. AGS—2114219. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Acknowledgments

OV acknowledges that portions of work were performed at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with NASA.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Azari, A. R., Biersteker, J. B., Dewey, R. M., Doran, G., Forsberg, E. J., Harris, C. D. K., et al. (2021). Integrating machine learning for planetary science: Perspectives for the next decade. *Submitted to the NRC Planetary and Astrobiology Decadal Survey* https://baas.aas.org/pub/2021n4i128/release/1?readingCollection=7272e5bb.

Camporeale, E., Chu, X., Agapitov, O. V., and Bortnik, J. (2019). On the generation of probabilistic forecasts from deterministic models. *Space weather.* 17, 455–475. doi:10.1029/2018sw002026

Camporeale, E. and Soc-ML-Helio, (2020). ML-helio: An emerging community at the intersection between heliophysics and machine learning. *J. Geophys. Res.* 125, e2019JA027502. doi:10.1029/2019JA027502

Dainotti, M. G., Bogdan, M., Narendra, A., Gibson, S. J., Miasojedow, B., Liodakis, I., et al. (2021). Predicting the redshift of γ-ray-loud AGNs using supervised machine learning. *Astrophys J* 920, 118. doi:10.3847/1538-4357/ac1748

Fozzard, R., Bradshaw, G., and Ceci, L. (1988). "A connectionist expert system that actually works," in Advances in neural information processing systems, Cambridge, 1 January 1988.

Joselyn J., Lundstedt H., and Trolinger J. (Editors) (1993). Artificial intelligence applications in solar terrestrial physics, Lund, Sweden, September 22–24, 1993.

Lundstedt, H. (1992). Neural networks and predictions of solar-terrestrial effects. *Planet. Space Sci.* 40, 457–464. doi:10.1016/0032-0633(92)90164-j

Lundstedt, H. (2006). Solar activity modelled and forecasted: A new approach. *Adv. Space Res.* 38, 862–867. doi:10.1016/j.asr.2006.03.041

Lundstedt, H. (1996). Solar origin of geomagnetic storms and predictions. *JATP* 58, 821–830. doi:10.1016/0021-9169(95)00105-0

Newell, P. T., Wing, S., Meng, C. I., and Sigillito, V. (1991). The auroral oval position, structure, and intensity of precipitation from 1984 onward: An automated on-line data base. *J. Geophys. Res.* 96, 5877–5882. doi:10.1029/90ja02450

Poduval, B., McPherron, R. L., Walker, R., Himes, M. D., Pitman, K. M., Azari, A. R., et al. (2022). AI-ready data in solar physics and space science: Concerns, mitigation and recommendations. *White Paper Submitted to the Decadal Survey for Solar and Space Physics (Heliophysics) 2024-2033* http://surveygizmoresponseuploads. s3. amazonaws. com/fileuploads/623127/6920789/107-1870187ec154eee48664bed68513f0cb_PoduvalBala.pdf.

Pulkkinen, A., Rastätter, L., Kuznetsova, M., Singer, H., Balch, C., Weimer, D., et al. (2013). Community-wide validation of geospace model ground magnetic field perturbation predictions to support model transition to operations. *Space weather.* 11, 369–385. doi:10.1002/swe.20056

Schulz, M., and Lanzerotti, L. J. (1974). "Particle diffusion in the radiation belts," in *Physics and chemistry in space* (Berlin: Springer).

Stix, M. (2002). *The sun*. Germany: Springer.

Stringer, G., Heuten, I., Salazar, C., and Stokes, B. (1996). Artificial neural network (ann) forecasting of energetic electrons at geosynchronous orbit. *Radiat. Belts Models Stand.* 97, 291–295.

Wing, S., Johnson, J. R., Camporeale, E., and Reeves, G. D. (2016). Information theoretical approach to discovering solar wind drivers of the outer radiation belt. *J. Geophys. Res.* 121, 9378–9399. doi:10.1002/2016JA022711

Wing, S., Johnson, J. R., Jen, J., Meng, C.-I., Sibeck, D. G., Bechtold, K., et al. (2005). Kp forecast models. *J. Geophys. Res.* 110, A04203. doi:10.1029/2004JA010500

Wing, S., Johnson, J., and Vourlidas, A. (2018). Information theoretic approach to discovering causalities in the solar cycle. *Astrophys. J.* 854, 85. doi:10.3847/1538-4357/aaa8e7

Wintoft, P., and Lundstedt, H. (1997). Prediction of daily average solar wind velocity from solar magnetic field observations using hybrid intelligent systems. *Phys. Chem. Earth* 22, 617–622. doi:10.1016/s0079-1946(97)00186-9

Wu, J.-G., and Lundstedt, H. (1997). Geomagnetic storm predictions from solar wind data with the use of dynamic neural networks. *J. Geophys. Res.* 102 (14), 14255–14268. doi:10.1029/97ja00975

Wu, J.-G., and Lundstedt, H. (1996). Prediction of geomagnetic storms from solar wind data using elman recurrent neural networks. *Geophys. Res. Let.* 23, 319–322. doi:10.1029/96GL00259

frontiers in Astronomy and Space Sciences

PERSPECTIVE
published: 18 February 2022
doi: 10.3389/fspas.2022.808629

The Need for a System Science Approach to Global Magnetospheric Models

Gian Luca Delzanno[1]* *and Joseph E. Borovsky*[2]

[1]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM, United States, [2]Space Science Institute, Boulder, CO, United States

This perspective advocates for the need of a combined system science approach to global magnetospheric models and to spacecraft magnetospheric data to answer the question "Do simulations behave in the same manner as the magnetosphere does?" (instead of the standard validation question "How well do simulations reproduce spacecraft data?"). This approach will 1) validate global magnetospheric models statistically, without the need for a direct comparison against spacecraft data, 2) expose the deficiencies of the models, and 3) provide physics support to the system analysis performed on the magnetospheric system.

Keywords: planetary magnetospheres, global magnetospheric models, system science, information theory, model validation

OPEN ACCESS

Edited by:
Olga Verkhoglyadova, NASA Jet Propulsion Laboratory (JPL), United States

Reviewed by:
Adnane Osmane, University of Helsinki, Finland
Arnaud Masson, European Space Astronomy Centre (ESAC), Spain

***Correspondence:**
Gian Luca Delzanno
delzanno@lanl.gov

Specialty section:
This article was submitted to Space Physics, a section of the journal Frontiers in Astronomy and Space Sciences

Received: *03 November 2021*
Accepted: *24 January 2022*
Published: *18 February 2022*

Citation:
Delzanno GL and Borovsky JE (2022) The Need for a System Science Approach to Global Magnetospheric Models. Front. Astron. Space Sci. 9:808629. doi: 10.3389/fspas.2022.808629

INTRODUCTION

The Helio2050 workshop was organized in May 2021 to develop a vision for Heliophysics (the Sun, the solar wind, and planetary magnetospheres and ionospheres) for the next 30 years. Acknowledging the tremendous progress made in understanding the various parts of the heliospheric system over many decades, one of the themes for the future that had strong support from diverse areas of the community is the need to understand the heliospheric system as a whole. The same considerations also apply to the Earth's magnetosphere. In fact, the idea of the magnetosphere as a "system of systems" is not new. For decades researchers have applied the tools of system science to data from solar wind, from magnetospheric spacecraft, and from geomagnetic indices and analyzed the correlations between causes (i.e., solar wind drivers) and effects (magnetospheric response). Reviews of magnetospheric system science are in Valdivia et al. (2005), Valdivia et al. (2013), and Borovsky and Valdivia (2018).

Here we are suggesting that system-science techniques be applied in parallel to 1) global magnetospheric simulations and 2) the actual magnetosphere. This methodology will result in a better assessment of the validity of the simulations and it will enable the identification of deficiencies in the simulation models. To validate the models, we will ask the question "Does the simulation behave in the same manner as the magnetosphere behaves?" rather than the standard validation question "How well does the simulation describe the data?". This methodology can also clarify the utility of system-science techniques for the magnetosphere, and help refining those techniques. A final motivation for this methodology is to open an avenue of communication between two diverse magnetospheric research communities: 1) the systems analysis community and 2) the more-mainstream reductionist community of data analysis, instrument designers, plasma and space physicists, and numerical simulators.

MAGNETOSPHERIC SYSTEM SCIENCE

The magnetosphere-ionosphere system exhibits many forms of activity when driven by the solar wind (cf. Borovsky and Valdivia, 2018): magnetospheric convection, morphology changes, substorms, aurora, ionospheric outflows, plasma-wave activity, radiation-belt intensification, and radio emission. Magnetospheric system science examines correlations and information flow between the solar wind and the magnetosphere and looks at statistical properties of the multiple behaviors of the solar-wind-driven magnetosphere. Much of the motivation for these methods comes from the science of systems. The earliest form of magnetospheric system analysis was correlation studies between the spacecraft measurements of the solar wind and geomagnetic indices (Snyder et al., 1963; Bargatze et al., 1985), a method that is still heavily used today, (e.g., McPherron et al., 2015): this methodology yields information about how the solar wind drives the magnetosphere and about various system reaction times. For the driving of the magnetosphere, cause-and-effect among the solar-wind variables can be better established using similar methods based on information transfer (cf. Wing and Johnson, 2019). State vector analysis has built on these simpler solar-wind/magnetosphere correlative studies (Fung and Shao, 2008; Borovsky and Osmane, 2019). Using the proper tools, analysis of magnetospheric time series (typically geomagnetic indices) can yield information about the statistics of magnetospheric dynamics through measurements of fractality, dimensionality, criticality, chaotic output: these time-series studies are discussed in multiple reviews [Voros, 1994; Lakhina, 1994; Klimas et al., 1996; Vassiliadis, 2000; Vassiliadis, 2006; Chapman et al., 2004; Valdivia et al., 2005; Valdivia et al., 2013; Dendy and Chapman, 2006; Sharma, 2010, 2014; Pavlos et al., 2011; and Stepanova and Valdivia, 2016. See also Watkins et al., 2001; Watkins et al., 2012; and Watkins, 2002]. A different type of time-series analysis identifies events in the time series and examines the statistics of event occurrences and amplitudes (Liou et al., 2018). Finally, there is a long history of building and analyzing mathematical (analog) models of the magnetosphere (Smith et al., 1986; Goertz et al., 1991; Goertz et al., 1993; Vassiliadis et al., 1993; Klimas et al., 1997; Klimas et al., 2004; Freeman and Morley 2004; Valdivia et al., 2006; Spencer et al., 2018). These models provide information 1) that can be used to test our physical understanding about how the solar-wind-driven system works, 2) that can inform us about which parameters in the solar wind are key to controlling the reaction of the magnetosphere-ionosphere system, 3) about the global modes of reaction of the magnetosphere to the solar wind, 4) about the flow of information into and through the system, and 5) about where in the system chaotic behaviors emerges. These system methods can improve our scientific knowledge of the magnetosphere (e.g., the uncovering of secondary modes of reaction of the Earth system to the solar wind (Borovsky and Osmane, 2019) and can uncover improved ways to predict space weather (e.g., the expectation of accurately predicting the reaction of the Earth-system to as-yet-unseen severe levels of solar-wind driving (Borovsky and Denton, 2018)). Note that, at present, system science methods do not appear to be used yet in their most general form for space weather prediction outside academia.

GLOBAL MAGNETOSPHERIC MODELS

In what at first might appear as an unrelated topic of magnetospheric research, global magnetospheric models have long been used to describe and understand the behavior of the Earth's magnetosphere. Initial efforts focused on a fluid magnetohydrodynamics (MHD) description of the solar wind and magnetospheric plasmas, owing to the limitations in available computer power (Gombosi et al., 2000; Raeder et al., 2001a; White et al., 2001; Lyon et al., 2004). More recently global magnetospheric models are evolving towards a description of the underlying kinetic plasma beyond MHD, acknowledging the importance of non-MHD physics for several key processes operating in the magnetosphere, such as solar-wind/magnetosphere coupling (day-side reconnection, plasma entry, Kelvin-Helmholz coupling), the ion foreshock, tail reconnection, and for wave-particle interactions [see the discussion in Palmroth et al., 2018]. This is in part because MHD becomes problematic for thin boundary layers such as those at the bow shock and the magnetopause. Examples of beyond-MHD approaches at various stages of development include more-sophisticated fluid models (Wang et al., 2018), hybrid approaches that treat ions kinetically and electrons as a massless fluid (Karimabadi et al., 2014; Lin et al., 2017; Palmroth et al., 2018; Omelchenko et al., 2021), spectral methods (Koshkarov et al., 2021) and MHD models locally coupled with kinetic solvers (Daldorff et al., 2014; Chen et al., 2017). Global magnetospheric models are also becoming more complex in terms of the number of sub-systems that they include. For instance, global MHD models have evolved to include ionospheric models (Fedder and Lyon, 1987; White et al., 2001; Raeder et al., 2001b; Wang et al., 2004; Ridley et al., 2004), ion outflow (Winglee, 2000; Glocer et al., 2009; Brambles et al., 2010), plasmaspheric models (Ouellette et al., 2016; Glocer et al., 2020), inner magnetospheric models to capture drift physics (Toffoletto et al., 2004; Welling and Ridley, 2010; Jordanova et al., 2018), and, as mentioned above, some embed kinetic solvers locally (Chen et al., 2017).

One critical aspect of global magnetospheric models is validation against spacecraft observations. Earlier works focused on applying global MHD codes to specific event challenges (Raeder et al., 1997; Ridley et al., 2002), which led to community-wide event challenges to assess the performance of different codes against observational data (see for instance Pulkkinen et al., 2013). This type of study is very useful in identifying the general trends of different models, in providing physics support and understanding magnetospheric reactions, and in providing comparisons with other codes. However, it is limited in its ability to achieve true validation in light of uncertainties in initial conditions, in particular the lack of knowledge of the actual solar wind hitting the magnetosphere (e.g., Borovsky, 2018; Walsh et al., 2019), boundary conditions, and lack of adequate physics that make it hard to really capture the local spatial and temporal variability of the magnetosphere.

Indeed, the magnetosphere is a high-Reynolds number system that can exhibit unpredictable and chaotic behavior.[1]. Attempts to reproduce all details of its spatial and temporal variability should be taken with a "grain of salt".

Recognizing the limitations just described, other efforts have taken a statistical approach to model validation. Some of these approaches still involve a direct comparison with data. For instance, Ridley et al. (2016) analyzed 662 global MHD simulations at the Community Coordinated Modeling Center to make statistical comparisons of different MHD codes against spacecraft magnetic field measurements. They concluded that models perform worse for higher geomagnetic activity and that coupling global MHD codes with inner magnetospheric models produced statistically better results (the latter conclusion agrees with Rastatter et al. (2013)). Other approaches do not involve a direct comparison with data but rather a 'behavioral' comparison against expressions derived from data. White et al. (2001) used the ISM code to study turbulent transport in the magnetotail under various IMF conditions and computed autocorrelation functions that were in reasonable agreement with autocorrelation functions calculated from ISEE-2 spacecraft measurements in the magnetotail. Specifically, the simulations could recover the general ordering of the decorrelation times for magnetic field component B_x, density n and magnetic field components B_y and B_z and the fact that the velocity components decorrelated more rapidly than the magnetic field components and density (Fig. 4 of White et al. (2001)) but could not recover the long tails seen in the data. El-Alaoui et al. (2013) studied plasma-sheet turbulence with MHD simulations and compared simulation power spectral densities against power spectral densities calculated from THEMIS spacecraft data, finding good agreement in the inertial range but not in the dissipative range. Gordeev et al. (2015) used different MHD models to evaluate several quantities representative of magnetospheric activity (examples include the subsolar magnetopause distance or the cross polar cap potential) against empirical relations obtained from spacecraft data. They found that no code provided satisfactory scores for all the magnetospheric variables considered. Haiducek et al. (2020) performed a month-long MHD simulation wherein over 100 substorms occurred: to validate the model for substorm occurrence, a distribution of substorm-to-substorm waiting times from the code was compiled and compared to equivalent distributions created from geomagnetic indices. The comparison showed a magnetospheric response in the code that was qualitatively similar to that observed for the real magnetosphere. The MHD simulation was also shown to have a small but statistically significant skill in predicting substorm occurrence times.

[1]Note, also, that collisionless or weakly-collisional plasmas can develop an effective viscosity due to kinetic physics that can be significantly larger than that induced by collisions (see, for instance, Squire et al. (2017)) and that, even if this might effectively lower the Reynolds number of the system, an MHD description would still be inadequate (see also the discussion in Borovsky and Gary (2009)).

TABLE 1 | Examples of equivalent quantities that could be compared between simulations and the magnetospheric systems.

Quantity in simulation	Quantity in magnetospheric system
Magnetospheric convection	Kp, am indices
Inner edge of electron plasma sheet	MBI index
Ion pressure	ion pressure
Ion-plasma-sheet number density	ion-plasma-sheet number density
Nightside electrojet current	AL index
Cross-polar-cap ionospheric current	PCI index
Flux of 1-MeV radiation-belt electrons	Flux of 1-MeV radiation-belt electrons
Flux of 130-keV substorm electrons	Flux of 130-keV substorm electrons
Power in electron precipitation	Power in electron precipitation
Power in ion precipitation	Power in ion precipitation
ULF wave intensity	ULF index

DISCUSSION: SYSTEM SCIENCE OF GLOBAL MAGNETOSPHERIC MODELS

In this perspective, we point out the need to apply system science tools to global magnetospheric models to understand if the system behavior of the global models is the same as the system behavior of the real magnetosphere and to overcome the limitations described above. There are clear advantages to this strategy. First, this approach offers the opportunity to validate the global models statistically, without attempting a direct comparison with spacecraft measurements in a high-Reynolds-number magnetosphere. Second, insight could be gained from a side-by-side statistical comparison of system science techniques applied to the outputs of global models and to spacecraft data. One could look at classic quantities of non-linear time series analysis (such as fractality, dimensionality, Lyapunov exponents, . . .) and check whether these quantities are the same in the models and in the real data. For those quantities that do not behave in the same manner, one can investigate why the behavior is different. From a correlation-analysis or information-analysis point of view, several natural questions immediately arise:

1) Are the same solar-wind variables important in the simulation as in the real system?
2) Is the derived driver function for the simulation similar to the derived driver function of the real system?
3) Are the time lags the same in the simulation and the real system?
4) Does the simulation show the same degrees of correlation as does the real system?
5) Does the simulation show the same modes of reaction to the solar wind as does the real system?
6) Does the code exhibit the same patterns of information flow as does the magnetosphere?

Third, as a corollary to the previous point, the system science of global models will facilitate exposing the deficiencies of the models. By turning on and off certain couplings in the simulations, one could ascertain how well the simulations reproduce the statistical correlations of the real system and what is the sensitivity to the various coupling elements. This

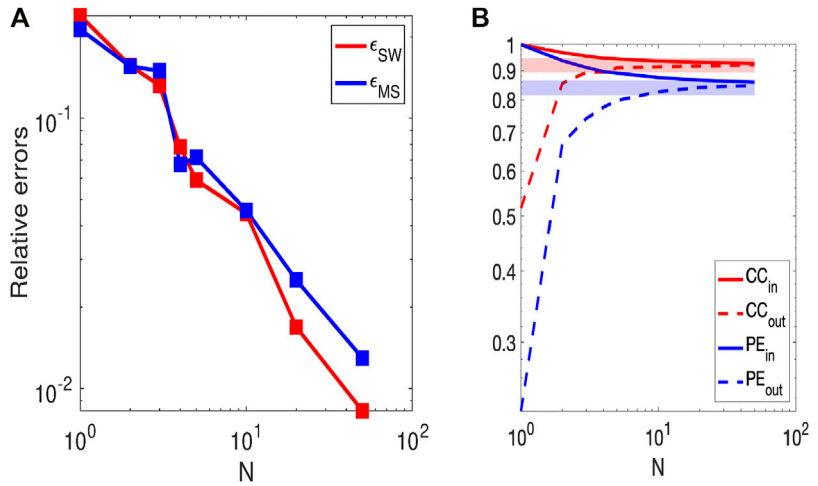

FIGURE 1 | (A) Relative errors for the mean solar wind state (red line, maximum over 8 state variables) and magnetospheric (blue line, maximum over 11 state variables) coefficients as a function of the number of data points per independent variable N used in the canonical correlation analysis. The errors are relative to the results of Borovsky and Denton (2018). **(B)** Prediction efficiency and correlation coefficient as a function of N, obtained for points inside (solid line, labelled as "in") and outside (dashed line, "out") of a given sample. The shaded areas are within ±3% of CC (red) and PE (blue) evaluated in BD2018 over the whole dataset. See the text for more details.

will also provide guidance on what parts of the global models need more improvement. Fourth, from the perspective of system science of the real data, it could provide the physics basis to understand the meaning of the driver functions and state vectors identified by system science tools.

To enable the application of system science tools to global models and its comparison against data, the first step is to determine a set of measurements from the global simulations and match them with an equivalent set of measurements in the magnetospheric system. **Table 1** shows examples of such equivalent quantities. Initially one could look at a single quantity in the simulations and the equivalent quantity in the magnetosphere to 1) compare the statistical behaviors of the pair of quantities, and 2) discern if the correlations with the solar wind are similar. Next, time-dependent state vectors comprised of multiple quantities could be created with the goal of 1) discerning whether the simulations exhibit the same collective modes of reaction to the solar wind as does the magnetosphere, 2) discerning whether the simulations have similar composite scalars as does the magnetosphere, and 3) discerning whether the simulations have the same high vector-vector correlations with the solar wind as the magnetosphere does.

An important question to consider is how much data would actually be needed to perform a meaningful system science analysis of global models. There are two distinct aspects to this point. The first is how much data from the solar wind input is necessary to obtain a magnetospheric response that is sufficiently representative of the variability of the environment. The second is the computational cost to obtain the necessary data through the simulations. To answer the first point, we turn to the analysis performed by Borovsky and Denton (2018) (hereafter "BD2018"). They used canonical correlation analysis (CCA) to correlate 8 solar wind state variables and 11 magnetospheric state variables for the years 1991–2007, a total of 102,672 hourly points for each state variable, i.e., 102,672*19 = 1,950,768 total points. They found a high prediction efficiency (PE) of 84% and a correlation coefficient (CC) of 0.92. We have performed the same canonical correlation analysis on a subset of the data to understand the minimum dataset that would give us a similar PE and CC. To do this, we select samples with N_{sample} = 19*N points randomly from the whole dataset; perform CCA on those N_{sample} points; construct S_1^{in} (S_1^{out}) and E_1^{in} (E_1^{out}) from the CCA coefficients for points inside (outside) the sample; compute CC_{in} (CC_{out}) between the solar wind state vector S_1^{in} (S_1^{out}) and the magnetospheric state vector E_1^{in} (E_1^{out}); compute the linear regression relating S_1^{in} to E_1^{in} and S_1^{out} to E_1^{out}; use the linear regression formula to predict values of E_1 from S_1, for the data points within and outside the sample; compute PE_{in} and PE_{out} as in BD2018. We also compute the error of the coefficients of each state variable relative to those found in BD2018. For a generic coefficient C_i, we define the relative error as $\varepsilon = max_i \frac{|C_i - C_i^{BD2018}|}{\sum_i |C_i^{BD2018}|}$. Note that we repeat this procedure 100 times and average the results, to reduce the noise associated with random sampling. The results are plotted in **Figure 1A**, where we show the relative error for the solar wind state vector (red line) and magnetospheric state vector (blue line) versus the number of points per variable N. One can see that in general there is a decreasing trend of the error and that with only 3 points per variable (i.e., 57 points) the error is fairly small, ~10%. **Figure 1B** show CC and PE versus N. CC_{in} and PE_{in} are monotonically decreasing functions of N (note that for N = 1, CC_{in} = PE_{in} = 1 because CCA can fit the data points exactly) while CC_{out} and PE_{out} are monotonically increasing functions of N. Asymptotically, all quantities converge to the

values of BD2018 computed from the whole dataset. N~3-4 (7) is sufficient for CC_{out} (PE_{out}) to be within 3% of the results of BD2018, i.e., to be within the shaded area of **Figure 1B**. These results are consistent with those of Hair et al. (2010) who indicate that CCA can be applied effectively with only 10 data points per independent variable, showing that CCA is extremely robust and does not need a lot of data. Finally, we also note that applying CCA on the data for January 2005 (i.e., the same time interval used by Haiducek et al. (2020) to study substorm onset with global MHD) yields $CC_{in} = 0.93$, $CC_{out} = 0.97$, $PE_{in} = 0.87$, and $PE_{out} = 0.79$, with $\epsilon_{SW} = 0.09$ and $\epsilon_{MS} = 0.13$. Although preliminary, these results are very encouraging as they show that fairly little data is sufficient to enable effective multi-variable correlation analysis. In terms of computational performance, we note that currently global MHD codes are sufficiently fast to enable system-science studies. For instance, the data from January 2005 is sufficient for meaningful CCA and so the simulation output from Haiducek et al. (2020) could already be used for this purpose. As another example, the GAMERA-REMIX code (Zhang et al., 2019; Sorathia et al., 2020), which combines the GAMERA global MHD solver and the REMIX ionospheric potential solver, runs at ~3,000 core-hours per hour of real time [K. Sorathia, private communication], implying that a simulation study that requires ~200 hourly points could be completed with ~600,000 core-hours. These performance numbers correspond to the high-resolution simulations, e.g., resolving plasma sheet mesoscale dynamics (Sorathia et al., 2021). This is a fairly small allocation on modern high-performance computing architectures. On the other hand, the cost of a single computational run of the more sophisticated global models under development is still very high. For instance, a representative simulation cost of the hybrid global code HYPERS is ~1-million core-hrs for a 1-hour-long simulation of the Earth's magnetosphere [Y. Omelchenko, private communication], extrapolated from the simulations presented in Omelchenko et al. (2021) for that specific resolution. As another example, the recent first 6D run of the hybrid global code Vlasiator cost ~15-million core-hrs for a 30-minutes-long simulation of the coupled solar wind—magnetosphere system using the Earth's dipole magnetic field [M. Palmroth, private communication]. Further computational optimization of the beyond-MHD global codes will be necessary to take full advantage of the upcoming exascale computing facilities and render statistical studies accessible with these codes. Note also that approaches targeting information theory have already been applied effectively without requiring as many simulation runs, e.g., Johnson et al. (2019) who are using transfer entropy to study causal relationships in a single global hybrid simulation run. We therefore conclude that a system science approach to global magnetospheric models is feasible with present-day tools and should be pursued. In general, it will be important to test a variety of system-science methods to obtain complementary information and understanding of the system.

As a final remark, we have focused this perspective on global magnetospheric models because of the general interest of the magnetospheric community to develop a holistic view of the magnetosphere. However, many of the same considerations are still applicable to the individual sub-systems and much could be learned from a side-by-side system science comparison of models and spacecraft data at the sub-system level.

DATA AVAILABILITY STATEMENT

The data set used to perform the analysis shown in Figure 1 can be found in the **Supplementary Material** information of Borovsky and Denton (2018). The data of Figure 1 is provided as **Supplementary Material**. Further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

GLD and JB both equally contributed to the ideas presented in the manuscript and to its writing.

FUNDING

GLD was supported by the Laboratory Directed Research and Development program at Los Alamos National Laboratory (LANL) under project 20220104DR. LANL is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (DOE) (Contract No. 89233218CNA000001). JB was supported at the Space Science Institute by the NSF GEM Program via grant AGS-2027569, by the NASA Heliophysics LWS program via award NNX16AB75G, by the NASA HERMES Interdisciplinary Science Program via grant 80NSSC21K1406, and by the NASA Heliophysics Guest Investigator Program via award NNX17AB71G.

ACKNOWLEDGMENTS

The authors wish to thank Humberto Godinez, Oleksandr Koshkarov, Slava Merkin, Yuri Omelchenko, Minna Palmroth, Vadim Roytershteyn, Kareem Sorathia and Simon Wing for useful conversations.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fspas.2022.808629/full#supplementary-material

REFERENCES

Bargatze, L. F., Baker, D. N., McPherron, R. L., and Hones, E. W. (1985). Magnetospheric Impulse Response for many Levels of Geomagnetic Activity. *J. Geophys. Res.* 90, 6387–6394. doi:10.1029/ja090ia07p06387

Borovsky, J. E., and Denton, M. H. (2018). Exploration of a Composite index to Describe Magnetospheric Activity: Reduction of the Magnetospheric State Vector to a Single Scalar. *J. Geophys. Res. Space Phys.* 123, 7384–7412. doi:10.1029/2018ja025430

Borovsky, J. E., and Gary, S. P. (2009). On Shear Viscosity and the Reynolds Number of Magnetohydrodynamic Turbulence in Collisionless Magnetized Plasmas: Coulomb Collisions, Landau Damping, and Bohm Diffusion. *Phys. Plasmas* 16, 082307. doi:10.1063/1.3155134

Borovsky, J. E., and Osmane, A. (2019). Compacting the Description of a Time-dependent Multivariable System and its Multivariable Driver by Reducing the State Vectors to Aggregate Scalars: the Earth's Solar-Wind-Driven Magnetosphere. *Nonlin. Process. Geophys.* 26, 429–443. doi:10.5194/npg-26-429-2019

Borovsky, J. E. (2018). The Spatial Structure of the Oncoming Solar Wind at Earth and the Shortcomings of a Solar-Wind Monitor at L1. *J. Atmos. Solar-Terrestrial Phys.* 177, 2–11. doi:10.1016/j.jastp.2017.03.014

Borovsky, J. E., and Valdivia, J. A. (2018). The Earth's Magnetosphere: A Systems Science Overview and Assessment. *Surv. Geophys.* 39, 817–859. doi:10.1007/s10712-018-9487-x

Brambles, O. J., Lotko, W., Damiano, P. A., Zhang, B., Wiltberger, M., and Lyon, J. (2010). Effects of Causally Driven Cusp O+ Outflow on the Storm Time Magnetosphere-Ionosphere System Using a Multifluid Global Simulation. *J. Geophys. Res.:Space Phys.* 115 (A9), A00J04. doi:10.1029/2010ja015469

Chapman, S. C., Dendy, R. O., and Watkins, N. W. (2004). Robustness and Scaling: Key Observables in the Complex Dynamic Magnetosphere. *Plasma Phys. Control Fusion* 46, B157–B166. doi:10.1088/0741-3335/46/12b/014

Chen, Y., Toth, G., Cassak, P., Jia, X., Gombosi, T. I., Slavin, J. A., et al. (2017). Global Three-Dimensional Simulation of Earth's Dayside Reconnection Using a Two-Way Coupled Magnetohydrodynamics with Embedded Particle-In-Cell Model: Initial Results. *J. Geophys. Res. Space Phys.* 122 (1010318–10), 335. doi:10.1002/2017ja024186

Daldorff, L. K. S., Tóth, G., Gombosi, T. I., Lapenta, G., Amaya, J., Markidis, S., et al. (2014). Two-way Coupling of a Global Hall Magnetohydrodynamics Model with a Local Implicit Particle-In-Cell Model. *J. Comput. Phys.* 268, 236–254. doi:10.1016/j.jcp.2014.03.009

Dendy, R. O., and Chapman, S. C. (2006). Characterization and Interpretation of Strongly Nonlinear Phenomena in Fusion, Space and Astrophysical Plasmas. *Plasma Phys. Control Fusion* 48, B313–B328. doi:10.1088/0741-3335/48/12b/s30

El-Alaoui, M., Richard, R. L., Ashour-Abdalla, M., Goldstein, M. L., and Walker, R. J. (2013). Dipolarization and Turbulence in the Plasma Sheet during a Substorm: THEMIS Observations and Global MHD Simulations. *J. Geophys. Res. Space Phys.* 118, 7752–7761. doi:10.1002/2013JA019322

Fedder, J. A., and Lyon, J. G. (1987). The Solar Wind-Magnetosphere-Ionosphere Current-Voltage Relationship. *Geophys. Res. Lett.* 14, 880–883. doi:10.1029/gl014i008p00880

Freeman, M. P., and Morley, S. K. (2004). A Minimal Substorm Model that Explains the Observed Statistical Distribution of Times between Substorms. *Geophys. Res. Lett.* 31, L128071–L128074. doi:10.1029/2004gl019989

Fung, S. F., and Shao, X. (2008). Specification of Multiple Geomagnetic Responses to Variable Solar Wind and IMF Input. *Ann. Geophys.* 26, 639–652. doi:10.5194/angeo-26-639-2008

Glocer, A., Toth, G., Gombosi, T., and Welling, D. (2009). Modeling Ionospheric Outflows and Their Impact on the Magnetosphere, Initial Results. *J. Geophys. Res.* 114, A05216. doi:10.1029/2009ja014053

Glocer, A., Welling, D., Chappell, C. R., Toth, G., Fok, M.-C., Komar, C., et al. (2020). A Case Study on the Origin of Near-Earth Plasma. *J. Geophys. Res. Space Phys.* 125, e2020JA028205. doi:10.1029/2020ja028205

Goertz, C. K., Shan, L.-H., and Smith, R. A. (1993). Prediction of Geomagnetic Activity. *J. Geophys. Res.* 98, 7673–7684. doi:10.1029/92ja01193

Goertz, C. K., Smith, R. A., and Shan, L.-H. (1991). Chaos in the Plasma Sheet. *Geophys. Res. Lett.* 18, 1639–1642. doi:10.1029/91gl01782

Gombosi, T. I., DeZeeuw, D. L., Groth, C. P. T., and Powell, K. G. (2000). Magnetospheric Configuration for Parker-spiral IMF Conditions: Results of a 3D AMR MHD Simulation. *Adv. Space Res.* 26, 139–149. doi:10.1016/s0273-1177(99)01040-6

Gordeev, E., Sergeev, V., Honkonen, I., Kuznetsova, M., Rastätter, L., Palmroth, M., et al. (2015). Assessing the Performance of Community-Available Global MHD Models Using Key System Parameters and Empirical Relationships. *Space Weather* 13, 868–884. doi:10.1002/2015SW001307

Haiducek, J. D., Welling, D. T., Morley, S. K., Ganushkina, N. Y., and Chu, X. (2020). Using Multiple Signatures to Improve Accuracy of Substorm Identification. *J. Geophys. Res. Space Phys.* 125, e2019JA027559. doi:10.1029/2019ja027559

Hair, J. F., Black, W. C., Babin, B. J., and Anderson, R. E. (2010). *Canonical Correlation: A Supplement to Multivariate Data Analysis*. Upper Saddle River, New Jersey: Pearson Prentice Hall Publishing.

Johnson, J., Cheng, L., Lin, Y., Wing, S., Wang, X., Perez, J. D., et al. (2019). *A System Science Approach to Understanding the Coupling between Tail Flows and Alfvenic Poynting Flux in a Global Hybrid Simulation*. American Geophysical Union, Fall meeting. Abstract #SM13F-3365.

Jordanova, V. K., Delzanno, G. L., Henderson, M. G., Godinez, H. C., Jeffery, C. A., Lawrence, E. C., et al. (2018). Specification of the Near-Earth Space Environment with SHIELDS. *J. Atmos. Solar-Terrestrial Phys.* 177, 148–159. doi:10.1016/j.jastp.2017.11.006

Karimabadi, H., Roytershteyn, V., Vu, H. X., Omelchenko, Y. A., Scudder, J., Daughton, W., et al. (2014). The Link between Shocks, Turbulence, and Magnetic Reconnection in Collisionless Plasmas. *Phys. Plasmas* 21 (6), 062308. doi:10.1063/1.4882875

Klimas, A. J., Uritsky, V. M., Vassiliadis, D., and Baker, D. N. (2004). Reconnection and Scale-free Avalanching in a Driven Current-Sheet Model. *J. Geophys. Res.* 109, A02218 1–14. doi:10.1029/2003ja010036

Klimas, A. J., Vassiliadis, D., and Baker, D. N. (1997). Data-derived Analogues of the Magnetospheric Dynamics. *J. Geophys. Res.* 102, 26993–27009. doi:10.1029/97ja02414

Klimas, A. J., Vassiliadis, D., Baker, D. N., and Roberts, D. A. (1996). The Organized Nonlinear Dynamics of the Magnetosphere. *J. Geophys. Res.* 101, 13089–13113. doi:10.1029/96ja00563

Koshkarov, O., Manzini, G., Delzanno, G. L., Pagliantini, C., and Roytershteyn, V. (2021). The Multi-Dimensional Hermite-Discontinuous Galerkin Method for the Vlasov-Maxwell Equations. *Comp. Phys. Commun.* 264, 107866. doi:10.1016/j.cpc.2021.107866

Lakhina, G. S. (1994). Solar Wind-Magnetosphere-Ionosphere Coupling and Chaotic Dynamics. *Surv. Geophys.* 15, 703–754. doi:10.1007/bf00666091

Lin, Y., Wing, S., Johnson, J. R., Wang, X. Y., Perez, J. D., and Cheng, L. (2017). Formation and Transport of Entropy Structures in the Magnetotail Simulated with a 3-D Global Hybrid Code. *Geophys. Res. Lett.* 44 (12), 5892–5899. doi:10.1002/2017gl073957

Liou, K., Sotirelis, T., and Richardson, I. (2018). Substorm Occurrence and Intensity Associated with Three Types of Solar Wind Structure. *J. Geophys. Res. Space Phys.* 123, 485–496. doi:10.1002/2017ja024451

Lyon, J. G., Fedder, J. A., and Mobarry, C. M. (2004). The Lyon-Fedder-Mobarry (LFM) Global MHD Magnetospheric Simulation Code. *J. Atmos. Solar-Terrestrial Phys.* 66, 1333–1350. doi:10.1016/j.jastp.2004.03.020

McPherron, R. L., Hsu, T.-S., and Chu, X. (2015). An Optimum Solar Wind Coupling Function for theALindex. *J. Geophys. Res. Space Phys.* 120, 2494–2515. doi:10.1002/2014ja020619

Omelchenko, Y. A., Chen, L.-J., and Ng, J. (2021). 3D Space-Time Adaptive Hybrid Simulations of Magnetosheath High-Speed Jets. *J. Geophys. Res. Space Phys.* 126, e2020JA029035. doi:10.1029/2020ja029035

Ouellette, J. E., Lyon, J. G., Brambles, O. J., Zhang, B., and Lotko, W. (2016). The Effects of Plasmaspheric Plumes on Dayside Reconnection. *J. Geophys. Res. Space Phys.* 121, 4111–4118. doi:10.1002/2016ja022597

Palmroth, M., Ganse, U., Pfau-Kempf, Y., Battarbee, M., Turc, L., Brito, T., et al. (2018). Vlasov Methods in Space Physics and Astrophysics. *Living Rev. Comput. Astrophys* 4 (1), 1. doi:10.1007/s41115-018-0003-2

Pavlos, G. P., Iliopoulos, A. C., Athanasiou, M. A., Karakatsanis, L. P., Tsoutsouras, V. G., Sarris, E. T., et al. (2011). Complexity in Space Plasmas: Universality of Non-equilibrium Physical Processes. *AIP Conf. Proc.* 1320, 77–81. doi:10.1063/1.3544341

Pulkkinen, A. (2013). Community-wide Validation of Geospace Model Ground Magnetic Field Perturbation Predictions to Support Model Transition to Operations. *SpaceWeather* 11, 369–385. doi:10.1002/swe.20056

Raeder, J., Berchem, J., Ashour-Abdalla, M., Frank, L. A., Paterson, W. R., Ackerson, K. L., et al. (1997). Boundary Layer Formation in the Magnetotail: Geotail Observations and Comparisons with a Global MHD Simulation. *Geophys. Res. Lett.* 24, 951. doi:10.1029/97gl00218

Raeder, J., Wang, Y., and Fuller-Rowell, T. J. (2001b). "Geomagnetic Storm Simulation with a Coupled Magnetosphere-Ionosphere-Thermosphere Model," in *Space Weather*. Editors P. Song, H. J. Singer, and G. L. Siscoe.

Raeder, J., Wang, Y. L., Fuller-Rowell, T. J., and Singer, H. J. (2001a). Global Simulation of Magnetospheric Space Weather Effects of the Bastille Day Storm. *Solar Phys.* 204, 325–338. doi:10.1023/a:1014228230714

Rastatter, L. (2013). Geospace Environment Modeling 2008–2009 challenge: Dst index. *Space Weather* 11, 187–205. doi:10.1002/swe.20036

Ridley, A. J., De Zeeuw, D. L., and Rastatter, L. (2016). Rating Global Magnetosphere Model Simulations through Statistical Data-Model Comparisons. *Space Weather* 14, 819–834. doi:10.1002/2016sw001465

Ridley, A. J., Gombosi, T., and De Zeeuw, D. L. (2004). Ionospheric Control of the Magnetospheric Configuration: Conductance. *Ann. Geophys.* 22, 567–584. doi:10.5194/angeo-22-567-2004

Ridley, A. J., Hansen, K. C., Toth, G., De Zeeuw, D. L., Gombosi, T. I., and Powell, K. G. (2002). University of Michigan MHD Results of the Geospace Global Circulation Model Metrics challenge. *J. Geophys. Res.* 107 (A10), 1290. doi:10.1029/2001ja000253

Sharma, A. S. (2014). Complexity in Nature and Data-Enabled Science: The Earth's Magnetosphere. *AIP Conf. Proc.* 1582, 35–45.

Sharma, A. S. (2010). The Magnetosphere: A Complex Driven System. *AIP Conf. Proc.* 1308, 120–131. doi:10.1063/1.3526148

Smith, R. A., Goertz, C. K., and Grossman, W. (1986). Thermal Catastrophe in the Plasma Sheet Boundary Layer. *Geophys. Res. Lett.* 13, 1380–1383. doi:10.1029/gl013i013p01380

Snyder, C. W., Neugebauer, M., and Rao, U. R. (1963). The Solar Wind Velocity and its Correlation with Cosmic-ray Variations and with Solar and Geomagnetic Activity. *J. Geophys. Res.* 68, 6361. doi:10.1029/jz068i024p06361

Sorathia, K. A., Merkin, V. G., Panov, E. V., Zhang, B., Lyon, J. G., Garretson, J., et al. (2020). Ballooning-Interchange Instability in the Near-Earth Plasma Sheet and Auroral Beads: Global Magnetospheric Modeling at the Limit of the MHD Approximation. *Geophys. Res. Lett.* 47, 18. doi:10.1029/2020GL088227

Sorathia, K. A., Michael, A., Merkin, V. G., Ukhorskiy, A. Y., Turner, D. L., Lyon, J. G., et al. (2021). The Role of Mesoscale Plasma Sheet Dynamics in Ring Current Formation. *Front. Astron. Space Sci.* 8, 761875. doi:10.3389/fspas.2021.761875

Spencer, E., Vadepu, S. K., Srinivas, P., Patra, S., and Horton, W. (2018). The Dynamics of Geomagnetic Substorms with the WINDMI Model. *Earth PlanetSpace* 70, 118. doi:10.1186/s40623-018-0882-9

Squire, J., Kunz, M. W., Quataert, E., and Schekochihin, A. A. (2017). Kinetic Simulations of the Interruption of Large-Amplitude Shear-Alfvén Waves in a High-β Plasma. *Phys. Rev. Lett.* 119, 155101. doi:10.1103/physrevlett.119.155101

Stepanova, M., and Valdivia, J. A. (2016). Contribution of Latin-American Scientists to the Study of the Magnetosphere of Earth. A Review. *Adv. Space Res.* 58, 1968–1985. doi:10.1016/j.asr.2016.03.023

Toffoletto, F. R., Sazykin, S., Spiro, R. W., Wolf, R. A., and Lyon, J. G. (2004). RCM Meets LFM: Initial Results of One-Way Coupling. *J. Atmos. Solar-terr. Phys.* 66, 1361–1370. doi:10.1016/j.jastp.2004.03.022

Valdivia, J. A., Rogan, J., Muñoz, V., Gomberoff, L., Klimas, A., Vassiliadis, D., et al. (2005). The Magnetosphere as a Complex System. *Adv. Space Res.* 35, 961. doi:10.1016/j.asr.2005.03.144

Valdivia, J. A., Rogan, J., Muñoz, V., Toledo, B. A., and Stepanova, M. (2013). The Magnetosphere as a Complex System. *Adv. Space Res.* 51, 1934. doi:10.1016/j.asr.2012.04.004

Valdivia, J. A., Rogan, J., Munoz, V., and Toledo, B. (2006). Hysteresis Provides Self-Organization in a Plasma Model. *Space Sci. Rev.* 122, 313–320. doi:10.1007/s11214-006-7846-2

Vassiliadis, D., Sharma, A. S., and Papadopoulos, K. (1993). An Empirical Model Relating the Auroral Geomagnetic Activity to the Interplanetary Magnetic Field. *Geophys. Res. Lett.* 20, 1731–1734. doi:10.1029/93gl01351

Vassiliadis, D. (2000). System Identification, Modeling, and Prediction for Space Weather Environments. *IEEE Trans. Plasma Sci.* 28, 1944–1955. doi:10.1109/27.902223

Vassiliadis, D. (2006). Systems Theory for Geospace Plasma Dynamics. *Rev. Geophys.* 44, RG2002 1–39. doi:10.1029/2004rg000161

Voros, Z. (1994). The Magnetosphere as a Nonlinear System. *Studia Geophysica et Geodaetica* 38, 168–186. doi:10.1007/bf02295912

Walsh, B. M., Bhakyapaibul, T., and Zou, Y. (2019). Quantifying the Uncertainty of Using Solar Wind Measurements for Geospace Inputs. *J. Geophys. Res.* 124, 3291–3302. doi:10.1029/2019ja026507

Wang, L., Germaschewski, K., Hakim, A., Dong, C., Raeder, J., and Bhattacharjee, A. (2018). Electron Physics in 3-D Two-Fluid 10-moment Modeling of Ganymede's Magnetosphere. *J. Geophys. Res. Space Phys.* 123, 2815–2830. doi:10.1002/2017ja024761

Wang, W., Wiltberger, M., Burns, A., Solomon, S., Killeen, T., Maruyama, N., et al. (2004). Initial Results from the Coupled Magnetosphere - Ionosphere - Thermosphere Model: Thermosphere - Ionosphere Responses. *J. Atm. Solar-terr. Phys.* 66, 1425–1441. doi:10.1016/j.jastp.2004.04.008

Watkins, N. W., Freeman, M. P., Chapman, S. C., and Dendy, R. O. (2001). Testing the SOC Hypothesis for the Magnetosphere. *J. Atmos. Solar-Terrestrial Phys.* 63 (13), 1435–1445. doi:10.1016/s1364-6826(00)00245-5

Watkins, N. W., Hnat, B., and Chapman, S. C. (2012). "On Self-Similar and Multifractal Models for the Scaling of Extreme Bursty Fluctuations in Space Plasmas," in *Extreme Events and Natural Hazards: The Complexity Perspective*. Editors A. S. Sharma, A. Bunde, V. P. Dimri, and D. N. Baker. doi:10.1029/2011gm001084

Watkins, N. W. (2002). Scaling in the Space Climatology of the Auroral Indices: Is SOC the Only Possible Description. *Nonlin. Process. Geophys.* 9, 389–397. doi:10.5194/npg-9-389-2002

Welling, D. T., and Ridley, A. J. (2010). Validation of SWMF Magnetic Field and Plasma. *Space Weather* 8 (S03002), 1–11. doi:10.1029/2009sw000494

White, W. W., Schoendorf, J. A., Siebert, K. D., Maynard, N. C., Weimer, D. R., Wilson, G. L., et al. (2001). "MHD Simulation of Magnetospheric Transport at the Mesoscale," in *Space Weather*. Editors P. Song, H. J. Singer, and G. L. Siscoe.

Wing, S., and Johnson, J. R. (2019). Applications of Information Theory in Solar and Space Physics. *Entropy* 21, 140. doi:10.3390/e21020140

Winglee, R. M. (2000). Mapping of Ionospheric Outflows into the Magnetosphere for Varying IMF Conditions. *J. Atmos. Sol.-Terr. Phys.* 62 (6), 527–540. doi:10.1016/s1364-6826(00)00015-8

Zhang, B., Sorathia, K. A., Lyon, J. G., Merkin, V. G., Garretson, J. S., and Wiltberger, M. (2019). Gamera: A Three-Dimensional Finite-Volume Mhd Solver for Non-orthogonal Curvilinear Geometries. *ApJS* 244, 20. doi:10.3847/1538-4365/ab3a4c

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer AO declared a past co-authorship with one of the authors JB to the handling Editor.

frontiers in Astronomy and Space Sciences

ORIGINAL RESEARCH
published: 04 March 2022
doi: 10.3389/fspas.2022.836215

Using Multivariate Imputation by Chained Equations to Predict Redshifts of Active Galactic Nuclei

Spencer James Gibson[1], *Aditya Narendra*[2], *Maria Giovanna Dainotti*[3,4]*, *Malgorzata Bogdan*[5,6], *Agnieszka Pollo*[7,8], *Artem Poliszczuk*[8], *Enrico Rinaldi*[9,10,11] and *Ioannis Liodakis*[12]

[1]Carnegie Mellon University, Pittsburgh, PA, United States, [2]Astronomical Observatory of Jagiellonian University, Kraków, Poland, [3]National Astronomical Observatory of Japan, Mitaka, Japan, [4]Space Science Institute, Boulder, CO, United States, [5]Department of Mathematics, University of Wroclaw, Wrocław, Poland, [6]Department of Statistics, Lund University, Lund, Sweden, [7]Astronomical Observatory of Jagiellonian University, Krakow, Poland, [8]National Centre for Nuclear Research, Warsaw, Poland, [9]Physics Department, University of Michigan, Ann Arbor, MI, United States, [10]Theoretical Quantum Physics Laboratory, RIKEN, Wako, Japan, [11]Interdisciplinary Theoretical and Mathematical Science Program, RIKEN (iTHEMS), Wako, Japan, [12]Finnish Centre for Astronomy with ESO (FINCA), University of Turku, Turku, Finland

OPEN ACCESS

Edited by:
Olga Verkhoglyadova,
NASA Jet Propulsion Laboratory (JPL), United States

Reviewed by:
Ali Luo,
National Astronomical Observatories (CAS), China
Jaime Perea,
Institute of Astrophysics of Andalusia, Spain

***Correspondence:**
Maria Giovanna Dainotti
mariagiovannadainotti@yahoo.it

Specialty section:
This article was submitted to Astrostatistics, a section of the journal Frontiers in Astronomy and Space Sciences

Received: 15 December 2021
Accepted: 24 January 2022
Published: 04 March 2022

Citation:
Gibson SJ, Narendra A, Dainotti MG, Bogdan M, Pollo A, Poliszczuk A, Rinaldi E and Liodakis I (2022) Using Multivariate Imputation by Chained Equations to Predict Redshifts of Active Galactic Nuclei. Front. Astron. Space Sci. 9:836215. doi: 10.3389/fspas.2022.836215

Redshift measurement of active galactic nuclei (AGNs) remains a time-consuming and challenging task, as it requires follow up spectroscopic observations and detailed analysis. Hence, there exists an urgent requirement for alternative redshift estimation techniques. The use of machine learning (ML) for this purpose has been growing over the last few years, primarily due to the availability of large-scale galactic surveys. However, due to observational errors, a significant fraction of these data sets often have missing entries, rendering that fraction unusable for ML regression applications. In this study, we demonstrate the performance of an imputation technique called Multivariate Imputation by Chained Equations (MICE), which rectifies the issue of missing data entries by imputing them using the available information in the catalog. We use the Fermi-LAT Fourth Data Release Catalog (4LAC) and impute 24% of the catalog. Subsequently, we follow the methodology described in Dainotti et al. (ApJ, 2021, 920, 118) and create an ML model for estimating the redshift of 4LAC AGNs. We present results which highlight positive impact of MICE imputation technique on the machine learning models performance and obtained redshift estimation accuracy.

Keywords: redshift, AGNs, BLLs, FSRQs, FERMI 4LAC, machine learning regressors, imputation, MICE

1 INTRODUCTION

Spectroscopic redshift measurement of Active Galactic Nuclei (AGNs) is a highly time-consuming operation and is a strong limiting factor for a large-scale extragalactic surveys. Hence, there is a pressing requirement for alternative redshift estimation techniques that provide reasonably good results Salvato et al. (2019). In current cosmological studies, such alternative redshift estimates, referred to as photometric redshifts, play a key role in our understanding of the Extragalactic Background Light (EBL) origins Wakely and Horan (2008)1, magnetic field structure in the intergalactic medium Marcotulli et al. (2020); Venters and Pavlidou (2013); Fermi-LAT Collaboration et al. (2018) and help in determining the bounds on various cosmological parameters Domínguez et al. (2019); Petrosian (1976); Singal et al. (2013b), Singal et al. (2012),

Singal et al. (2014); Singal (2015); Singal et al. (2013a); Chiang et al. (1995); Ackermann et al. (2015); Singal et al. (2013b); Ackermann et al. (2012).

One technique that has gained significant momentum is the use of machine learning (ML) to determine the photometric redshift of AGNs Brescia et al. (2013), Brescia et al. (2019); Dainotti et al. (2021); Nakoneczny et al. (2019); Jones and Singal (2017); Cavuoti et al. (2014); Fotopoulou and Paltani (2018); Logan and Fotopoulou (2020); Yang et al. (2017); Zhang et al. (2019); Curran (2020); Nakoneczny et al. (2019); Pasquet-Itam and Pasquet (2018); Jones and Singal (2017). Large AGN data sets derived from all-sky surveys like the Wide-field Infrared Survey Explorer (WISE) Brescia et al. (2019); Ilbert et al. (2008); Hildebrandt et al. (2010); Brescia et al. (2013); Wright et al. (2010); D'Isanto and Polsterer (2018) and Sloan Digital Sky Survey (SDSS) Aihara et al. (2011) have played a significant role in the proliferation of ML approaches. However, the quality of the results from an ML approach depends significantly on the size and quality of the training data: the data on which the ML models learn the underlying relationship to predict the redshift. Unfortunately, almost all of these large data sets suffer from the issue of missing entries, which can lead to a considerable portion of the data being discarded.

This is especially problematic in catalogs of smaller size, such as in the case of gamma-ray loud AGNs.

Using the Fermi Fourth Data Release Catalog's (4LAC) gamma-ray loud AGNs Ajello et al. (2020); Abdollahi et al. (2020), Dainotti et al. (2021) demonstrated that ML methods lead to promising results, with a 71% correlation between the predicted and observed redshifts. However, in that study, the training set consists of only 730 AGNs, and a majority of the data (50%) are discarded due to missing entries. More specifically, we have several reasons why the sources are missing also in relation to the variables we consider. Regarding the missing values of the Gaia magnitudes: this could be either because the sources are too faint and thus they undergo the so called Malmquist bias effect (only the brightest sources are visible at high-z) or the coordinates are not accurate enough and the cross-matching is failing to produce a counterpart (the latter is not that likely, the former is much more likely).

Regarding the variables observed in γ-rays: here the source is detected, but it is faint in gamma-rays and again we have the Malmquist bias effect in relation to the detector threshold of Fermi-LAT and/or it does not appear variable and/or the spectral fitting fails to produce values, hence the missing values.

Regarding the multi-wavelength estimates (ν, νf_ν): these depend on the availability of multi-wavelength data from radio to X-rays. If sufficient data exists then a value can be estimated, so the missing values are most likely sources that have not been observed by telescopes. In other words, this does not mean that the sources are necessarily faint, they could be bright, but just no telescope performed follow-up observations.

There is also the possibility to explain the missing values because of the relativistic effects that dominate blazar emission. The relativistic effects, quantified by a parameter called the Doppler factor, boost the observed flux across all frequencies, but also shorten the timescales making sources appear more variable. It has been shown that sources detected in γ-rays have higher Doppler factors and are more variable Liodakis et al. (2017), Liodakis et al. (2018). This would suggest that sources observed more off-axis, i.e., lower Doppler factor, would have a lower γ-ray flux and appear less variable. Therefore introduce more missing values as we have discussed above.

In this study, we address this issue of missing entries using an imputation technique called Multivariate Imputation by Chained Equations (MICE) Van Buuren and Groothuis-Oudshoorn (2011). This technique was also recently used by Luken et al. (2021) for redshift estimation of Radio-loud AGNs.

Luken et al. (2021) test multiple imputation techniques, MICE included, to determine the best tool for reliably imputing missing values. Their study considers the redshift estimation of radio-loud galaxies present in the Australia Telescope Large Area Survey (ATLAS). However, in contrast to our approach where we impute actual missing information in the catalog, they manually set specific percentages of their data as missing and test how effective various imputation techniques are. Their results demonstrate distinctly that MICE is the best imputation technique, leading to the least root mean square error (RMSE) and outlier percentages for the regression algorithms they have tested.

In our study, we are using the updated 4LAC catalog, and using MICE imputations to fill in missing entries, we achieve a training data set which is 98% larger than the one used in Dainotti et al. (2021). We achieve results on this more extensive training set that are comparable to Dainotti et al. (2021) while attaining higher correlations. Furthermore, we are using additional ML algorithms in the SuperLearner ensemble technique, as compared to Dainotti et al. (2021).

Section 2 discusses the specifics of the extended 4LAC data set: how we create the training set, which predictors are used and which outliers are removed. In **Section 3** we discuss the MICE imputation technique, the SuperLearner ensemble with a brief description of the six algorithms used in this analysis, followed by the different feature engineering techniques implemented. Finally, we present the results in **Section 4**, followed by the discussion and conclusions in **Section 5**.

2 SAMPLE

This study uses the Fermi Fourth Data Release Catalog (4LAC), containing 3,511 gamma-ray loud AGNs, 1764 of which have a measured spectroscopic redshift. Two categories of AGNs dominate the 4LAC catalog, BL Lacertae (BLL) objects and Flat Spectrum Radio Quasars (FSRQ). To keep the analysis consistent with Dainotti et al. (2021), we remove all the non-BLL and non-FSRQ AGNs.

These AGNs have 13 measured properties in the 4LAC catalog; however, we only use 11 and a categorical variable that distinguishes BLLs and FSRQs. The two omitted properties in the analysis are *Highest_Energy* and *Fractional_Variability* because 42.5% of the entries are missing, and there is insufficient information to impute them reliably. We consider imputation of predictors which have

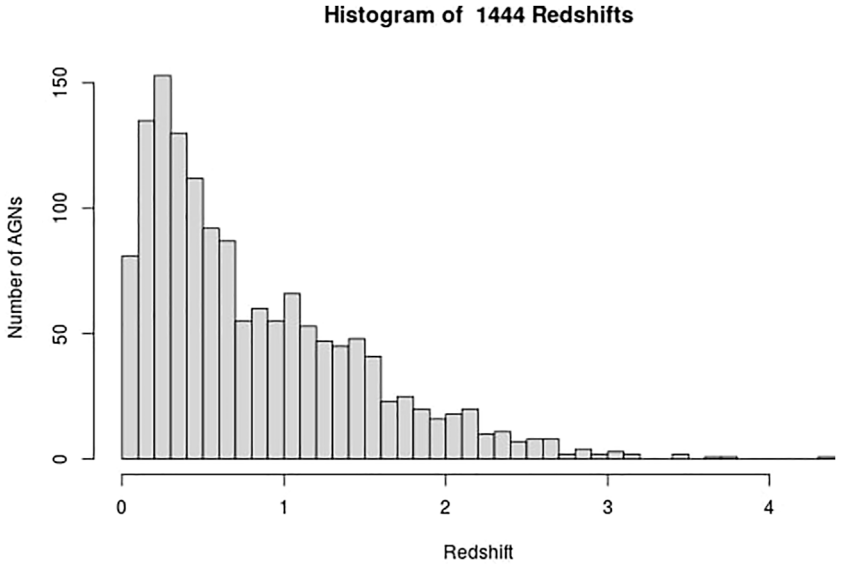

FIGURE 1 | Redshift distribution of the training set.

missing entries in less than 18% of the data. The remaining 11 properties and the categorical variables are *Gaia_G_Magnitude*, *Variability_Index*, *Flux*, *Energy_Flux*, *PL_Index*, *νfν*, *LP_Index*, *Significance*, *Pivot_Energy*, *ν*, and *LP_β* and *LabelNo*, serve as the predictors for the redshift in the machine learning models and are defined in Dainotti et al. (2021) and Ajello et al. (2020). However, some of these properties are not used as they appear in the 4LAC, since they span several orders of magnitude. The properties *Flux*, *Energy_Flux*, *Significance*, *Variability_Index*, *ν*, *νfν*, and *Pivot_Energy* are used in their base-10 logarithmic form. In the categorical variable *LabelNo* we assign the values 2 and 3 to BLLs and FSRQs, respectively. We are not training the ML models to predict the redshift directly. Instead, we train the models to predict $1/(z + 1)$, where z is the redshift. Such a transformation of the target variable is crucial as it helps improve the model's performance. In addition, $1/(z + 1)$ is known as the scale factor and has a more substantial cosmological significance than redshift itself. We remove AGNs with an $LP\_\beta < 0.7$, $LP\_Index > 1$, and $LogFlux > -10.5$, as they are outliers of their respective distributions. These steps lead us to a final data sample of 1897 AGNs, out of which 1,444 AGNs have a measured redshift (see **Figure 1**). These AGNs form the training sample, while the remaining 453 AGNs, which do not have a measured redshift, form the generalization sample.

3 METHODOLOGY

Here we present the various techniques implemented in the study, definitions of the statistical metrics used, and a comprehensive step-by-step description of our procedure to obtain the results. We use the following metrics to measure the performance of our ML model:

- Bias: Mean of the difference between the observed and predicted values.
- σ_{NMAD}: Normalized median absolute deviation between the predicted and observed measurements.
- r: Pearson correlation coefficient between the predicted and observed measurements.
- Root Mean Square Error (RMSE) between the predicted and observed redshift
- Standard Deviation σ between the predicted and observed redshift

We present these metrics for both $\Delta z_{\rm norm}$ and Δz, which are defined as:

$$\Delta z = z_{observed} - z_{predicted} \tag{1}$$

$$\Delta z_{norm} = \frac{\Delta z}{1 + z_{observed}} \tag{2}$$

We also quote the catastrophic outlier percentage, defined as the percentage of predictions that lies beyond the 2σ error. The metrics presented in this study are the same as in Dainotti et al. (2021), allowing for easy comparison.

3.1 Procedure

Here we provide a walk-through of how the final results are obtained. First, we remove all the non-BLL and non-FSRQ AGNs from the 4LAC data set, in addition to outliers, and end up with 1897 AGNs for the total set. Then, we impute the missing entries using MICE (see **Section 3.2**). Having obtained a complete data set, we split it into the training and the generalization sets, depending on whether the AGNs have or do not have a measured redshift value. We aim to train an ensemble model that is the least complex and best suited to the data at hand. For this purpose, we need to test many different algorithms with ten-

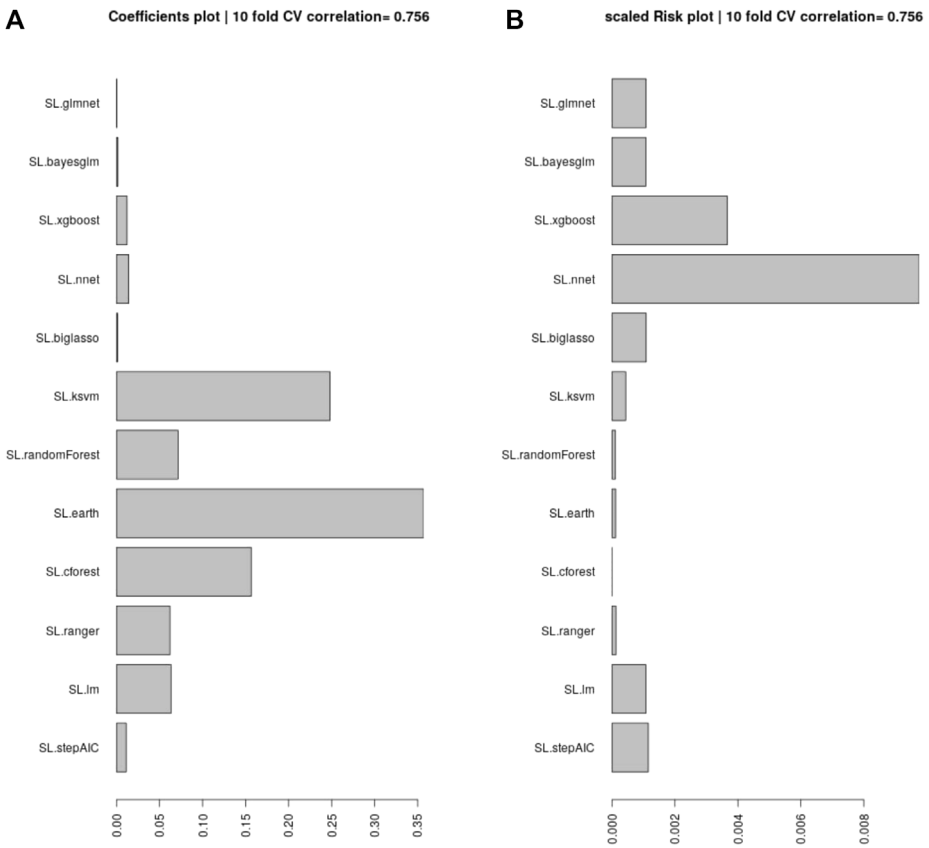

FIGURE 2 | (A): The coefficients assigned by SuperLearner to the algorithms tested. We select the algorithms that have a coefficient above 0.05 to be incorporated into our ensemble. **(B)**: The RMSE error (risk) of each of the algorithms, scaled to show the minimum risk algorithm at 0, which is Cforest. These values are average over one hundred iterations of 10fCV.

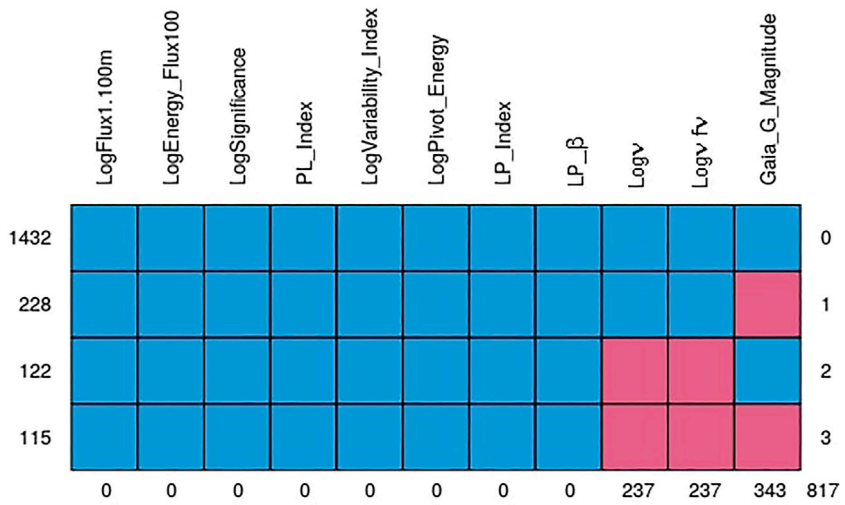

FIGURE 3 | The pattern of the missing data. The blue cells represent complete values, while the pink ones indicate where we have missing data. The first row shows that there are 1,432 AGNs without missing values. Second row shows that there are 228 data points with *Gaia_G_Magnitude* missing. Third row shows that there are 122 data points with Logν and Log$\nu f\nu$ missing. And finally, the last row shows that there are 115 data points with missing values in *Gaia_G_Magnitude*, Logν and Log$\nu f\nu$. The columns indicate that there are 237 missing values in Logν and Log$\nu f\nu$, and 343 missing values in *Gaia_G_Magnitude*. The remaining predictors have no missing entries.

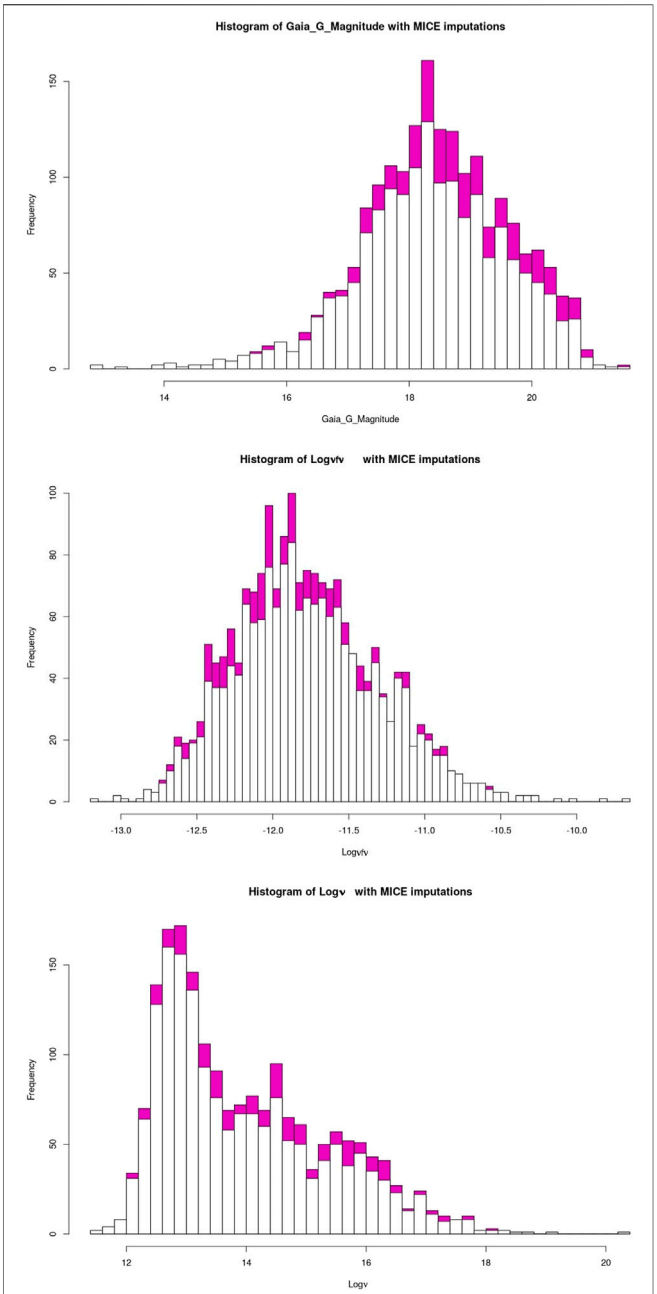

FIGURE 4 | The white bars show the initial distribution of the variables. The magenta bars plotted on top of it are the MICE imputed values. The top plot shows the distribution of *Gaia_G_Magnitude* with and without MICE. The central plot shows this for Logνf_ν, and the bottom plot shows this for Logν.

fold cross-validation (10fCV). Cross-validation is a resampling procedure that uses different portions of the data, in this case 10, to train and test a model, and find out which algorithm performs the best in terms of the previously defined metrics. However, since there is inherent randomness in how the folds are created during 10fCV, we perform 10fCV one hundred times and average the results to derandomize and stabilize them. This repeated k-fold cross-validation technique is standard in evaluating ML models. In each of the one hundred iterations of 10fCV, we train a SuperLearner model (see **Section 3.3**) on the training set using the twelve algorithms shown in **Figure 2**. Finally, averaging over the one hundred iterations, we obtained the coefficients and risk measurements associated with each SuperLearner ensemble model, as well as the individual algorithms. Following the previous step, we pick six algorithms that have coefficients greater than 0.05 (see **Section 3.4** for information about these algorithms).

With the six best ML algorithms, we create an ensemble with SuperLearner and perform the 10fCV one hundred times once more. The final cross-validated results are again an average of these one hundred iterations.

Next, we proceed to show the results obtained without the repeated cross-validation procedure. For this, we simply select a fixed validation set by choosing the last 111 AGNs from the 1,444 AGNs of the previously used training data. Now, with the new training set of 1,333 AGNs, we train a SuperLearner model, with the algorithms being the same as in the cross-validation step, and we predict the redshift of the validation set. We then calculate the same statistical metrics for these results as we did for the cross-validated results. The results on this fixed validation set provide a representative of the performance of the SuperLearner model, which we have explored in more details (and in a more computationally expensive way) during the repeated cross-validation procedure.

3.2 Multivariate Imputation by Chained Equations

Multivariate Imputation by Chained Equations (MICE) is a method for imputing missing values for multivariate data Van Buuren and Groothuis-Oudshoorn (2011); Luken et al. (2021). The multivariate in MICE highlights its use of multiple variables to impute missing values. The MICE algorithm works under the assumption that the data are missing at random (MAR). MAR was first detailed in the paper Rubin (1976). It implies that errors in the system or with users cause the missing entries and not intrinsic features of the object being measured. Furthermore, MAR implies the possibility that the missing entries can be inferred by the other variables present in the data Schafer and Graham (2002). Indeed, this is a strong assumption, and it is our first step to deal with missing data. However, we know that selection biases play an important role for the flux detection. Although this problem is mitigated for the gamma-ray sources, for the G-band magnitude, one can argue that, e.g., BL Lacs are systematically fainter than FSRQs and below the Gaia limiting magnitude. A more in-depth analysis to take this problem into account is worthwhile, but this is beyond the scope of the current paper.

With this assumption, MICE attempts to fill in the absent entries using the complete variables in the data set iteratively. We impute the missing variables 20 times with each iteration of MICE consisting of multiple steps. General practice is to perform the imputation ten times as in Luken et al. (2021) and Van

TABLE 1 | Composition of the training and generalization sets, and Redshift properties on the training set.

Type	Training set	Generalization set	Redshift median	Redshift minimum	Redshift maximum
BLLs	721	450	0.336	3.7×10^{-5}	2.82
FSRQ	723	3	1.12	0.097	4.313
Total	1,444	453	0.628	3.7×10^{-5}	4.313

Buuren and Groothuis-Oudshoorn (2011), but we perform it twenty times to stabilize the imputation.

Here, we use the method "midastouch"—a predictive mean matching (PMM) method Little and Rubin (2019). It works by initializing a feature's missing entries with its mean and then estimating them by training a model using the rest of the complete data. For each prediction, a probability is assigned based on its distance from the value imputed for the desired entry. The missing entry is imputed by randomly drawing from the observed values of the respective predictor, weighted according to the probability defined previously.

The process is repeated for each missing entry until all have been refitted. This new complete table is used as a basis for the next iteration of MICE, where the same process is repeated until the sequence of table converges or a set number of iterations is achieved.

3.3 SuperLearner

SuperLearner Van der Laan et al. (2007) is an algorithm that constructs an ensemble of ML models predictions using a cross-validated metric and a set of normalized coefficients. By default Superlearner uses a ten-fold cross-validation procedure. It outputs a combination of user-provided ML models such that the RMSE of the final prediction is minimized by default Polley and Van der Laan (2010) (or any other user-defined metric defining the expected risk of the task at hand). In our setup, SuperLearner achieves this using 10fCV, where the training data is divided into ten equal portions or folds, the models are trained on nine folds, and the 10th fold is used as a test set. The models predict the target variable of the test set, and based on the RMSE of their predictions, SuperLearner assigns a coefficient. If an algorithm has a lower RMSE in 10fCV, it will be assigned a higher coefficient. Finally, it creates the ensemble as a linear combination of the constituent models multiplied by their respective coefficients. Note that this 10fCV is an internal procedure of model selection to build the SuperLearner ensemble model, and it is separate from the repeated cross-validation procedure which we described in **Section 3.1** and which is used to evaluate the performance and final results.

3.4 The Machine Learning Algorithms Used in Our Analysis

Following Dainotti et al. (2021) we analyze the coefficients assigned by SuperLearner to 12 ML algorithms, and pick those with a value greater than 0.05. In **Figure 2**, we show all the ML algorithms tested, and their coefficients. We pick the six algorithms above the 0.05 cutoff, which are: Enhanced Adaptive Regression Through Hinge (EARTH), KSVM, Cforest, Ranger, Random Forest, and Linear Model. We provide brief explanations for each of them below.

Enhanced Adaptive Regression Through Hinges (EARTH) is an algorithm that allows for better modeling of predictor interaction and non-linearity in the data compared to the linear model. It is based on the Multivariate Adaptive Regression Splines method (MARS) Friedman and Roosen (1995). EARTH works by fitting a sum or product of hinges. Hinges are part-wise linear fits of the data that are joined such that the sum-of-squares residual error is minimized with each added term.

KSVM is an R implementation of the Support Vector Regression method (SVR). Similar to Support Vector Machine (SVM) Cortes and Vapnik (1995), SVR uses a kernel function to send its inputs to a higher-dimensional space where the data is linearly separable by a hyper-plane. SVR aims to fit this hyper-plane such that the prediction error is within a pre-specified threshold. For our purposes, KSVM uses the Gaussian kernel with the default parameters.

The Random Forest algorithm Breiman (2001); Ho (1995) seeks to extend decision trees capabilities by simultaneously generating multiple, independent decision trees. For regression tasks, Random Forest will return the average of the outputs of each of the generated decision trees. An advantage of Random Forest over decision trees is the reduction in the variance. However, Random Forest often suffers from low interpretability.

The Ranger algorithm is similar to Random Forest with the difference of extremely randomized trees (ERTs) Geurts et al. (2006) and quicker implementation.

Similar to Random Forest, the Cforest algorithm Hothorn et al. (2006) builds conditional inference trees that perform splits on significance tests instead of information gain.

We use the ordinary least squares (OLS) linear model found in the SuperLearner package. This model aims to minimize the mean squared error.

Note that we are using the default hyperparameter settings for all the algorithms.

3.5 Feature Engineering

Feature engineering is a broad term that incorporates two techniques: feature selection and feature creation. Feature selection is a method where the best predictors of a response variable are chosen from a larger pool of predictors. There exist multiple methods to perform feature selection. We are using the Least Absolute Selection and Shrinkage Operator (LASSO) method. Feature selection is an essential part of any ML study as it reduces the dimensionality of the data and minimizes the risk

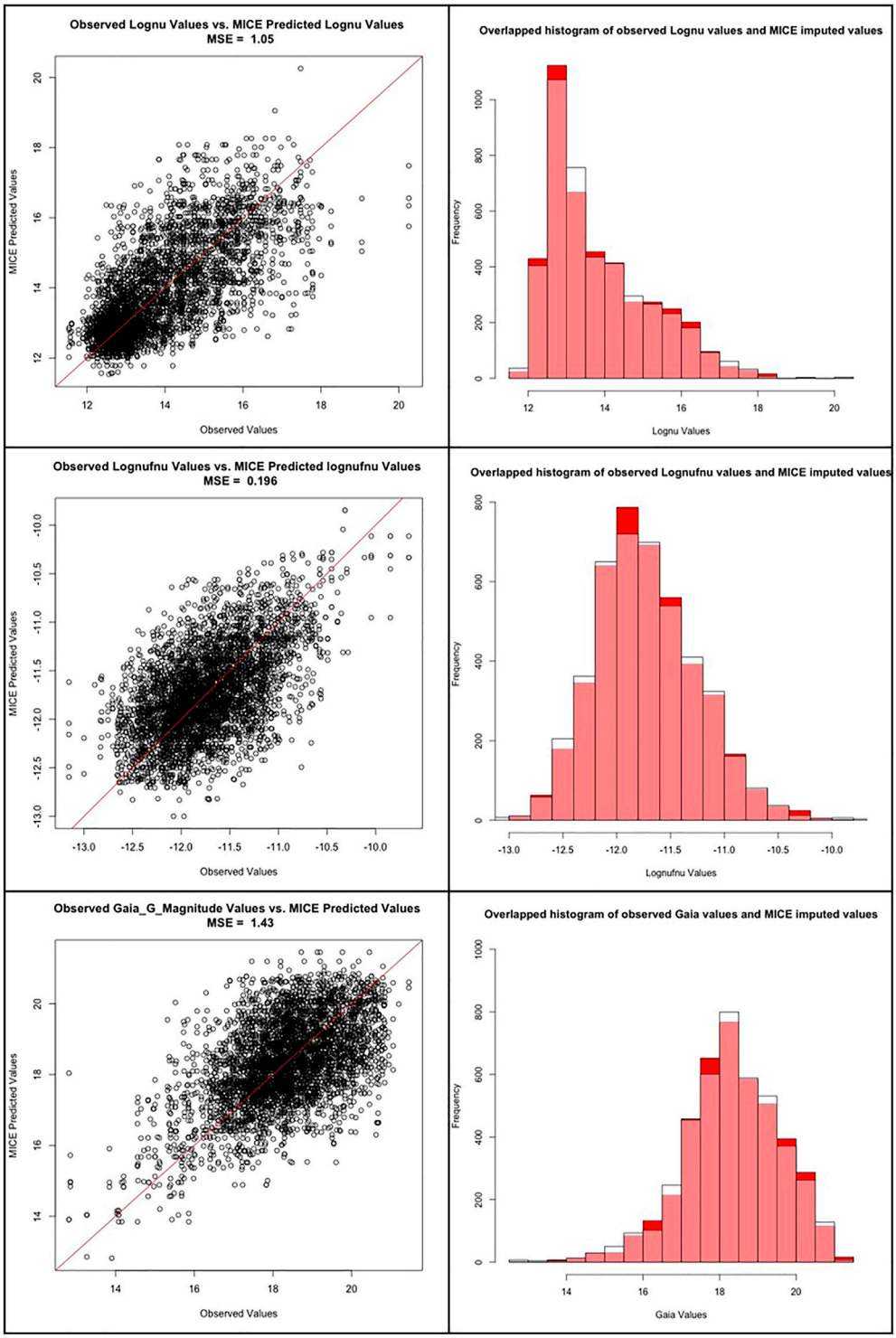

FIGURE 5 | First row: The scatter plot between the observed and predicted MICE values for the Logν predictor, followed by the overlapped histogram distributions of the same. Second row: The scatter plot between the observed and predicted MICE values for the Log$\nu f\nu$ predictor, followed by the overlapped histogram distributions of the same. Third row: The scatter plot between the observed and predicted MICE values for the *Gaia_G_Magnitude* predictor, followed by the overlapped histogram distributions of the same.

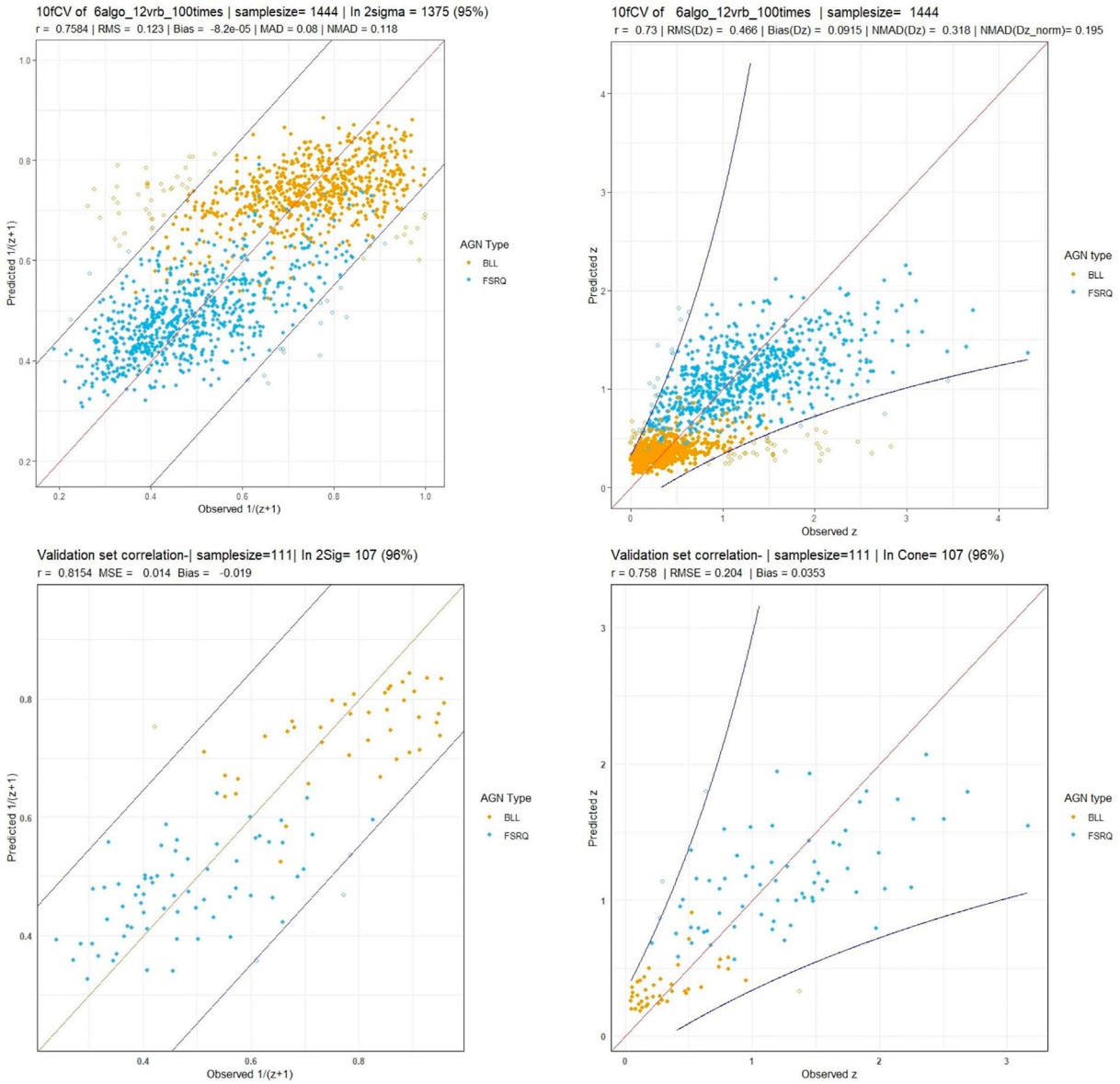

FIGURE 6 | These plots are for the O1 predictors case. Top left and right panels: The correlation plots between the observed and predicted redshift from 10fCV in the $\frac{1}{z+1}$ and linear scales, respectively. Bottom left and right panels: The validation set correlation plots in the $\frac{1}{z+1}$ and linear scales, respectively.

of overfitting. Feature creation is a technique where additional features are created from various combinations of existing properties. These combinations can be cross-multiplications, higher-order terms, or ratios. Feature creation can reveal hidden patterns in the data that ML algorithms might not be able to discern and consequently boost the performance.

In machine learning, some of the methods used by SuperLearner are linear by nature (BayesGLM, Lasso, elastic-net). Adding quadratic and multiplicative terms allows us to model some types of non-linear relationships. Interactions among variables are very important and can boost the prediction when used. The phrase "interaction among the variables" means the influence of one variable on the other; however, not in an additive way, but rather in a multiplicative way. In our feature engineering procedure, we build these interactions by cross-products and squares of the initial variables. It is common that adding O2 predictors aids results since they may contain information not available in the O1 predictors.

In this study, we create 66 new features, which, as mentioned, are the cross-products and squares of the existing features of the 4LAC catalog. We denote the existing predictors of the 4LAC catalog as Order-1 (O1) predictors and the new predictors as Order-2 (O2). Thus, we expand the set of predictors from the initial eleven O1 predictors to a combined seventy-eight O1 and O2 predictors.

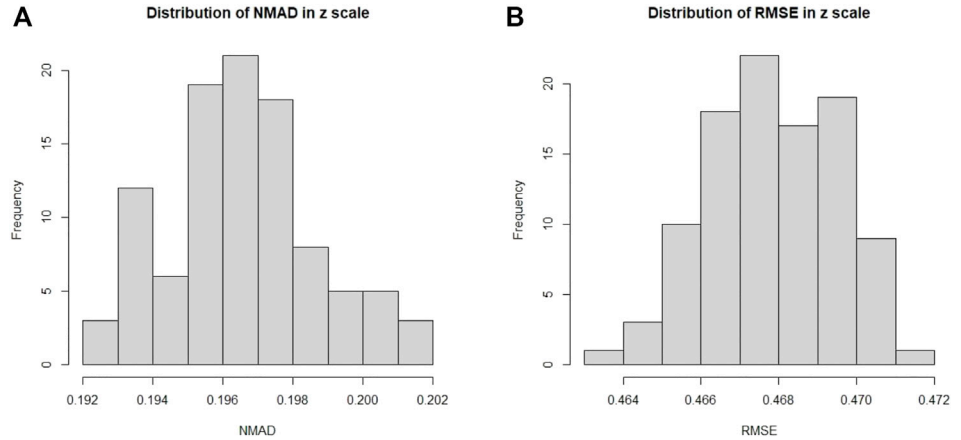

FIGURE 7 | Here we present the distribution across one hundred iterations of 10fCV of the σ_{NMAD} and RMSE for the O1 case. **(A)**: Distribution of σ_{NMAD} in linear scale. Note that σ_{NMAD} is denoted as NMAD in the plots. **(B)**: Distribution of RMSE in linear scale.

For features selection, LASSO Tibshirani (1996) is used. It works by constraining the ℓ^1 norm of the coefficient vector to be less than or equal to a tuning parameter λ while fitting a linear model to the data. The predictors that LASSO chooses have a non-zero coefficient for the largest λ value with the property that the corresponding prediction error is within one standard deviation of the minimum prediction error Friedman et al. (2010); Birnbaum (1962); Hastie and Tibshirani. (1987), Hastie and Tibshirani. (1990); Friedman et al. (2010). This study performs LASSO feature selection on a fold-by-fold basis during external 10fCV. Optimal features are picked using LASSO for nine of the ten folds, and the predictions on the 10th fold are performed using these selected features. This step is iterated such that for every combination of nine folds, an independent set of features is picked. This usage of LASSO is in contrast to Dainotti et al. (2021), where the best features are picked for the entire training set. Our updated technique ensures that during the 10fCV, LASSO only picks the best predictors based on the training data, and the test set does not affect the models. This feature selection method is applied to both the O1 and O2 predictor sets.

4 RESULTS

The quality of the MICE imputations depends on the information density of the entire data set. Hence, to ensure the best possible imputations we use all 1897 AGNs which remain after the removal of outliers and non-BLL and non-FSRQ AGNs. The pattern of the missing entries in our data set is shown in **Figure 3**, and they are present in only three predictors, namely, Logν, Log$\nu f\nu$, and *Gaia_G_Magnitude* (see **Sec. 2**). There are 237 AGNs which have missing values in both Logν and Log$\nu f\nu$, and 343 AGNs have a missing value in *Gaia_G_Magnitude*. MICE is used to fill the missing values of these AGNs. In **Figure 4** we show the distributions of Logν, Log$\nu f\nu$, and *Gaia_G_Magnitude* with and without MICE. The quality of the MICE imputations can be evaluated in part by comparing the original distribution of a variable and its distribution with imputations. If the imputations alter the distribution, the results cannot be trusted and would require additional precautions or measures to deal with the missing values. However, as can be discerned from the plots (**Figure 4**), the MICE imputations are indeed following the underlying distribution for the three predictors, and hence we confidently incorporate them into our analysis. We impute 465 data points, 24% of our data set, resulting in a training sample of 1,444 AGNs and a generalization sample of 453 AGNs. The two sets are detailed in **Table. 1**.

4.1 Multivariate Imputation by Chained Equations Reliability Analysis

In the work by Luken et al. (2021), they present an extensive analysis of the reliability of MICE imputations. However, since they use a different dataset than ours, a similar investigation regarding the performance of MICE is essential. Thus, we take 1,432 AGNs from our catalog with no missing entries and randomly dropped 20% of the entries from each of the three predictors which have missing entries, namely: Logν, Log$\nu f\nu$, and *Gaia_G_Magnitude*. We then impute these dropped entries using MICE, as described in **Section 3.2**. This process is repeated fifteen times, and each time a different set of random entries are dropped. Furthermore, as we can see in **Figure 5**, the observed vs predicted values for Logν, Log$\nu f\nu$ and *Gaia_G_Magnitude* are concentrated about the $y = x$ line, with little variance. The mean squared error (MSE), defined as the, of the observed values vs the MICE imputed values for Logν, Log$\nu f\nu$, and *Gaia_G_Magnitude* were 1.05, 0.196, and 1.43, respectively. Thus, the MSEs are all small, which provides evidence that the MICE imputed effectively. Note that if MICE imputes effectively, then the imputed values and observed values should come from the same distribution for each of the three variables. To check this, we performed a Kolmogorov-Smirnov (KS) test on the observed vs MICE imputed values for each of the three

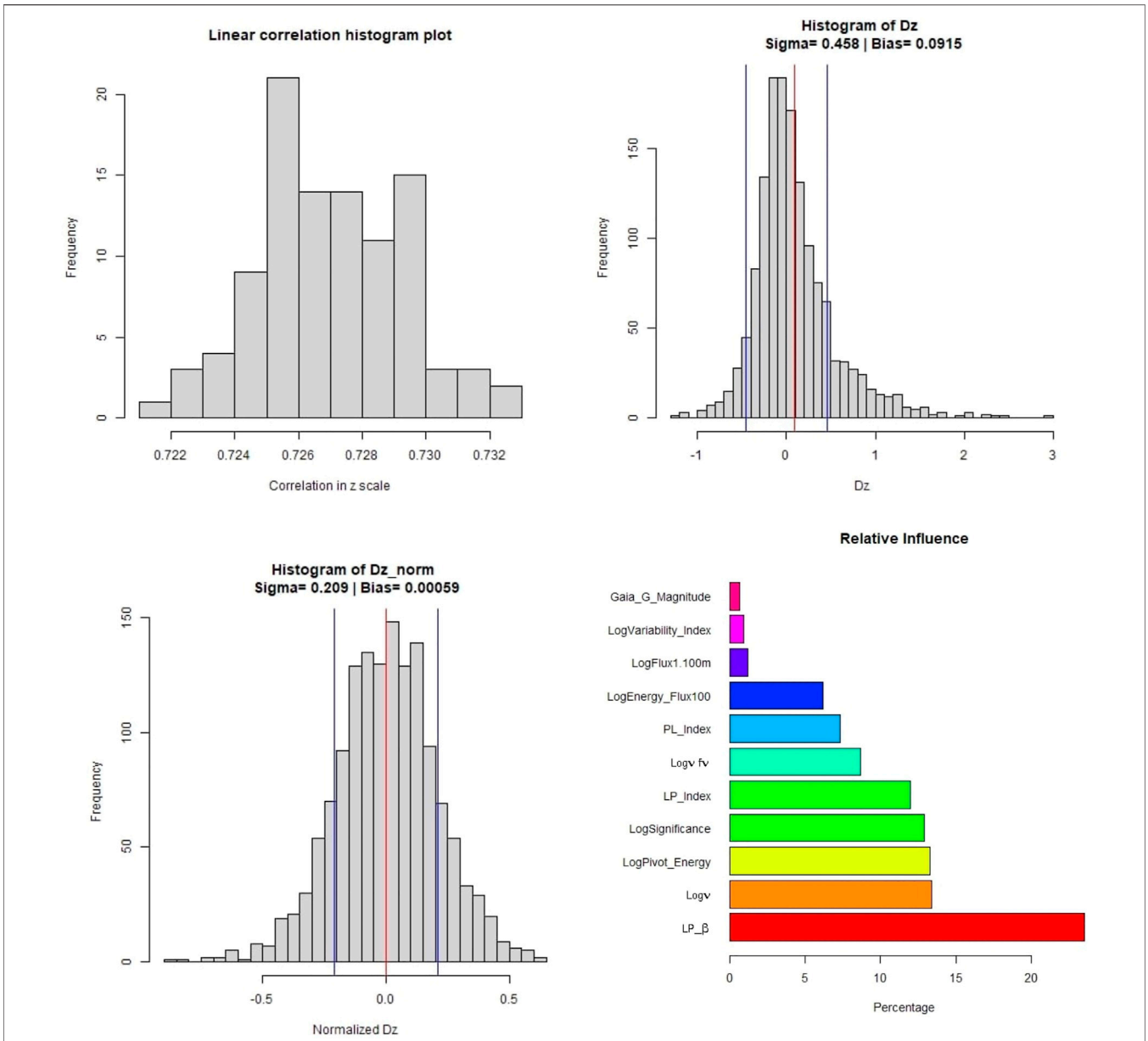

FIGURE 8 | These plots are for the O1 predictor case. Top left panel: Distribution of the correlations in linear scale from the 100 iterations of 10fCV. Top right panel: Distribution of Δz (Dz in the plots) with average bias (red) and sigma lines (blue). Bottom left panel: Distribution of the Δz_{norm} (Normalized Dz in the plots) with the average bias (red) and sigma values (blue). Bottom right panel: Relative influence of the O1 predictors. The suffix of *Sqr* implies the square of the respective predictor.

variables with missing entries. The *p*-values of the KS test for Logν, Log$\nu f\nu$, and *Gaia_G_Magnitude* were 0.744, 0.5815, and 0.6539, respectively. Since each of these *p*-values is above 0.05, we cannot reject the null-hypothesis; namely, we conclude that the observed values and the MICE imputed values come from same distributions for any of the three variables. As shown in **Figure 5**, the overlapped histogram of the observed vs MICE imputed values for Logν, Log$\nu f\nu$ and *Gaia_G_Magnitude* are each very similar, which reinforces the findings of the KS test - namely, that they are from the same distribution. This provides additional proof for the accuracy, and reliability of the MICE imputations.

4.2 With O1 Variables

The O1 variable set consists of 12 predictors, including the categorical variable *LabelNo*, which distinguishes between BLLs and FSRQs. LASSO chooses the best predictors from within this set for each fold in the 10fCV as explained in **Section 3.5**.

Using this feature set with the six algorithms mentioned, we obtain a correlation in the $1/(z+1)$ scale of 75.8%, a σ of 0.123, an RMSE of 0.123, and a σ_{NMAD} of 0.118. In the linear redshift scale (z scale), we obtain a correlation of 73%, an RMSE of 0.466, a σ of 0.458, a bias of 0.092, and a σ_{NMAD} of 0.318. In the normalized

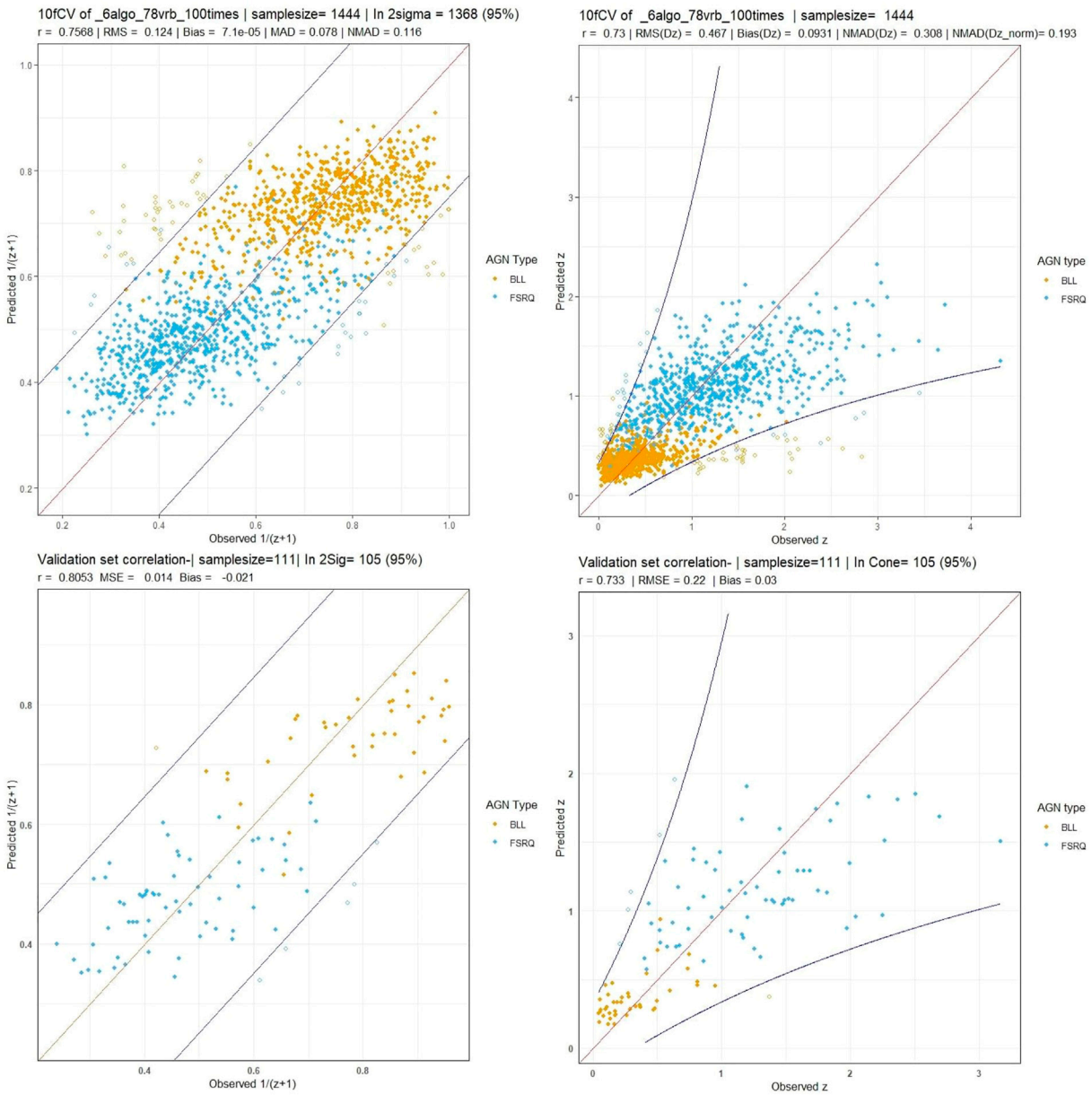

FIGURE 9 | These plots are for the O2 predictor case. Top left and right panels: Correlation plots between observed vs cross-validated redshift in the $\frac{1}{z+1}$ and linear scale, respectively. Bottom left and right panels: The validation set correlation plots in the $\frac{1}{z+1}$ and linear scale, respectively.

scale ($\Delta z_{\rm norm}$), the RMSE obtained is 0.209, bias is 6×10^{-3}, and σ_{NMAD} is equal to 0.195. The correlation plots are shown in **Figure 6**, with the left panel showing the correlation in the $1/(z + 1)$ scale and the right panel showing the correlation in the z scale. We obtain a low 5% catastrophic outlier percentage in this scenario. The lines in blue depict the 2σ curves for each plot, where the σ is calculated in the $1/(z + 1)$ scale.

In **Figure 7**, we present the distributions of σ_{NMAD} and RMSE across the one hundred iterations. Note that σ_{NMAD} is written as NMAD in the plots for brevity.

In **Figure 8**, we present the distributions of various parameters and the normalized relative influence plot of the 11 predictors - *LabelNo* is excluded, as its a categorical variable. The top left panel shows the variation in the linear correlation obtained from the one hundred iterations. The top right panel shows the distribution of Δz along with the σ (blue vertical line) and bias (red vertical line) values. The bottom left panel shows the distribution of the $\Delta z_{\rm norm}$ along with the bias and σ presented similarly. Finally, the barplot in the bottom right panel shows the relative influence of the 11 predictors used. *LP_β* has the highest

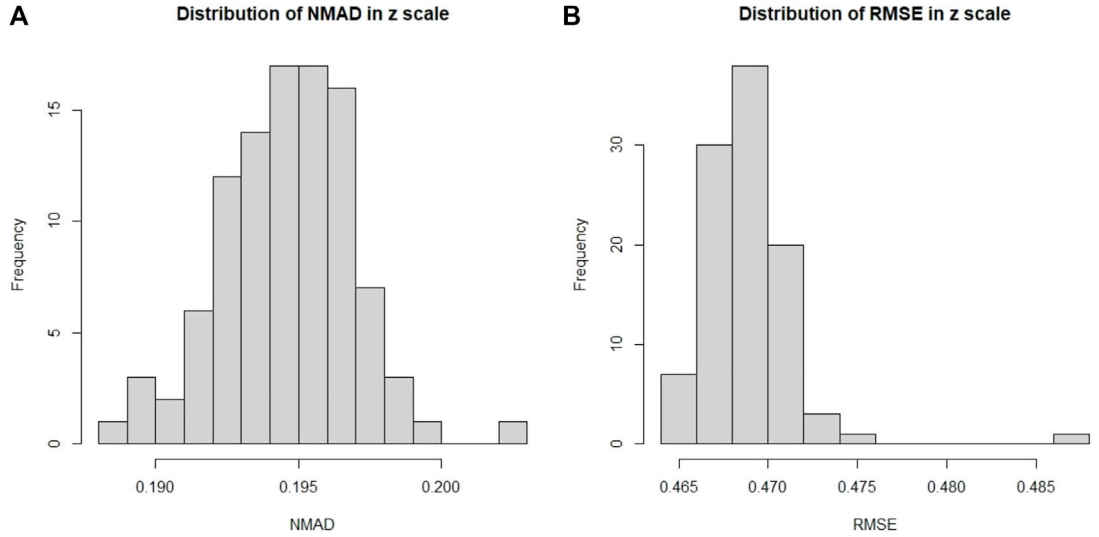

FIGURE 10 | Here we present the distribution of the σ_{NMAD} and RMSE for the O2 case. **(A)**: Distribution of σ_{NMAD} in linear scale. Note that σ_{NMAD} is denoted as NMAD in the plots. **(B)**: Distribution of RMSE in linear scale.

influence, followed by Logν, *LogPivot_Energy*, and *LogSignificance*. Surprisingly, *Gaia_G_Magnitude* has the least influence at ≈ 1%, in contrast to Dainotti et al. (2021), where we found it to be quite significant at ≈ 11% influence. The difference we obtain from this analysis and the previous one of Dainotti et al. (2021) lies in the data set and that MICE had not been used.

4.3 With O2 Variables

The O2 variables, 78 in total, are made from cross-products of the O1 variables. As in the O1 case, LASSO feature selection is performed on a fold-by-fold basis, after which the SuperLearner ensemble with the six algorithms previously mentioned is trained and makes predictions. The cross-validation and validation correlation plots are presented in **Figure 9**.

As shown in the previous section, we have correlation plots in the $1/(z+1)$ scale and the z scale. In the $1/(z+1)$ scale, we get a correlation of 75.6%, RMSE of 0.124, and σ_{NMAD} of 0.116. In the z scale, we obtain a correlation of 73%, RMSE of 0.467, and σ_{NMAD} of 0.308. We obtain the statistical parameters for Δz: an RMSE of 0.467, a σ of 0.458, a bias of 0.093, and a σ_{NMAD} of 0.308. For $\Delta z_{\rm norm}$, we obtain an RMSE of 0.21, a bias of 7×10^{-4}, and a σ_{NMAD} of 0.193. We have a similar catastrophic outlier percentage (5%) as the O1 variable case, although the number of AGNs predicted outside the 2σ cone is seven AGNs more. This discrepancy can be attributed to the randomness inherent in our calculations and additional noise introduced by the O2 predictors.

In **Figure 10**, we show the distributions of σ_{NMAD} and RMSE. Note that there is an outlier during the analysis, which leads to the unusually high RMSE value seen in the distribution.

Figure 11 shows the distribution plots for various parameters. The top left panel shows the distribution of the correlations across the one hundred iterations. There is an outlier in the distribution of the correlation plot, corresponding to the distributions of RMSE in **Figure 10**. This scenario only happens with the O2 variable set and with MICE imputations. Apart from this fluctuation, most of the correlations lie around 73%. The histogram distribution plots for Δz (top right) and $\Delta z_{\rm norm}$ (bottom left) show a similar spread as in the case of the O1 variable set. We only present predictors with influence greater than 0.5% in the relative influence plot. In this case, *PL_Index* turns out to have the highest influence, over 20%, followed by *LogSignificance*, *LogPivot_Energy*, and *LogEnergy_Flux*100.

We note that out of the 11 O1 predictors with relative influences, only 3 have less than 5% influence, and out of the 78 O2 predictors, only 4 have greater than 5% influence. Thus, the majority of the O2 predictors do not seem to provide much additional information about the redshift.

In **Tables 2** and **3** we provide a comparision between the results obtained in the two experiments we have here with MICE, and one without MICE imputations. The latter results have been taken from Narendra et al. (2022).

5 DISCUSSIONS AND CONCLUSION

In Dainotti et al. (2021), the correlation between the observed and predicted redshift achieved with a training set of 730 AGNs was 71%, with RMSE of 0.434, σ_{NMAD} ($\Delta z_{\rm norm}$) of 0.192, and a catastrophic outlier of 5%. Here, with the use of an updated 4LAC catalog, O1 predictors, and the MICE imputation technique, along with additional ML algorithms in the SuperLearner ensemble, we achieve a correlation of 73% between the observed and predicted redshift, an RMSE of 0.466, σ_{NMAD} ($\Delta z_{\rm norm}$) of 0.195 and a catastrophic outlier of 5%. Although the RMSE and σ_{NMAD} ($\Delta z_{\rm norm}$) are increasing by 7 and 1.5%, respectively, we are able to maintain the

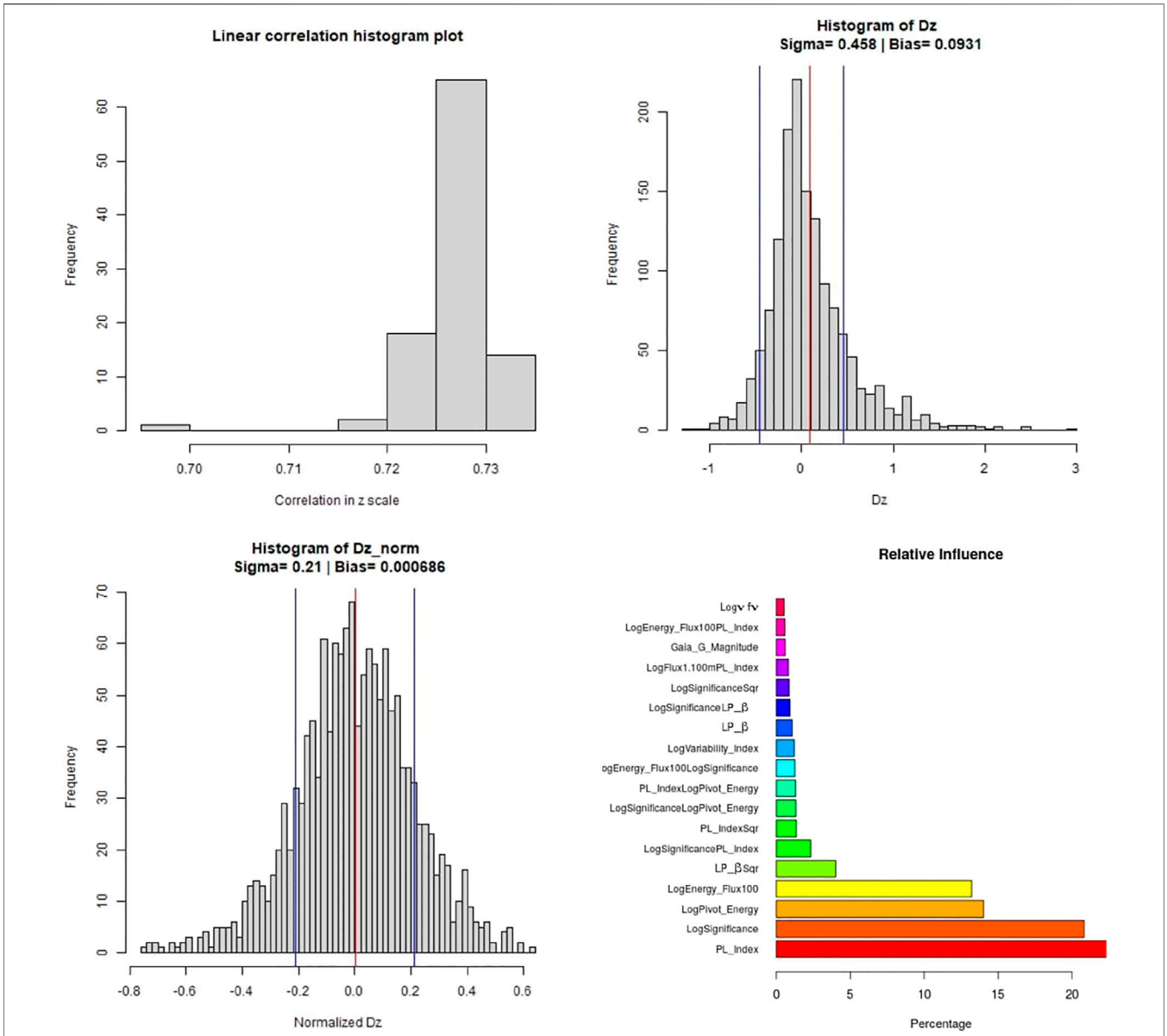

FIGURE 11 | These plots are for the O2 predictor case. Top left panel: Distribution of the correlations in linear scale from the 100 iterations of 10fCV. Top right panel: Distribution of Δz (Dz in the plots) with average bias (red) and sigma lines (blue). Bottom left panel: Distribution of the Δz_{norm} (Normalized Dz in the plots) with the average bias (red) and sigma values (blue). Bottom right panel: Relative influence of the O2 predictors, above cutoff of 0.1.

catastrophic outliers at 5%, while increasing the correlation by 3%. These results are achieved with a data sample 98% larger than the one used by Dainotti et al. (2021). Note that this achievement is not trivial, as a larger data set does not guarantee favourable results.

With the O2 predictor set, we obtain a similar correlation of 73% between the predicted and observed redshifts. However, compared to the O1 case, the RMSE goes up by 0.2%–0.467 and the σ_{NMAD} (Δz_{norm}) goes down by 1% to 0.193. The catastrophic outlier percentage is maintained at 5% in both cases.

The most influential O1 predictors in this study were *LP_β*, Log*ν*, *LogPivot_Energy*, *LogSignificance*, *LP_Index*, *PL_Index*, and *LogEnergy_Flux*, each of which has a relative influence greater than 5%. *LP_β* was also the most influential predictor in Dainotti et al. (2021), followed by *LogPivot_Energy*, *LogSignificance*, *LogEnergy_Flux*, and Log*ν*. The main difference in the relative influences of the predictors in these studies is that in the O1 case with MICE, *LP_Index* and *PL_Index* are the 5th and 7th most influential predictors, respectively, while in Dainotti et al. (2021), they were not influential.

Among the O2 predictors, *PL_Index* is the most influential, followed by *LogSignificance*, *LogPivot_Energy*, and *LogEnergy_Flux*, each of which has a relative influence

TABLE 2 | Comparision of the statistical metrics across the different experiments performed. These have been calculated for the $1/(z + 1)$ scale.

Metric	SL with O1	SL with O2	Without MICE
r	0.758	0.757	0.781
RMSE	0.123	0.124	0.119
Bias	-8.2×10^{-5}	7.1×10^{-5}	4×10^{-4}
σ_{NMAD}	0.118	0.116	0.113
σ	0.209	0.210	0.119

TABLE 3 | Comparision of the statistical metrics across the different experiments performed. These have been calculated for the z scale.

Metric	SL with O1	SL with O2	Without MICE
r	0.73	0.73	0.74
RMSE (Δz)	0.466	0.467	0.467
Bias (Δz)	0.091 5	0.093 1	0.095
σ_{NMAD} (Δz)	0.318	0.308	0.321
σ (Δz)	0.458	0.458	0.458
Bias (Δz_{norm})	5.9, ×, 10^{-4}	6.9, ×, 10^{-4}	9.6×10^{-4}
σ_{NMAD} (Δz_{norm})	0.195	0.193	0.195
σ (Δz_{norm})	0.209	0.210	0.208

greater than 5%. Note that the only O2 predictors with influence greater than 5% are those we have just listed and they are also O1 predictors. When additional variables are added it is not guaranteed that the most influential variables will be kept the same. This is true for both parametric and non-parametric models. The influence is a measure of how much your improvement in the prediction changes when you remove one variable in relation to the presence of the other variables. Thus, these measures depend on the other variables in the model and are different when O2 variables are added. We can conclude from these results that the O1 predictors contain most of the predictive information for redshift, in the case of the 4LAC catalog. Furthermore, we note that obtaining results with the O2 set takes more time than with the O1 set due to the larger list of predictors. However, in other catalogs, such O2 predictors might perform better and be an avenue worth exploring in the future.

Here, we use MICE on the O1 variables, because this allows MICE to act on three variables which present missing entries. In this way, we can control the effectiveness of MICE and the results. We agree with the referee that imputing the MICE in the cross products would imply an imputation on variables that are currently not defined and most importantly would allow more uncertainty when the cross products would involve for example two variables with missing entries. If we had used MICE in the O2 parameters we would have had a large number of imputation which would be less controllable. From these results, we can discern that the MICE imputation technique is a robust method to mitigate the issue of missing entries in a catalog while maintaining the predictive power of the data.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://fermi.gsfc.nasa.gov/ssc/data/access/lat/10yr_catalog/.

AUTHOR CONTRIBUTIONS

SG and AN were responsible for improving parts of the initial code, the generation of the final results, and the writing of the manuscript. This project was initiated by MD. The original code was developed by her and MB. It was later modified to suit the specific needs of this project. MD also participated in all the discussions regarding the project was instrumental in shaping the outcome, and determining the structure of the manuscript. MB helped with the statistical results of the work. Dr. Agniezska Pollo participated in crucial discussions regarding the direction of the research. AP helped in editing the manuscript along with discussions about the results. ER helped in editing the structure of the manuscript along with discussions about the procedures that were adopted along with the idea of using LASSO feature selection inside the cross-validation. IL helped us in acquiring and explaining the data and its features, along with corresponding labels of BLL and FSRQ. He also helped in making the manuscript clearer.

FUNDING

Funding for the DPAC is provided by national institutions, in particular the institutions participating in the Gaia MultiLateral Agreement (MLA). The Gaia mission website is https://www.cosmos.esa.int/gaia. The Gaia archive website is https://archives.esac.esa.int/gaia. This research was supported by the Polish National Science Centre grant UMO-2018/30/M/ST9/00 757 and by Polish Ministry of Science and Higher Education grant DIR/WK/2018/12. This research was also supported by the MNS2021 grant N17/MNS/000057 by the Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University.

ACKNOWLEDGMENTS

This work presents results from the European Space Agency (ESA) space mission, Gaia. Gaia data are being processed by the Gaia Data Processing and Analysis Consortium (DPAC). MD thanks Trevor Hastie for the interesting discussion on overfitting problems. We also thank Raymond Wayne for the initial computation and discussions about balanced sampling techniques which will be implemented in subsequent papers. We also thank Shubbham Bharadwaj for helping create four of the plots with correct fonts and symbols.

REFERENCES

Abdollahi, S., Acero, F., Ackermann, M., Ajello, M., Atwood, W. B., Axelsson, M., et al. (2020). Fermi Large Area Telescope Fourth Source Catalog. *ApJS* 247, 33. doi:10.3847/1538-4365/ab6bcb

Ackermann, M., Ajello, M., Albert, A., Atwood, W. B., Baldini, L., Ballet, J., et al. (2015). Multiwavelength Evidence for Quasi-Periodic Modulation in the Gamma-Ray Blazar PG 1553+113. *Astrophysical J. Lett.* 813, L41. doi:10.1088/2041-8205/813/2/L41

Ackermann, M., Ajello, M., Allafort, A., Baldini, L., Ballet, J., Bastieri, D., et al. (2012). GeV Observations of Star-forming Galaxies with the Fermi Large Area Telescope. *Astrophysical J.* 755, 164. doi:10.1088/0004-637X/755/2/164

Aihara, H., Prieto, C. A., An, D., Anderson, S. F., Aubourg, É., Balbinot, E., et al. (2011). Erratum: The Eighth Data Release of the sloan Digital Sky Survey: First Data from SDSS-III. *Astrophysical J. Suppl. Ser.* 193, 29. doi:10.1088/0067-0049/195/2/26

Ajello, M., Angioni, R., Axelsson, M., Ballet, J., Barbiellini, G., Bastieri, D., et al. (2020). The Fourth Catalog of Active Galactic Nuclei Detected by the Fermi Large Area Telescope. *ApJ* 892, 105. doi:10.3847/1538-4357/ab791e

Birnbaum, A. (1962). On the Foundations of Statistical Inference. *J. Am. Stat. Assoc.* 57, 269–306. doi:10.1080/01621459.1962.10480660

Breiman, L. (2001). Random Forests. *Machine Learn.* 45, 5–32. doi:10.1023/a:1010933404324

Brescia, M., Cavuoti, S., D'Abrusco, R., Longo, G., and Mercurio, A. (2013). Photometric Redshifts for Quasars in Multi-Band Surveys. *ApJ* 772, 140. doi:10.1088/0004-637x/772/2/140

Brescia, M., Salvato, M., Cavuoti, S., Ananna, T. T., Riccio, G., LaMassa, S. M., et al. (2019). Photometric Redshifts for X-ray-selected Active Galactic Nuclei in the eROSITA Era. *Monthly Notices R. Astronomical Soc.* 489, 663–680. doi:10.1093/mnras/stz2159

Cavuoti, S., Brescia, M., D'Abrusco, R., Longo, G., and Paolillo, M. (2014). Photometric Classification of Emission Line Galaxies with Machine-Learning Methods. *Monthly Notices R. Astronomical Soc.* 437, 968–975. doi:10.1093/mnras/stt1961

Chiang, J., Fichtel, C. E., Von Montigny, C., Nolan, P. L., and Petrosian, V. (1995). The Evolution of Gamma-Ray–loud Active Galactic Nuclei. *ApJ* 452, 156. doi:10.1086/176287

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Machine Learn.* 20, 273–297. doi:10.1023/a:1022627411411

Curran, S. J. (2020). QSO Photometric Redshifts from SDSS, WISE, and GALEX Colours. *Monthly Notices R. Astronomical Soc. Lett.* 493, L70–L75. doi:10.1093/mnrasl/slaa012

Dainotti, M. G., Bogdan, M., Narendra, A., Gibson, S. J., Miasojedow, B., Liodakis, I., et al. (2021). Predicting the Redshift of γ-Ray-loud AGNs Using Supervised Machine Learning. *ApJ* 920, 118. doi:10.3847/1538-4357/ac1748

D'Isanto, A., and Polsterer, K. L. (2018). Photometric Redshift Estimation via Deep Learning. Generalized and Pre-Classification-Less, Image Based, Fully Probabilistic Redshifts. *aap* 609, A111. doi:10.1051/0004-6361/201731326

Domínguez, A., Wojtak, R., Finke, J., Ajello, M., Helgason, K., Prada, F., et al. (2019). A New Measurement of the Hubble Constant and Matter Content of the Universe Using Extragalactic Background Light $γ$-ray Attenuation. *ApJ* 885, 137. doi:10.3847/1538-4357/ab4a0e

Fermi-LAT CollaborationAbdollahi, S., Ackermann, M., et al. (2018). *Science* 362, 1031.

Fotopoulou, S., and Paltani, S. (2018). CPz: Classification-Aided Photometric-Redshift Estimation. *A&A* 619, A14. doi:10.1051/0004-6361/201730763

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* 33, 1. doi:10.18637/jss.v033.i01

Friedman, J. H., and Roosen, C. B. (1995). An Introduction to Multivariate Adaptive Regression Splines. *Stat. Methods Med. Res.* 4, 197–217. doi:10.1177/096228029500400303

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely Randomized Trees. *Machine Learn.* 63, 42–63. doi:10.1007/s10994-006-6226-1

Hastie, T. J., and Tibshirani, R. J. (1990). *Generalized Additive Models*, 43. Boca Raton: CRC Press.

Hastie, T., and Tibshirani, R. (1987). Generalized Additive Models: Some Applications. *J. Am. Stat. Assoc.* 82, 371–386. doi:10.1080/01621459.1987.10478440

Hildebrandt, H., Arnouts, S., Capak, P., Moustakas, L. A., Wolf, C., Abdalla, F. B., et al. (2010). PHAT: PHoto-zAccuracy Testing. *A&A* 523, A31. doi:10.1051/0004-6361/201014885

Ho, T. K. (1995). "Random Decision Forests," in Proceedings of the Third International Conference on Document Analysis and Recognition Volume 1, ICDAR '95, Montreal, QC, Canada, August 14-16, 1995 (USA: IEEE Computer Society), 278.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *J. Comput. Graphical Stat.* 15, 651–674. doi:10.1198/106186006x133933

Ilbert, O., Capak, P., Salvato, M., Aussel, H., McCracken, H., Sanders, D. B., et al. (2008). Cosmos Photometric Redshifts with 30-Bands for 2-deg^2. *Astrophysical J.* 690, 1236–1249. doi:10.1088/0004-637X/690/2/1236

Jones, E., and Singal, J. (2017). Analysis of a Custom Support Vector Machine for Photometric Redshift Estimation and the Inclusion of Galaxy Shape Information. *A&A* 600, A113. doi:10.1051/0004-6361/201629558

Liodakis, I., Hovatta, T., Huppenkothen, D., Kiehlmann, S., Max-Moerbeck, W., and Readhead, A. C. S. (2018). Constraining the Limiting Brightness Temperature and Doppler Factors for the Largest Sample of Radio-Bright Blazars. *ApJ* 866, 137. doi:10.3847/1538-4357/aae2b7

Liodakis, I., Pavlidou, V., Hovatta, T., Max-Moerbeck, W., Pearson, T. J., Richards, J. L., et al. (2017). Bimodal Radio Variability in OVRO-40 M-Monitored Blazars. *MNRAS* 467, 4565–4576. doi:10.1093/mnras/stx432

Little, R. J., and Rubin, D. B. (2019). *Statistical Analysis with Missing Data*, 793. Hoboken: John Wiley & Sons.

Logan, C. H. A., and Fotopoulou, S. (2020). Unsupervised star, Galaxy, QSO Classification. *A&A* 633, A154. doi:10.1051/0004-6361/201936648

Luken, K. J., Padhy, R., and Wang, X. R. (2021). Missing Data Imputation for Galaxy Redshift Estimation. *arXiv:2111.13806*.

Marcotulli, L., Ajello, M., and Di Mauro, M. (2020). "The Density of Blazars above 100 MeV and the Origin of the Extragalactic Gamma-ray Background," in American Astronomical Society Meeting Abstracts (American Astronomical Society Meeting Abstracts), 405–406.235

Nakoneczny, S. J., Bilicki, M., Solarz, A., Pollo, A., Maddox, N., Spiniello, C., et al. (2019). Catalog of Quasars From the Kilo-Degree Survey Data Release 3. *aap* 624, A13. doi:10.1051/0004-6361/201834794

Narendra, A., Gibson, S. J., Dainotti, M. G., Bogdan, M., Pollo, A., Liodakis, I., et al. (2022). Predicting the Redshift of Gamma-ray Loud AGNs Using Supervised Machine Learning: Part 2. *arXiv:2201.05374*.

Pasquet-Itam, J., and Pasquet, J. (2018). Deep Learning Approach for Classifying, Detecting and Predicting Photometric Redshifts of Quasars in the Sloan Digital Sky Survey Stripe 82. *A&AAstronomy & Astrophysics* 611, A97. doi:10.1051/0004-6361/201731106

Petrosian, V. (1976). Surface Brightness and Evolution of Galaxies. *ApJ* 209, L1. doi:10.1086/182253

Polley, E. C., and Van der Laan, M. J. (2010). Super Learner in Prediction. U.C. Berkeley Division of Biostatistics Working Paper Series. Bepress. Available at: https://biostats.bepress.com/ucbbiostat/paper266

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika* 63, 581–592. doi:10.1093/biomet/63.3.581

Salvato, M., Ilbert, O., and Hoyle, B. (2019). The many Flavours of Photometric Redshifts. *Nat. Astron.* 3, 212–222. doi:10.1038/s41550-018-0478-0

Schafer, J. L., and Graham, J. W. (2002). Missing Data: Our View of the State of the Art. *Psychol. Methods* 7, 147–177. doi:10.1037/1082-989x.7.2.147

Singal, J. (2015). A Determination of the Gamma-ray Flux and Photon Spectral index Distributions of Blazars from theFermi-LAT 3LAC. *Mon. Not. R. Astron. Soc.* 454, 115–122. doi:10.1093/mnras/stv1964

Singal, J., Ko, A., and Petrosian, V. (2014). Gamma-Ray Luminosity and Photon Index Evolution of FSRQ Blazars and Contribution to the Gamma-Ray Background. *Astrophysical J.* 786, 109.

Singal, J., Ko, A., and Petrosian, V. (2013a). Flat Spectrum Radio Quasar Evolution and the Gamma-ray Background. *Proc. IAU* 9, 149–152. doi:10.1017/s1743921314003597

Singal, J., Petrosian, V., and Ajello, M. (2012). Flux and Photon Spectral Index Distributions Offermi-Lat Blazars and Contribution to the Extragalactic Gamma-ray Background. *ApJ* 753, 45. doi:10.1088/0004-637x/753/1/45

Singal, J., Petrosian, V., and Ko, A. (2013b). Cosmological Evolution of the FSRQ Gamma-ray Luminosity Function and Spectra and the Contribution to the Background Based on Fermi-LAT Observations. *AAS/High Energy Astrophysics Division#* 13, 300–307.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B (Methodological)* 58, 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

Van Buuren, S., and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* 45, 1–67. doi:10.18637/jss.v045.i03

Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. 2007, Super learner. *Stat. Appl. Genet. Mol. Biol.*, 6. doi:10.2202/1544-6115.1309

Venters, T. M., and Pavlidou, V. (2013). Probing the Intergalactic Magnetic Field with the Anisotropy of the Extragalactic Gamma-ray Background. *MNRAS* 432, 3485–3494. doi:10.1093/mnras/stt697

Wakely, S. P., and Horan, D. (2008). "TeVCat: An online catalog for Very High Energy Gamma-Ray Astronomy," in International Cosmic Ray Conference, 1341–1344.3

Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., Ressler, M. E., Cutri, R. M., Jarrett, T., et al. (2010). The Wide-field Infrared Survey Explorer (WISE): Mission Description and Initial On-orbit Performance. *Astronomical J.* 140, 1868–1881. doi:10.1088/0004-6256/140/6/1868

Yang, Q., Wu, X.-B., Fan, X., Jiang, L., McGreer, I., Green, R., et al. (2017). Quasar Photometric Redshifts and Candidate Selection: A New Algorithm Based on Optical and Mid-infrared Photometric Data. *Astronomical J.* 154, 269. doi:10.3847/1538-3881/aa943c

Zhang, K., Schlegel, D. J., Andrews, B. H., Comparat, J., Schäfer, C., Vazquez Mata, J. A., et al. (2019). Machine-learning Classifiers for Intermediate Redshift Emission-Line Galaxies. *ApJ* 883, 63. doi:10.3847/1538-4357/ab397e

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Astronomy and Space Sciences

ORIGINAL RESEARCH
published: 10 March 2022
doi: 10.3389/fspas.2022.838442

Identification of Flux Rope Orientation *via* Neural Networks

Thomas Narock[1]*, *Ayris Narock*[2,3], *Luiz F. G. Dos Santos*[4] *and Teresa Nieves-Chinchilla*[2]

[1]Center for Data, Mathematical, and Computational Sciences, Goucher College, Baltimore, MD, United States, [2]NASA Goddard Space Flight Center, Greenbelt, MD, United States, [3]ADNET Systems Inc., Bethesda, MD, United States, [4]CIRES, University of Colorado, Boulder, CO, United States

Geomagnetic disturbance forecasting is based on the identification of solar wind structures and accurate determination of their magnetic field orientation. For nowcasting activities, this is currently a tedious and manual process. Focusing on the main driver of geomagnetic disturbances, the twisted internal magnetic field of interplanetary coronal mass ejections (ICMEs), we explore a convolutional neural network's (CNN) ability to predict the embedded magnetic flux rope's orientation once it has been identified from *in situ* solar wind observations. Our work uses CNNs trained with magnetic field vectors from analytical flux rope data. The simulated flux ropes span many possible spacecraft trajectories and flux rope orientations. We train CNNs first with full duration flux ropes and then again with partial duration flux ropes. The former provides us with a baseline of how well CNNs can predict flux rope orientation while the latter provides insights into real-time forecasting by exploring how accuracy is affected by percentage of flux rope observed. The process of casting the physics problem as a machine learning problem is discussed as well as the impacts of different factors on prediction accuracy such as flux rope fluctuations and different neural network topologies. Finally, results from evaluating the trained network against observed ICMEs from Wind during 1995–2015 are presented.

Keywords: flux rope, neural network, machine learning, space weather, magnetic field

OPEN ACCESS

Edited by:
Bala Poduval,
University of New Hampshire,
United States

Reviewed by:
Stefano Markidis,
KTH Royal Institute of Technology,
Sweden
Emilia Kilpua,
University of Helsinki, Finland

***Correspondence:**
Thomas Narock
Thomas.Narock@goucher.edu

Specialty section:
This article was submitted to Space Physics, a section of the journal Frontiers in Astronomy and Space Sciences

Received: *17 December 2021*
Accepted: *01 February 2022*
Published: *10 March 2022*

Citation:
Narock T, Narock A, Dos Santos LFG and Nieves-Chinchilla T (2022) Identification of Flux Rope Orientation via Neural Networks. Front. Astron. Space Sci. 9:838442. doi: 10.3389/fspas.2022.838442

1 INTRODUCTION

Coronal mass ejections (CMEs) are one of many manifestations of our dynamic Sun. CMEs are responsible for the transport of large quantities of solar mass into the interplanetary medium at very high speeds and in various directions. CMEs are commonly referred to as interplanetary coronal mass ejections (ICMEs) after leaving the solar atmosphere and reaching the interplanetary medium. ICMEs are the main drivers of geomagnetic activity at Earth as well as at other planets and spacecraft throughout the heliosphere (Baker and Lanzerotti, 2008; Kilpua et al., 2017a). *In situ* observations of ICMEs frequently find them to have a combination of an increase in magnetic field strength, low proton plasma temperature, β_{plasma} below 1, and monotonic rotation of the magnetic field components (Burlaga, 1988). These characteristics are commonly referred to as a Magnetic Cloud (MC) (Burlaga et al., 1981; Klein and Burlaga, 1982). CME eruption theories (Vourlidas, 2014) suggest that a twisting internal magnetic signature—referred to as a flux rope—is always present. While commonly observed, not all ICMEs show the signatures of an internal structure characterized by a flux rope, perhaps resulting from changes during interplanetary evolution (Jian et al., 2006; Manchester et al., 2017). Yet, flux ropes are sufficiently prevalent that they can aid in

space weather forecasting. The observed magnetic field profile depends on a flux rope's orientation and where the spacecraft traverses the structure. The latitudinal and longitudinal deflections of CMEs happen in the lower corona and are not expected to change greatly throughout the interplanetary medium. If flux rope orientation and the spacecraft's crossing trajectory can be determined early enough, this can lead to advanced forecasting as the remaining portion of the flux rope's magnetic field structure can be inferred from physics-based models. The flux rope's internal magnetic field structure is prone to couple with Earth's upper magnetosphere triggering magnetic reconnection processes and allowing the injection of solar magnetic energy into the magnetospheric system. Orientation determines the magnetic field profile observed at Earth and, thus, the geo-effectiveness of the flux rope making early determination of a flux rope's orientation a vital requisite for space weather forecasting. A major challenge to developing such a forecasting system is that information about the internal magnetic structure of ICMEs is often limited to 1D observations of a single spacecraft crossing the structure. This leaves a considerable amount of uncertainty about the three-dimensional structure of the ICME.

Various physics-based flux rope models exist [for example, Lepping et al. (1990) and Nieves-Chinchilla et al. (2019)] that can be used to reconstruct the internal ICME magnetic configuration and provide information on orientation, geometry, and other magnetic parameters such as the central magnetic field. Recent *in situ* observations (Kilpua et al., 2017b; Nieves-Chinchilla et al., 2018; Nieves-Chinchilla et al., 2019; Rodríguez-García et al., 2021), and references therein] are continuing to complement earlier studies (Gosling et al., 1973; Burlaga et al., 1981; Klein and Burlaga, 1982) and enhance our understanding of ICMEs, MCs, and flux ropes. Meanwhile, an increase of space- and ground-based data availability has led to more interest in applications of machine learning within the space weather community [see (Camporeale, 2019), and references therein]. Nguyen et al. (2018) have explored machine learning techniques for automated identification of ICMEs and dos Santos et al. (2020) used a deep neural network to create a binary classifier for flux ropes in the solar wind, determining whether a flux rope was or was not present in a given interval. Recently, Reiss et al. (2021) use machine learning to predict the minimum Bz value as a magnetic cloud was sweeping past a spacecraft.

We aim to assess a neural network's ability to predict a flux rope's orientation after an ICME is identified. This work is an attempt to understand if a neural network can predict a flux rope's orientation having only seen a portion of the event. If the full magnetic field profile of the flux rope can reliably be reconstructed when the spacecraft is only partially through the flux rope this can provide advanced warning of impending geomagnetic disturbance. Yet, as machine learning is relatively new to space weather, the accuracy of these forecasts, and more generally, which neural network topologies to utilize, are unclear. We begin with a set of exploratory experiments to quantify the capabilities of neural networks in this regard. The results of these experiments then serve as a baseline to begin exploring forecasting.

Here, we extend the binary classifier work of dos Santos et al. (2020) and explore a neural network's ability to predict the orientation, impact parameter, and chirality of an already identified flux rope. We extend the capabilities presented in Reiss et al. (2021) by reconstructing the entire three dimensional magnetic field profile. The neural network is trained using simulated magnetic field measurements over a range of spacecraft trajectories and flux rope orientations. Moreover, we report on the prediction accuracy of the neural network as a function of percentage of flux rope observed. To connect this proof of concept to its potential for real-world use, we also present results from evaluating the neural network on flux ropes observed by the Wind spacecraft. In performing these experiments, we highlight the multiple ways in which this space weather forecasting problem can be cast as a machine learning application and the implications those choices have on prediction accuracy.

In **Section 2** we present our methodology. We describe the flux rope analytical model and the generation of our synthetic data set. **Section 2** also details our neural network designs and training process. **Section 3** presents our results first from the full duration synthetic flux ropes, then from partial duration flux ropes, and ultimately from application to flux ropes observed from the Wind spacecraft. We present a discussion of these results in section 4 along with concluding remarks.

2 METHODOLOGY

The task of predicting a flux rope's key defining parameters from magnetic field measurements can be cast as a supervised machine learning problem. This is an approach in which the goal is to learn a function that maps an input to an output based on numerous input-output pairs. There are currently not enough *in situ* observed flux ropes (inputs) with known key parameters (outputs) to train a neural network. Instead, we choose to use a physics-based flux rope model to produce a synthetic training dataset.

2.1 Synthetic Data

The circular-cylindrical flux rope model of Nieves-Chinchilla et al. (2016) (N-C model) is used to simulate the magnetic field signature of flux ropes at numerous orientations and spacecraft trajectories. The N-C model takes as input the following parameters:

H, Chirality of the flux rope; right-handedness is designated with 1, left-handedness with -1

Y_0, Impact parameter; The perpendicular distance from the center of the flux rope to the crossing of the spacecraft expressed as a percentage of the flux rope's radius

ϕ, Longitude orientation angle of the flux rope

θ, Latitude orientation angle of the flux rope

R, Radius of flux rope

V_{sw}, Bulk velocity of the solar wind

C_{10}, A measure of the force free structure. A value of 1 indicates a force free flux rope

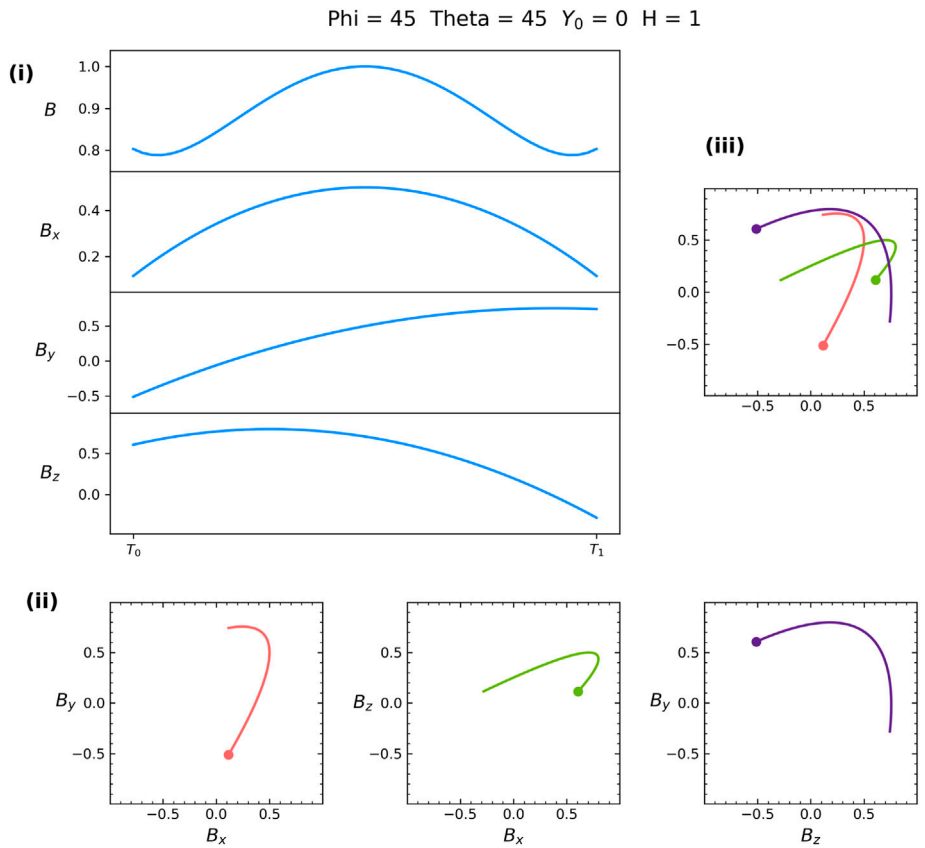

FIGURE 1 | A synthetic flux rope example generated using $\phi = 45$, $\theta = 45$, $Y_0 = 0$, and $H = +1$. (i) The total magnetic field and the magnetic field components. (ii) Three hodograms of the magnetic field components. The dot represents the starting point of the simulated flux rope crossing. Flux rope classification with 2D CNNs require three images, which can (iii) be combined for a single set of convolutions or (ii) have convolutions applied separately.

The output of the N-C model is the magnetic field profile (Bx, By, Bz components) that would be observed for spacecraft traversing a flux rope with the given input parameters. An illustration of this is shown in **Figure 1** where panel (i) shows the N-C model output visualized as a time series and panel (ii) depicts the same output as hodograms. All flux ropes were simulated using a solar wind speed (V_{sw}) of 450 km/s, a radius (R) of $0.07AU$, and with poloidal normalization. The C_{10} parameter was held constant at 1, which imposes a force free structure. The model was run for all combinations of longitude (ϕ) $\in [5^{\circ}, 355^{\circ}]$, latitude ($\theta$) $\in [-85^{\circ}, 85^{\circ}]$, and impact parameter ($Y_0$) $\in [0\%, 95\%]$. This is done in 5° and 5% increments and with both chirality options, $H \in \{-1, 1\}$. We exclude combinations involving $\phi = 180^{\circ}$ as the model is not always defined in this instance. This results in 98,000 combinations.

The fixed bulk velocity of 450 km/s and fixed radius of $0.07AU$ describe a "typical" flux rope observed at Earth based on fittings in Nieves-Chinchilla et al. (2019). Magnetic field profiles of this "typical" flux rope have been shown (dos Santos et al., 2020) to scale with changes in speed and size. In other words, magnetic field profiles are very similar when orientation is held constant and speed and radius are varied. The only variation in the profiles is duration, which is not a factor for us as all flux ropes are interpolated to 50 points. This relationship allows us to only simulate a subset of all possible speeds and sizes drastically reducing the training data set size and minimizing training time.

The output from each of these 98, 000 combinations is then used to generate 10 exemplars of this event in different percentages of completion - from 10 to 100% in steps of 10%. For example, first a 50-point trace through a flux rope defined by the parameter combination is generated (100% completion, **Figure 1(i)**). The first 5 points are interpolated to 50 points to create the 10% completion exemplar. Similarly the first 10 points are used to create the 20% exemplar, the first 15 points for the 30% exemplar, etc. The final dataset contains 980, 000 exemplars - a mixture of full duration and partially observed events. These simulated partial flux ropes are useful to understand how much of the flux rope needs to be observed before reliable autonomous predictions can be made. The ability to predict in the absence of the complete flux rope is very desirable in the context of space weather forecasting.

2.2 Convolutional Neural Networks

Simply put, a convolution is the application of a filter to an input that results in an activation. Repeatedly applying the same filter to

an input–for example, by sliding a small dimensional filter across an image - results in a map of activations called a feature map. The feature map then indicates the locations and strength of a detected feature in the input. Convolutions are the major building blocks of convolutional neural networks (CNNs) (LeCun and Bengio, 1995), which use a training dataset to learn a set of highly specific filters from the input that lead to the most accurate output predictions. The innovation of the CNN is in not having to handcraft the filters, but rather automatically learning the optimal set of filters during the training process.

The CNN is the basis of the neural network architectures explored in this work. The training phase consists of showing the network the input-output pairs of simulated flux rope magnetic field vectors (input) and the corresponding key parameters used to create this simulated data trace (output). The key parameters represented in this training are ϕ, θ, Y_0, and H. From repeated exposure to input-output pairs the network learns the filters that lead to the most optimal predictions. These neural networks require all inputs to be of the same size, which does not pose a problem when working with synthetic data. *In situ* observations from spacecraft, however, reveal a diverse set of events ranging from a few hours to multiple days. These need to be thoughtfully processed for use as input to the CNN. One could average or interpolate *in situ* events to ensure all input magnetic field time series are of the same length. Alternatively, dos Santos et al. (2020) showed an innovative technique of representing flux ropes as hodograms. Flux ropes of any duration can be cast as a set of three consistently sized images (see **Figure 1(ii)**), which can then serve as input to a CNN. This technique also leverages a wide swath of existing literature in the computer vision field (particularly in the area of handwritten digit classification) that can be helpful in fine-tuning the CNN architecture. Over the next several sections, we present a series of experiments evaluating multiple CNN architectures. Specifically, we compare the predictions from convolutions applied directly to magnetic field time series to predictions made from convolutions applied to hodograms of those magnetic field time series. We do so under two scenarios. First, we develop a baseline for a CNN's capability to predict flux rope orientation by training the architectures with only the 98,000 exemplars of full duration flux ropes. We separately train another copy of each of the three aforementioned architectures with the complete set of 980,000 full and partial duration flux ropes to assess CNN usage in a time-predictive capacity.

2.2.1 CNN Architectures

Representing flux ropes as hodograms was inspired by work in handwritten digit classification (dos Santos et al., 2020). Yet, flux ropes provide a more challenging version of this computer vision problem. The input for handwritten digit classification is always a single image; however, flux ropes require a set of three images (hodograms) to capture the entirety of their magnetic field configurations. An initial research question is then how to feed three hodograms as input to a CNN. In the approach chosen for our first architecture, we stack the images (**Figure 1(iii)**) and do a single two-dimensional convolution across the resulting tensor. In our second tested architecture, we apply two-dimensional convolutions to each of the three hodograms separately (**Figure 1(ii)**) and then concatenate the resulting feature maps.

The architecture schematic for the stacked approach is shown in **Figure 2(i)**. An input layer of dimension [100, 100, 3] passes through two rounds of 2D Convolution with a 3×3 kernel size. The resulting layer of dimension [100, 100, 64] undergoes a 2×2 Max Pooling to transform to dimensions [50, 50, 64]. This layer is then Flattened and Fully Connected to each of four output layers. This 2D CNN with one input ends up with 979,398 trainable parameters.

The architecture for applying two-dimensional convolutions to each of the three hodograms separately and then concatenating the resulting feature maps is shown in **Figure 2(ii)**. Each prong of the initial part of this network involves the same transformations as in the previously described network, with the exception that each of the three input layers is of dimension [100, 100, 1]. Additionally, the Flattened layers at the end of these individual pipelines are then concatenated before being Fully Connected to the four output layers. This architecture has the advantage that salient features in specific hodograms can become more apparent in the feature maps. Yet, this comes at the cost of a more complex neural network. With 2,936,454 trainable parameters, this CNN has significantly more weights that need training.

Finally, we tested an architecture that did not rely on hodogram images. Instead we apply 1D convolutions directly to magnetic field time series. This approach is depicted in **Figure 2(iii)** and results in the smallest CNN with 216,518 trainable parameters. The input layer of size [1, 50, 3] has a 1D Convolution with kernel size 5 applied twice, resulting in a layer of dimension [1, 50, 64]. Max Pooling with a kernel size 2 then creates a layer of dimension [1, 25, 64] before this is flattened to a vector of size 1,600. This layer is then Fully Connected to a layer of size 128 and then to each of the four output layers.

Hyperparameters for all of these architectures were found by doing a simple grid search. Our focus was on comparison of architectures and we acknowledge there may still be room for hyperparameter optimization.

2.2.2 CNN Tuning and Training

Neural networks learn by minimizing a loss function, which typically involves some measure of difference between current predictions and expected outputs. Angles can challenge neural network predictions in that loss functions, such as mean squared error (MSE), completely miss the circular nature of angles. For example, if a flux rope's longitudinal value is 0°, then predictions of 350° and 10° are both off by 10°. Yet, MSE will miss this relation and penalizes the 350° prediction more than the 10° prediction. To combat this, we predict $(\sin(\measuredangle), \cos(\measuredangle))$ with tanh activation to enforce outputs to be in $[-1, +1]$. We then post-process the CNN's predictions with arctan to convert to degrees. This approach is applied across all three CNN architectures when predicting ϕ and θ.

A challenge also arises in that predicting the real-valued parameters ϕ, θ, and Y_0 is a regression problem while determining the binary parameter, H, is a classification problem. We address this by training four separate loss functions in each CNN. For ϕ and θ we predict the pair $(\sin(\measuredangle), \cos(\measuredangle))$ and train using the MSE loss function. Impact parameter is also trained using MSE while chirality is defined as a two class classification problem and trained using binary cross entropy.

In our first experiment, the 98,000 full duration synthetic flux ropes were randomly divided into 60% training, 20% validation,

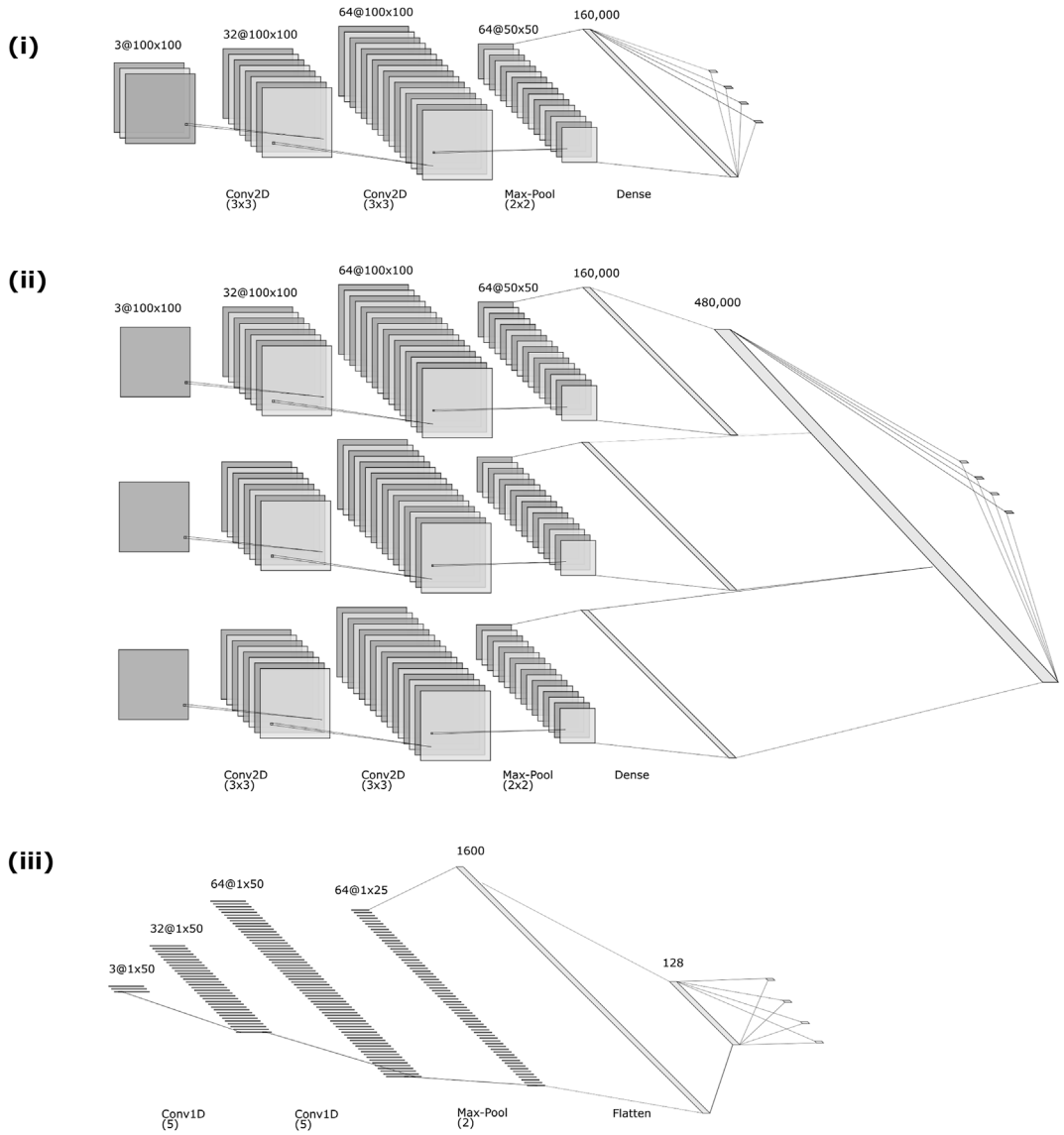

FIGURE 2 | CNN architecture schematics. (i) 2D CNN with one input which uses stacked hodograms; (ii) 2D CNN with three inputs, which performs individual convolutions over each of the three hodograms, and (iii) CNN architecture for 1D convolutions over time series.

and 20% testing sets. This resulted in 58, 800 synthetic flux ropes used for training, 19,600 used for validation, and 19, 600 used for testing. The training set was used in a supervised learning fashion with the Adam optimizer (Kingma et al., 2015) with the validation set used during the training process to avoid overfitting. All networks were set to train over 500 epochs, but the 2D CNNs had early stopping from criteria on the validation set at around 35 to 50 epochs. The 1D CNN had a training time of 12 min and both the one input and three input 2D CNNs had training times approaching 4–6 h.

The setup of the second experiment, in which we train over all full and partial flux ropes, was similar. A 60/20/20 split was used, with validation criteria used for early stopping and evaluation on the testing set. Again, the 1D CNN trained over all 500 epochs while the 2D networks reached early stopping within 50 epochs. The 1D CNN took just over 2 h to train, while the 2D CNNs completed in 6–10 h. It should be noted that all percentages of a particular flux rope configuration were included in an input batch. Also, an important consideration in this scenario is that the networks will be seeing multiple inputs that share the same output. All neural networks were constructed, trained, and tested using Python 3.8.10, Keras 2.4.3 (Chollet, 2015), TensorFlow 2.3.1 (Abadi et al., 2015), Numpy 1.18.5 (Harris et al., 2020), and Scipy 1.7.1 (Virtanen et al., 2020).

2.3 Wind Spacecraft

The final segment of this work is to evaluate the trained CNNs on flux ropes observed by the Wind spacecraft. This application of the CNNs on non-synthetic data helps us understand the limitations of the flux

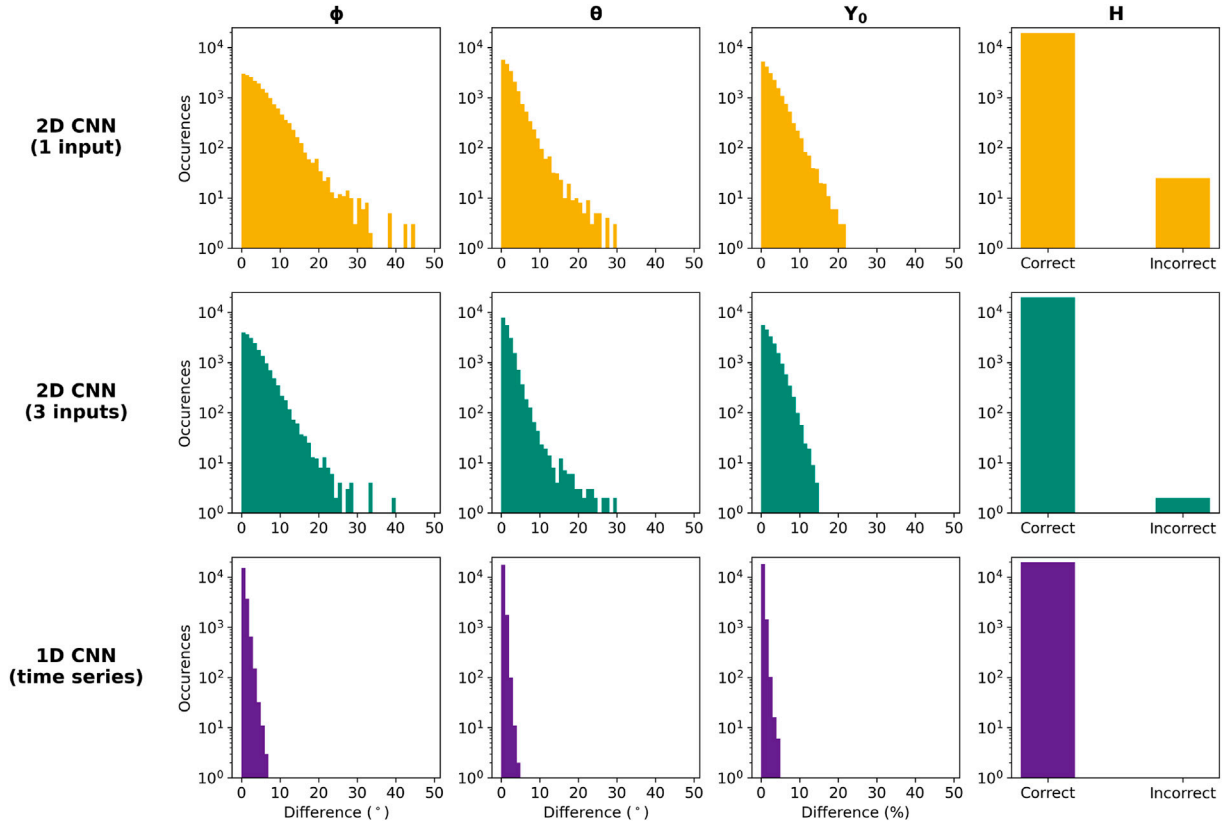

FIGURE 3 | The parameter prediction error for the synthetic test set of full duration flux rope crossings. The first three columns show histograms of the differences between predicted and modeled ϕ, θ, and Y_0 values. The last column displays the number of correct and incorrect chirality predictions. The 1D CNN predicted all angles within 10°, all Y_0 within 10% and achieved 100% accuracy for H. While the 2D CNNs perform this well in most cases, they exhibit a much wider range in error.

rope analytical model and the transition to actual space weather forecasting. Nieves-Chinchilla et al. (2018) carried out a comprehensive study of the internal magnetic field configurations of ICMEs observed by Wind at 1AU in the period 1995-2015. In this analysis, the term magnetic obstacle (MO) is adopted as a more general term than magnetic cloud in describing the magnetic structure embedded in an ICME. The authors used the Magnetic Field Instrument (MFI) (Lepping et al., 1995) and Solar Wind Experiment (SWE) (Ogilvie et al., 1995) to manually set the boundaries of the MO through visual inspection. All MO events were sorted into three broad categories based on the magnetic field rotation pattern: events without evident rotation (E), those with single magnetic field rotation (F), and those with more than one magnetic field rotation (Cx). More recently, Nieves-Chinchilla et al. (2019) presented an in-depth classification, which further classified the F types events into F-, Fr, and F+ based on the angular span of the magnetic field rotation. These events were then manually fit with the Circular-Cylindrical N-C model by a human expert. Of the events cataloged and fit, those that were classified as the Fr type tended to be the ones that could best be fit with the N-C model. Because we restricted out training set of synthetic data to flux rope cases with a $Y_0 > 0$, we also restrict our Wind test event cases to this criteria. We use this subset of 75 Wind Fr type events to evaluate our neural network predictions on actual flux rope observations. We compare the human-fit key parameters to the neural network predictions. While we have high confidence in the human expert's fit values, we acknowledge that they are not definitive. Other experts may parameterize the event slightly differently. Instead of using the human expert as ground-truth, we are interested in seeing if a neural network, trained on the same physical model that the human expert used, will arrive at similar flux rope orientations. The average correlation coefficient is used to compare human and neural network fits to the Wind magnetic field profiles.

As noted earlier, the 1D CNN is configured to input vectors of size 50 and trained on normalized synthetic data, requiring some pre-processing for use with real-event data. We begin with the 1-min resolution MFI data for each the 75 Wind events and apply a 5-point moving average smoothing followed by interpolation to 50 points evenly spaced in time.

3 RESULTS

3.1 Full Duration Synthetic Flux Ropes

Results of applying the neural networks trained on full duration flux ropes to the testing set of full duration flux ropes are shown in **Figure 3**. The ϕ, θ, and Y_0 panels display histograms of the difference between the neural network's predictions and the true values used to create the simulated instance for longitude,

TABLE 1 | Median of the parameter differences shown in **Figure 3**.

	Median difference		
CNN	**2D (1)**	**2D (3)**	**1D**
ϕ(°)	3.65	2.67	0.54
θ(°)	1.86	1.31	0.37
Y_0 (%)	2.13	1.93	0.34

latitude, and impact parameter, respectively. The *H* panel shows the number of correct and incorrect chirality predictions. Subsequent figures use the same color scheme (2D CNN with 1 input in orange, 2D CNN with 3 inputs in green, and 1D CNN in purple) for clarity. **Table 1** lists the median values of these difference distributions. The 2D CNN with a single input channel has the highest median difference across all three of the real-valued key parameters, as well as the most skewness. The 1D CNN shows the least skewness and the lowest median difference values across the parameters. The 2D CNN with three inputs falls in between, but with median difference and skewness more similar to the other 2D network than the 1D. A similar trend is seen in the *H* predictions, with the one input 2D network having the most incorrect classifications and the 1D network making no incorrect classifications. Taken together, it is evident that the 1D CNN, which is applied to the time series directly, gives more accurate predictions across all four output parameters.

While the 1D CNN gives the most accurate predictions, all three architectures give reasonably useful predictions for the vast majority of cases. The bulk of the prediction errors are less than 15° for ϕ and θ and under 10% for Y_0 for both of the 2D CNNs. **Figure 4** illustrates the prediction errors as a function of Y_0. The 2D CNNs using hodograms as input have the most significant ϕ and θ prediction errors, which occur at large Y_0. In contrast, the 1D CNN more accurately predicts ϕ and θ over the entire range of simulated Y_0. Clearly, the architecture of the neural network plays a role in prediction accuracy and leads to an important trade off. The two-dimensional networks, by using hodogram input, remove time from the training process. This makes little difference with the synthetic training data but is an advantage when working with data from time-varying, real ICME events, as the data can be used with less manipulation in pre-processing. Yet, this comes at the cost of less accurate predictions at large spacecraft impact parameters (Y_0). The trade off is that the simpler and more accurate 1D network comes with the added complexity of determining the most appropriate data transformations to fit the measured time-series to the prescribed input array dimensions of the network.

Our CNNs were each designed with four loss functions and our analysis up to this point has looked at each predicted parameter individually. We now turn our attention to evaluating the predictions as a set. To do so, we use the predicted ϕ, θ, Y_0, and *H* to reconstruct the magnetic field

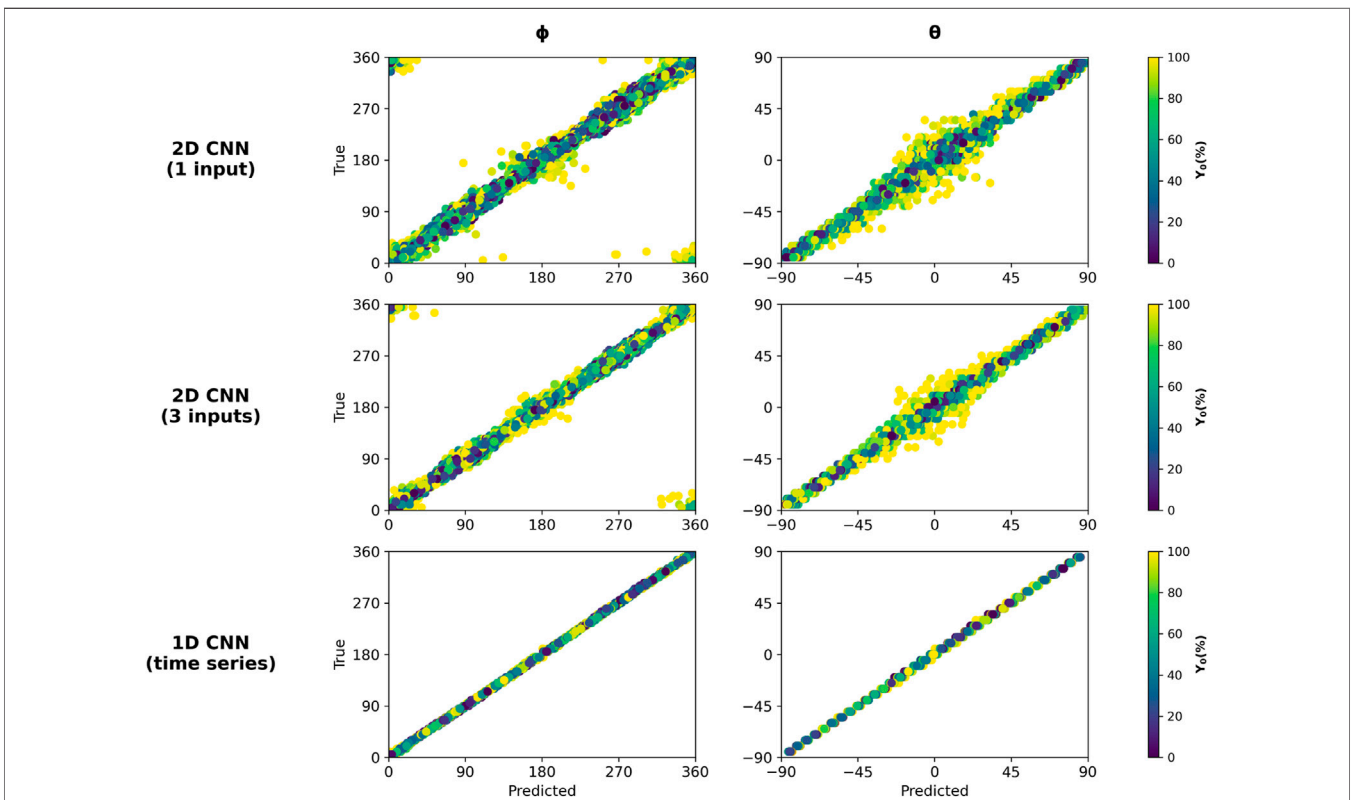

FIGURE 4 | Latitude and longitude predictions vs. true values as a function of spacecraft impact parameter when evaluated on synthetic data test set. The 1D CNN performs similarly well across the entire range of Y_0 while the 2D CNNs show a larger discrepancy in parameter predictions at high impact parameters of Y_0 > 80%.

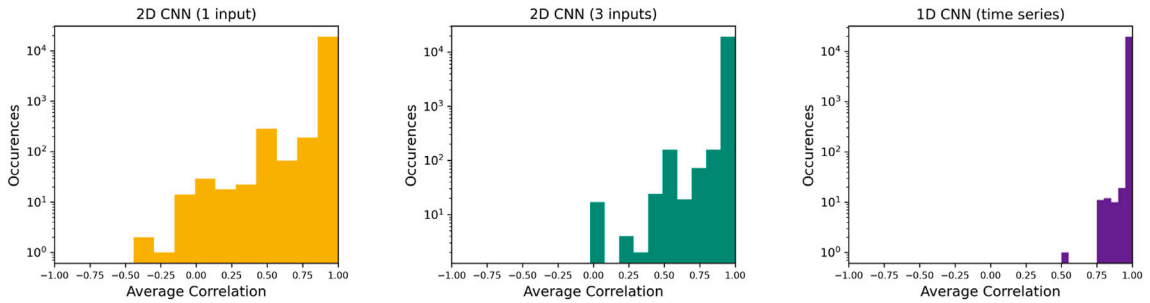

FIGURE 5 | Correlation coefficient histograms on full duration, synthetic data test set for each neural network architecture. Each set of parameters {ϕ, θ, Y_0, H} model a spacecraft's traversal of a flux rope. In these comparisons, the magnetic field trace modeled by the predicted parameters is correlated with the trace modeled by true parameters. Again, we see that all three architectures predict highly correlated results in the vast majority of cases but with the 2D CNNs exhibiting a significantly wider distribution.

time series with the analytical model and correlate it with the simulated magnetic field used as input for the CNN. For analysis, we use the average correlation coefficient, r, defined as:

$$r = \frac{r_x + r_y + r_z}{3} \tag{1}$$

where r_x is the Pearson's correlation between the simulated and reconstructed b_x components, r_y is the Pearson's correlation between the simulated and reconstructed b_y components, and r_z is the Pearson's correlation between the simulated and reconstructed b_z components.

Figure 5 shows the average correlation values for each of the three CNN architectures. While the bulk of the simulated data predicted by the 2D CNNs have high average correlation, over 0.75, there is a long tail of predictions with much lower correlation. The single input 2D CNN even makes some predictions that lead to negative correlations. As in the individual key parameters, we see the 1D CNN applied to the time series outperforming the 2D CNNs. In the case of the 1D CNN, we find only one correlation value below 0.75. Further analysis reveals that this event occurred at a simulated ϕ value of 175°. The 1D neural network predicted a ϕ value of 181°. Although the neural network predictions were fairly accurate (within 5°, 5%, and correct H), this small deviation in ϕ changed the spacecraft's trajectory through the flux rope leading to a negative correlation in the b_z component. Because the 2D CNNs are impacted by their difficulties making predictions at large spacecraft impact parameters, we see many of the poor average correlation coefficients in the 2D CNNs at large spacecraft impact parameters.

3.2 Partial Duration Synthetic Flux Ropes

In our second experiment, we retrained a second version of each of the three neural networks, this time using the full set of 980,000 full and partial duration flux ropes. Like the difference comparisons shown in **Figure 3** and **Table 1**, **Table 2** provides summary statistics of ϕ, θ, and Y_0 prediction error as a function of percentage of flux rope observed. All three models make fairly accurate predictions even when seeing just 10% of the flux rope and then continue to improve their prediction accuracy up to a point. After this point, the key parameter accuracy gets worse as higher percentages of the flux ropes are fed to the networks. The level of observation giving the lowest median errors for each CNN is highlighted in yellow, with the next lowest medians highlighted in green. Additionally, all three models were able to predict the correct H over 99% of the time at all percentages of flux rope observed.

It is worth noting that all three networks perform worse at 100% duration when trained with partial duration flux ropes as compared to these same networks trained only with full duration flux ropes. The introduction of partial flux ropes into the training produces more error (see **Table 1** and **Table 2**). We suspect this is due to multiple inputs now producing the same output. It remains for future research to conduct a more in depth analysis into how to combat this.

As with the networks trained only with full duration flux ropes, the 1D CNN gives better predictions across all parameters. We see a familiar pattern emerge in the 2D CNNs; they have difficulty predicting spacecraft impact parameter and more often predict chirality incorrectly. This in turn leads to greater inaccuracies in ϕ and θ predictions. Given that the 1D CNN out performed the 2D CNNs in both training experiments, we focus only on the 1D architecture when evaluating network performance on actual spacecraft measurements.

3.3 Application to Wind Catalog Flux Ropes

To assess the transfer-ability of this technique to real-time use, we applied the 1D CNN trained on full duration flux ropes to the 75 selected Wind events described in **Section 2.3** with the data processed in two ways. The first approach, which we label Full Resolution, is where we simply use the window smoothing before interpolating the event down to 50 points. The second approach, called Downsampled, first applies 15 min averaging before smoothing and interpolation. The idea being that the Downsampled approach would further reduce fluctuations found inside Wind flux ropes. Comparing Full Resolution and Downsampled would help us isolate the impacts of fluctuations. The difference histograms in **Figure 6** show the result of comparing the fit parameters from Nieves-Chinchilla et al. (2019) (N-C) with the neural network

TABLE 2 | Median parameter differences by percentage of flux rope observed for the neural network architectures when trained using partial duration crossings. Cells highlighted in yellow indicate the lowest error for each (CNN, parameter) pair and cells highlighted in green, the next two lowest errors. The overall performance of the 1D CNN continues to be significantly better than the 2D CNNs. The 2D CNNs make their best predictions when seeing less of the flux rope crossing.

	2D (1)			2D (3)			1D		
% Observed	ϕ	θ	Y_0	ϕ	θ	Y_0	ϕ	θ	Y_0
10	7.67°	5.85°	7.33%	5.49°	3.92°	5.61%	2.10°	1.23°	1.28%
20	6.62°	4.89°	6.62%	4.94°	3.47°	5.15%	1.58°	1.02°	1.08%
30	6.25°	4.58°	6.20%	4.70°	3.28°	4.98%	1.37°	0.90°	0.94%
40	6.29°	4.38°	5.99%	4.60°	3.24°	4.96%	1.23°	0.84°	0.84%
50	5.96°	4.28°	5.87%	4.61°	3.24°	4.95%	1.14°	0.81°	0.79%
60	6.02°	4.22°	5.75%	4.70°	3.31°	4.96%	1.11°	0.80°	0.76%
70	6.17°	4.23°	5.79%	4.74°	3.33°	5.06%	1.04°	0.76°	0.72%
80	6.13°	4.36°	5.89%	4.83°	3.37°	5.10%	1.01°	0.75°	0.71%
90	6.38°	4.45°	6.13%	4.93°	3.41°	5.25%	1.04°	0.76°	0.75%
100	7.07°	4.81°	6.52%	5.17°	3.61°	5.51%	1.10°	0.79°	0.83%

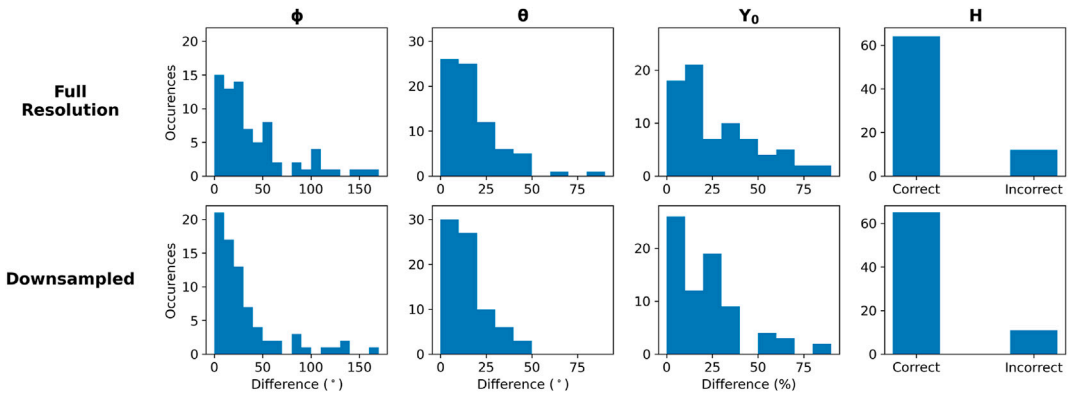

FIGURE 6 | The 1D CNN parameter prediction error for the Wind event test set of full duration flux rope crossings. Two sets of predictions were generated: One from processing 1 min Wind MFI measurements and the second from processing the Wind MFI measurements down-sampled to the 15 min averages. The human-fit parameter values from the published ICME catalog were compared against neural network predictions. The overall error magnitude is greater than when tested on synthetic input but shows the same trend. Predictions improve when the down-sampled input is used.

predicted parameters. We note that our neural network was trained with force free synthetic flux ropes (C_{10} parameter equal to 1). The N-C fittings allowed for deviations from a force free flux rope. This difference likely played a role in the discrepancies between neural network predictions and the human expert's fits.

Most ϕ predictions were within 50° of the hand-fit value but the maximum error was over 150°. The θ errors tend to be less than 25° with a maximum around 80°. Most Y_0 predictions are within 30% of the comparison values with a maximum near 80%. Across all these real-valued key parameters, the predictions made from the Downsampled input display a less skewed error distribution with a higher percentage of the predictions having relatively small error. The network produced similar results predicting chriality (H) when fed with Full Resolution and Downsampled input.

We extend this comparison with analysis of average correlation coefficient. We display correlation between the interpolated Wind observations and the magnetic field vectors generated using the N-C fit parameters as well as the correlation between interpolated Wind observations and magnetic field vectors generated from neural network predictions. **Figure 7** column 1 shows the distribution of average correlation between the human-fit model and Wind data. Column 2 is the distribution of average correlation between the CNN fit and Wind data. Displayed in column 3 is the difference histogram showing the neural network correlation minus the hand-fit correlation for each of the Wind flux rope events. Positive values indicate the neural network produced a statistically more reliable fit. Panel 7(i) shows these distributions for all of the 75 events. Panel 7(ii) shows the distributions when we consider only the events in which the CNN predicted the same chirality as the human-fit. We see good agreement in average correlation coefficients when the predictions are used to reconstruct the magnetic field time series. The shape of the distributions are similar to those from the comparison with human expert fits and an event by event comparison with human expert fits leads to a difference histogram nearly centered at zero. When we look at the Downsampled neural network predictions with

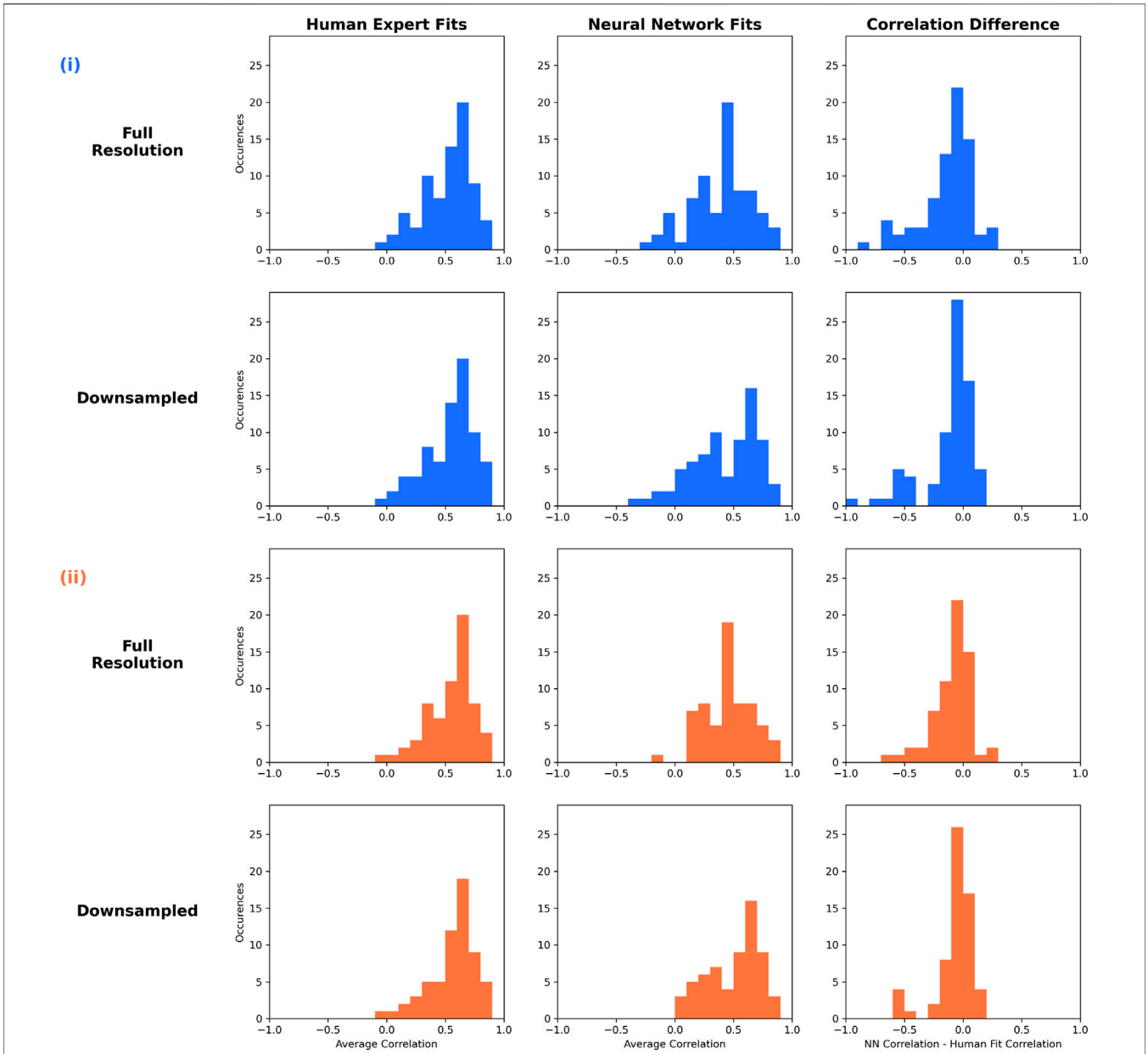

FIGURE 7 | Correlation distributions and comparisons for the Wind event test set of full duration flux rope crossings. Column 1 shows the correlation score between the human-fit parameters and the Wind measurements. Column 2 shows the correlation between the 1D CNN prediction and the Wind measurements. The third column displays the difference between the human-fit correlations and the CNN prediction correlations. (i) Includes all 75 Wind test cases. (ii) Includes cases where the CNN Y_0 prediction matched human expert's chirality only. When the CNN predicts Y_0 correctly, the correlation to Wind data is similar to that of the human-expert.

chirality prediction matching the chirality of the human expert (**Figure 7(ii)**) we see no negative correlations.

We next applied the network trained on full and partial duration flux ropes to the aforementioned subset of 75 Wind Fr events. A summary of the results are shown in **Table 3** where we list median differences between network predictions and hand-fit values as a function of flux rope observed. Also shown are the percentage of events where predicted chirality and hand-fit chirality match. The median difference in longitude ranges from 58° to 89°; in latitude from 31° to 50°; and in impact parameter from 36 to 53%. The network predicted the chirality correctly between 52 and 65% of the time.

3.4 Number of Wind Events to Train a Network

Experimenting with synthetic and real flux ropes raised an interesting question: How many real flux ropes are needed to train a neural network and how many suitable flux ropes are available for such a study? We can not answer this question

TABLE 3 | Wind event summary statistics as a function of percentage of flux rope observed for the 1D network trained with both full and partial duration flux ropes. Human-fit parameters are compared to neural network predictions and the ϕ, θ, and Y_0 columns are median differences between the two. The H column is the percentage of events where the chirality prediction matches hand-fit value.

% Observed	ϕ	θ	Y_0 (%)	H (%)
10	89°	50°	36	63
20	66°	42°	39	52
30	69°	32°	53	60
40	64°	37°	39	60
50	70°	33°	37	60
60	69°	37°	51	60
70	73°	31°	44	64
80	73°	34°	53	65
90	73°	33°	47	56
100	58°	42°	44	60

TABLE 4 | Percentage of synthetic flux rope predictions with an average correlation coefficient of 0.75 or greater as a function of training set size. The 1D CNN was used for training. Each training set size was repeated three time, each time taking a different random sample. The percentages reported are the average of the three repetitions. The SD column lists the standard deviation of the three repetitions.

# Flux ropes in training	% ≥ 0.75	SD
29,440	99	0.5
14,592	99	0.4
7,168	98	0.05
3,584	97	0.4
1,792	97	1.3
1,024	97	0.6
512	95	0.3
256	93	0.6
128	88	0.88

conclusively. As discussed in a previous section, and elaborated on below, neural networks trained on synthetic events do not transfer perfectly to Wind. However, we can perform one additional experiment to roughly gauge an answer.

We re-used the train-validation-test split of our synthetic flux ropes mentioned in **Section 2.2**. We then set up a loop of nine iterations. In each iteration, we randomly selected a diminishing subset of the training data, trained a 1D CNN with that subset, and then evaluated the trained model on the testing set of 19,600 synthetic flux ropes. The subsets were selected at random to simulate what happens in practice where we cannot dictate the orientation of flux ropes observed by a spacecraft. The testing consisted of using the trained CNN to make orientation predictions for each of the 19,600 test flux ropes, use those orientation predictions to create the corresponding magnetic field profiles, and correlate those magnetic field profiles with the magnetic field profiles of the test flux rope. As an evaluation metric, we computed the percentage of the test correlations greater than or equal to 0.75. Within each iteration, the subsetting-training-prediction-correlation workflow was repeated three times to investigate how the random subsetting might impact the results.

Table 4 lists the results. Over ninety percent of testing events have an average correlation coefficient above 0.75 as long as the training set size is over 200 events. Put another way, a 1D CNN trained with roughly 200 events produces average correlation coefficients on par with the 2D 3-input CNN (middle panel of 5). We do note, however, that our experiment is based on training the network with a specific flux rope model and simulated (synthetic) flux ropes. The specific flux rope model chosen will play a role as more complex descriptions of flux ropes (i.e., taking into account compression/expansion) will have more output parameters, which in turn will impact accuracy. In addition, these synthetic flux ropes do not take into account the turbulent fluctuations found in real flux ropes - a further source of prediction error. Nevertheless, it is interesting to note that we may be tantalizingly close to a neural network trained on real observations. There are 151 Wind events in Nieves-Chinchilla et al. (2019) that could potentially be used in training. The HELIO4CAST ICME catalog version 2.1 (Moestl et al., 2020) has over 1,000 ICMEs identified from multiple spacecraft. It is unknown how many of these ICMEs have associated flux ropes. Once identified, assuming there are enough, those flux ropes will need to be fit by human experts to provide labeled data for supervised learning. Nevertheless, our experiment provides the intriguing result that a few hundred more events may be all that is needed. Existing ICME catalogs may hold enough events that a concerted effort could lead to training set of real flux ropes in the coming years.

4 DISCUSSION AND CONCLUSION

Our experiments have demonstrated that convolutional neural networks are capable of providing extremely reliable characterizations of flux ropes from synthetic data. A trained network can use the structure of simulated magnetic field vectors to learn filters that map to accurate flux rope key parameter predictions; successfully inferring large scale, 3D information from single-point measurements.

When trained only on examples of full duration flux ropes, all three architectures predict key parameters of a flux rope which correlate well with the input data; however the best performing is the 1D network that feeds on time series data. Although the 2D networks that use hodogram style input do not see the same, perfect accuracy in predicting the chirality as the 1D network, the difference is statistically minor. The biggest weakness in the hodogram-input CNNs is when interpreting flux rope traces generated with a high spacecraft impact parameter. It is possible that similarities in hodogram shape profile between low- and high-valued Y_0 are activating similar filters in the 2D networks and leading to poor predictions in these cases.

When we extend the synthetically trained networks to include both partial and full duration traces through flux ropes, we still find this approach highly accurate. The CNNs are capable of making reliable predictions having only seen a fraction of the full flux rope. Although the overall discrepancy between the true and predicted values is higher than when done with only full duration traces, all median differences are well within a tolerable limit. In

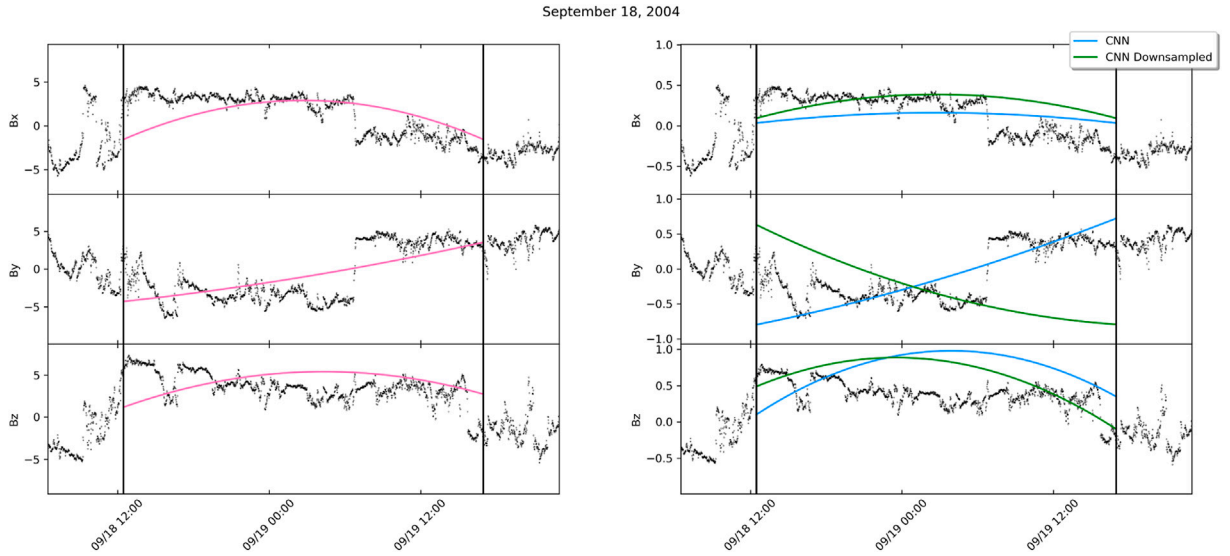

FIGURE 8 | Wind MFI measurements for September 18–19, 2004 overlaid with simulated spacecraft flux rope crossings. The left panel shows the crossing described by the human-fit parameters. The right panel shows the crossings described by the CNN predictions when it is given both the Full and Downsampled input.

these idealized, synthetic, circular-cylindrical flux ropes even the poorest performing network is able to predict orientation angles with a median error under 8° after only observing 10% of a simulated spacecraft crossing. The 1D network here, at only 10% observed, is able to give predictions with lower median difference than the 2D networks do when trained and tested with only full flux ropes. All three models show a peak performance at some point prior to seeing 100% of the flux rope crossing, perhaps due to some similarity in shape between low percentage of observation and high percentage. It is interesting to note that the 2D networks hit their peak predictive point earlier than the 1D CNN, 2D one input at 50–60% and 2D three input even earlier at 40–50%. This suggests that research into where the convolutional network is looking (for example, with the Grad-CAM method (Selvaraju et al., 2017)) can help us further understand the benefits and limitations of hodograms and time series as inputs. Future research will examine where the network is focusing its attention and if this can be exploited for more accurate predictions earlier in the forecasting process.

With the success of the 1D CNN in real-time forecasting from idealized synthetic data, we evaluated this trained 1D network on partial Wind event data. Overall, the neural network struggles to reproduce the accuracy achieved on the synthetic data set. Unlike the synthetic case, we see no trend towards a peak performance point dependent on the amount of flux rope observed. When looked at on a case-by-case basis, there are a few specific events in which the neural network is able to make accurate predictions after only seeing a fraction of the flux rope. In general, however, the median difference in angle and impact parameter prediction falls well outside any tolerance levels for useful prediction and the chirality is only correct approximately 60% of the time. Clearly, the partial-trained CNN cannot be transferred as-is to real-time application, but insight can be found by examining the results of the full duration network evaluated with Wind events.

Applying the 1D CNN trained only on full duration synthetic flux ropes to *in situ* Wind events, we again see the individual parameter predictions show significant deviation from hand fit values. However, we note lower median differences and higher H accuracy than when the network trained on both full and partial events was applied to Wind. Using down-sampled input improves this even further. Yet, by looking at the average correlation scores we see that the flux rope analytical model is robust to small deviations - small changes in longitude in particular do not lead to significant differences in reconstructed time series. We also find the neural network robust to variation in solar wind speed, expansion/compression, duration, and to some degree, magnetic field fluctuations. The neural networks were trained on synthetic data that was all generated with a simulated solar wind velocity of 450 km/s and simulated flux rope radius of 0.07 AU; yet, are able to offer reasonable predictions for Wind Fr events having significant differences in solar wind speed, expansion/compression, duration, and magnetic field fluctuations.

The neural network gives reliable predictions in a number of events and exhibits a distribution of average correlations that is qualitatively similar to those from the human expert. As evident in the right-most column of **Figure 7**, the neural network results in better average correlation in nearly half of the 75 events. When we consider only cases in which the network prediction for H matches the human-fit H the correlation to Wind data is even greater.

Analysis reveals two primary reasons the neural network performs less accurately on Wind events; incorrect physical model (Wind flux ropes not fitting the circular cylindrical

assumptions) and internal physical processes (such as fluctuations and discontinuities) that alter the expected magnetic field profile of a smooth flux rope. An example of a flux rope with magnetic field fluctuations and a discontinuity is shown in the event with MO beginning on 18 September 2004 in **Figure 8**. Down-sampling the Wind magnetic field data from 1-min to 15-min prior to interpolating to 50 points reduces the difference between neural network predictions and hand-fit values. The down-sampling further smooths out the magnetic field time series removing small-scale fluctuations. However, down-sampling cannot account for all observed internal physical processes that lead to a deviation from the expected smooth flux rope profile. The September 2004 event illustrates how differences in data processing can have a strong effect on the resulting prediction. In this particular example, the predictions made from the Wind data without prior averaging match the hand fit predictions well, while those from the down-sampled input clearly lost important information. The choice of 15-min averaging was arbitrary and is presented here to highlight how data pre-processing can have both positive and negative impacts on prediction accuracy. It remains for future research to systematically address fluctuations and determine an optimal input resolution.

Of the total 151 Wind Fr events in Nieves-Chinchilla et al. (2019), only 41% were classified as a flux rope by the neural network developed in dos Santos et al. (2020) when trained with no fluctuations. This same network classified 84 and 76% as flux ropes when trained with synthetic data augmented with 5 and 10% Gaussian fluctuations, respectively. In other words, some of the Wind events on which we do poorly finding good parametrization, would not have been considered a flux rope by the first step of an automated fitting workflow. At present, magnetic field fluctuations are not fully accounted for in flux rope analytical models and pre-processing of neural network input data does not fully address the discrepancy between synthetic and spacecraft observed flux ropes. Accurately accounting for fluctuations in measured data appears to be a significant factor for improving an automated space weather forecasting pipeline. Early experimentation with 5% Gaussian fluctuations in our study did not lead to significant improvement. Solar wind and flux rope turbulence is known to be non-Gaussian. Yet, at present, a complete understanding of turbulence leads analytical models lacking in this regard. We choose to not introduce non-realistic fluctuations and instead will explore physics-based turbulence enhancements to the analytical model in future research.

The ultimate source of prediction error in any CNN is in the inputs not matching any of the learned filters. In the case of Wind events, we notice that the neural network trained with only full duration flux ropes incorrectly predicts chirality in nearly 20% of Wind events. This leads to poor correlation coefficients as the reconstructed time series do not match the Wind observations. Yet, across all implementations of the CNNs with synthetic data the CNNs overwhelmingly identify the correct chirality. This indicates that the convolutional filters the network learned to predict chirality do not transfer to Wind events; that the filters learned to focus on a quality in the synthetic data that is not shared in the real observations. Interestingly, down-sampling has no effect on chirality predictions. We believe this source of error is related to the physical model chosen to simulate the flux ropes. Wind flux ropes show deviations from the circular cylindrical assumption. This opens the door to tantalizing future evaluations of physics-based flux rope models using an ensemble of neural networks, each trained with a different physical model.

Partial duration predictions and real-time forecasting are not really feasible at this time due in large part to features in the real data that are not present in the training set, though the concept of using CNNs to infer 3D geometric parameters from an *in situ* measurement have been borne out. Additionally, the neural networks have helped highlight the limitations of the physics-based model and even suggested better fittings of some Wind flux ropes. Future work will include implementing a single, physics-based loss function into the CNN to replace the four separate loss functions in the current design as well as enhancing analytical flux rope models to produce training data that includes more realistic turbulence and asymmetry.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: study conception and design: TN, AN, LS, and TN-C; flux rope analytical model development and coding: TN-C developed the analytical model and initial version of the code. AN ported the code to language/version used in this study; synthetic data generation: TN; neural network design and implementation; TN, AN, LS; neural network training and testing: TN; analysis and interpretation of results: TN, AN, LS, and TN-C; draft manuscript preparation: TN, AN, LS, and TN-C . All authors reviewed the results and approved the final version of the manuscript.

FUNDING

LS was supported by NASA Grant 80NSSC20K1580.

ACKNOWLEDGMENTS

The first author would like to acknowledge Ron Lepping, Daniel Berdichevsky, and Chin-Chun Wu for their many helpful discussions surrounding flux rope physics during earlier projects involving flux rope detection and analysis.

REFERENCES

Baker, D. N., and Lanzerotti, L. J. (2008). A Continuous L1 Presence Required for Space Weather. *Space Weather* 6, 1. doi:10.1029/2008SW000445

Burlaga, L. F. (1988). Period Doubling in the Outer Heliosphere. *J. Geophys. Res.* 93, 4103–4106. doi:10.1029/JA093iA05p04103

Burlaga, L., Sittler, E., Mariani, F., and Schwenn, R. (1981). Magnetic Loop Behind an Interplanetary Shock: Voyager, Helios, and Imp 8 Observations. *J. Geophys. Res.* 86, 6673–6684. doi:10.1029/JA086iA08p06673

Camporeale, E. (2019). The Challenge of Machine Learning in Space Weather: Nowcasting and Forecasting. *Space Weather* 17, 1166–1207. doi:10.1029/2018SW002061

[Software] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. Accessed on: January 2021.

[Software] Chollet, F. (2015). Keras. Available at: https://keras.io (Accessed on: January 2021).

dos Santos, L. F. G., Narock, A., Nieves-Chinchilla, T., Nuñez, M., and Kirk, M. (2020). Identifying Flux Rope Signatures Using a Deep Neural Network. *Sol. Phys.* 295, 131. doi:10.1007/s11207-020-01697-x

Gosling, J. T., Pizzo, V., and Bame, S. J. (1973). Anomalously Low Proton Temperatures in the Solar Wind Following Interplanetary Shock Waves-Evidence for Magnetic Bottles? *J. Geophys. Res.* 78, 2001–2009. doi:10.1029/JA078i013p02001

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array Programming with NumPy. *Nature* 585, 357–362. doi:10.1038/s41586-020-2649-2

Jian, L., Russell, C. T., Luhmann, J. G., and Skoug, R. M. (2006). Properties of Interplanetary Coronal Mass Ejections at One AU during 1995 - 2004. *Sol. Phys.* 239, 393–436. doi:10.1007/s11207-006-0133-2

Kilpua, E., Koskinen, H. E. J., and Pulkkinen, T. I. (2017). Coronal Mass Ejections and Their Sheath Regions in Interplanetary Space. *Living Rev. Sol. Phys.* 14, 5–83. doi:10.1007/s41116-017-0009-6

Kilpua, E., Koskinen, H. E. J., and Pulkkinen, T. I. (2017). Coronal Mass Ejections and Their Sheath Regions in Interplanetary Space. *Living Rev. Sol. Phys.* 14, 33. doi:10.1007/s41116-017-0009-6

Kingma, D., and Ba, J. (2015). "Adam: A Method for Stochastic Optimization," in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015. Editors Y. Bengio and Y. LeCun (Conference Track Proceedings).

Klein, L. W., and Burlaga, L. F. (1982). Interplanetary Magnetic Clouds at 1 au. *J. Geophys. Res.* 87, 613–624. doi:10.1029/JA087iA02p00613

LeCun, Y., and Bengio, Y. (1995). "Convolutional Networks for Images, Speech, and Time Series," in *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA: MIT Press, 3361, 1995.

Lepping, R. P., Acũna, M. H., Burlaga, L. F., Farrell, W. M., Slavin, J. A., Schatten, K. H., et al. (1995). The Wind Magnetic Field Investigation. *Space Sci. Rev.* 71, 207–229. doi:10.1007/BF00751330

Lepping, R. P., Jones, J. A., and Burlaga, L. F. (1990). Magnetic Field Structure of Interplanetary Magnetic Clouds at 1 AU. *J. Geophys. Res.* 95, 11957–11965. doi:10.1029/JA095iA08p11957

Manchester, W., Kilpua, E., Liu, Y., Lugaz, N., Riley, P., Torok, P. T., et al. (2017). The Physical Processes of Cme/icme Evolution. *Space Sci. Rev.* 212 (3-4), 1159. doi:10.1007/s11214-017-0394-0

Moestl, C., Weiss, A., Bailey, R., and Reiss, M. (2020). Helio4cast Interplanetary Coronal Mass Ejection Catalog v2.1. figshare. Dataset. doi:10.6084/m9.figshare.6356420.v11

Nguyen, G., Fontaine, D., Aunai, N., Vandenbossche, J., Jeandet, A., Lemaitre, G., et al. (2018). "Machine Learning Methods to Identify Icmes Automatically," in 20th EGU General Assembly, EGU2018, Proceedings from the Conference Held, Vienna, Austria, 4-13 April, 2018, 1963.

Nieves-Chinchilla, T., Jian, L. K., Balmaceda, L., Vourlidas, A., dos Santos, L. F. G., and Szabo, A. (2019). Unraveling the Internal Magnetic Field Structure of the Earth-Directed Interplanetary Coronal Mass Ejections during 1995 - 2015. *Sol. Phys.* 294, 89. doi:10.1007/s11207-019-1477-8

Nieves-Chinchilla, T., Linton, M. G., Hidalgo, M. A., Vourlidas, A., Savani, N. P., Szabo, A., et al. (2016). A Circular-Cylindrical Flux-Rope Analytical Model for Magnetic Clouds. *Astrophysical J.* 823, 27. doi:10.3847/0004-637X/823/1/27

Nieves-Chinchilla, T., Vourlidas, A., Raymond, J. C., Linton, M. G., Al-haddad, N., Savani, N. P., et al. (2018). Understanding the Internal Magnetic Field Configurations of ICMEs Using More Than 20 Years of Wind Observations. *Sol. Phys.* 293, 25. doi:10.1007/s11207-018-1247-z

Ogilvie, K. W., Chornay, D. J., Fritzenreiter, R. J., Hunsaker, F., Keller, J., Lobell, J., et al. (1995). SWE, A Comprehensive Plasma Instrument for the Wind Spacecraft. *Space Sci. Rev.* 71, 55–77. doi:10.1007/BF00751326

Reiss, M. A., Möstl, C., Bailey, R. L., Rüdisser, H. T., Amerstorfer, U. V., Amerstorfer, T., et al. (2021). Machine Learning for Predicting the Bz Magnetic Field Component from Upstream *In Situ* Observations of Solar Coronal Mass Ejections. *Space Weather* 19. doi:10.1029/2021SW002859

Rodríguez-García, L., Gómez-Herrero, R., Zouganelis, I., Balmaceda, L., Nieves-Chinchilla, T., Dresing, N., et al. (2021). The Unusual Widespread Solar Energetic Particle Event on 2013 August 19: Solar Origin and Particle Longitudinal Distribution. *Astron. Astrophysics* 653, A137. doi:10.1051/0004-6361/202039960

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, October 22–29, 2017, 618–626. doi:10.1109/ICCV.2017.74

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* 17, 261–272. doi:10.1038/s41592-019-0686-2

Vourlidas, A. (2014). The Flux Rope Nature of Coronal Mass Ejections. *Plasma Phys. Control Fusion* 56, 064001. doi:10.1088/0741-3335/56/6/064001

Conflict of Interest: Author AN is employed by ADNET Systems Inc. and is contracted to NASA/Goddard Space Flight Center where a portion of her employment is to carry out scientific software development and data analysis support. Despite the affiliation with a commercial entity, author AN has no relationships that could be construed as a conflict of interest related to this work.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Astronomy and Space Sciences

REVIEW
published: 14 March 2022
doi: 10.3389/fspas.2022.826947

Recent Applications of Bayesian Methods to the Solar Corona

Iñigo Arregui[1,2]*

[1]Instituto de Astrofísica de Canarias, Tenerife, Spain, [2]Departamento de Astrofísica, Universidad de La Laguna, Tenerife, Spain

Solar coronal seismology is based on the remote diagnostics of physical conditions in the corona of the Sun by comparison between model predictions and observations of magnetohydrodynamic wave activity. Our lack of direct access to the physical systems of interest makes information incomplete and uncertain so our conclusions are at best probabilities. Bayesian inference is increasingly being employed in the area, following a general trend in the space sciences. In this paper, we first justify the use of a Bayesian probabilistic approach to seismology diagnostics of solar coronal plasmas. Then, we report on recent results that demonstrate its feasibility and advantage in applications to coronal loops, prominences and extended regions of the corona.

Keywords: Sun: corona, Sun: magnetic fields, magnetohydrodynamics (MHD), waves, solar coronal seismology, bayesian statistics

OPEN ACCESS

Edited by:
Bala Poduval,
University of New Hampshire,
United States

Reviewed by:
Peng-Fei Chen,
Nanjing University, China
Ajay Tiwari,
Northumbria University,
United Kingdom

***Correspondence:**
Iñigo Arregui
iarregui@iac.es

Specialty section:
This article was submitted to Space Physics, a section of the journal Frontiers in Astronomy and Space Sciences

Received: *01 December 2021*
Accepted: *02 February 2022*
Published: *14 March 2022*

Citation:
Arregui I (2022) Recent Applications of Bayesian Methods to the Solar Corona. Front. Astron. Space Sci. 9:826947. doi: 10.3389/fspas.2022.826947

1 INTRODUCTION

The aim of this paper is to give a rationale for the use of Bayesian methods in the study of the solar corona and to show recent applications in the area of solar coronal seismology. Coronal seismology aims to infer difficult to measure physical parameters in magnetic and plasma structures, such as coronal loops and prominence plasmas, by a combination of observations of wave activity and theoretical models, usually under the MHD approximation (Uchida, 1970; Roberts et al., 1984). Because of our lack of direct access to the physical systems of interest information is incomplete and uncertain. As a consequence, solar atmospheric seismology deals with inversion problems that are probabilistic in nature and our conclusions can only be probabilities at best. A prototypical example is the determination of the magnetic field strength in coronal loops from the observational measurement of the kink speed of transverse oscillations (Nakariakov and Ofman, 2001). Only after assumptions about the loop plasma density and the density contrast one can derive the magnetic field. Since the values of the density and density contrast have probabilistic distributions, the derived magnetic field has a probabilistic distribution.

Bayesian analysis is increasingly being used in astrophysics. **Figure 1** shows the number of Bayesian astrophysics papers as a function of year. The first studies (already 50 years ago) dealt with both technical problems, such as the construction of image restoration algorithms (Richardson, 1972), as well as with procedures for formalising the evaluation of astrophysical hypotheses by comparison between theoretical predictions and observational data (Sturrock, 1973). It took two more decades for the Bayesian approach to be adopted in solar physics. Initial solar applications were focused on statistical analyses of solar neutrino data (Gates et al., 1995), followed by studies on solar flare prediction (Wheatland, 2004), the analysis of solar global oscillations (Marsh et al., 2008), and the inversion of magnetic and thermodynamic properties of the solar atmosphere from the analysis of spectro-polarimetric data (Asensio Ramos et al., 2007). The first study that made use of Bayesian analysis in coronal seismology was by Arregui and Asensio Ramos (2011), who inferred coronal loop physical parameters from observed periods and damping times of their transverse oscillations. In the

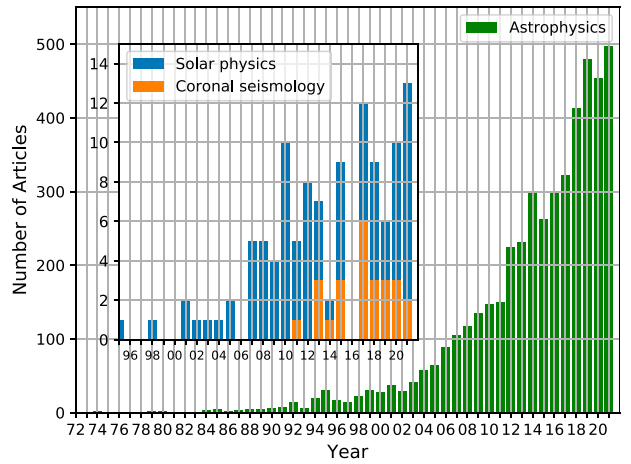

FIGURE 1 | Bar plots representing the yearly number of referred articles employing Bayesian analysis techniques in the field of astrophysics, the branch of solar physics, and the research area of coronal seismology. At the time of writing (November 2021), the total number of articles amounts to 4678 in astrophysics, 114 in solar physics, and 25 in coronal seismology. Source: NASA Astrophysics Data System (ADS), digital library operated by the Smithsonian Astrophysical Observatory (SAO).

last decade, about 25 studies in coronal seismology have made use of Bayesian techniques. They deal with parameter inference, model comparison, and model averaging applications to gain information on the magnetic field and the plasma conditions in structures in the solar corona and in solar prominences. Here, we discuss some recent developments in the area.

The layout of the article is the following. **Section 2** describes the basic principles and the tools used to perform parameter inference and model comparison in the Bayesian framework. In **Section 3**, first, results on the inference of physical parameters in coronal loops and prominence plasmas are described. Then, examples are shown on the application of model comparison to the assessment of the damping mechanism(s) of coronal waves. A summary is presented in **Section 4**.

2 BASIC PRINCIPLES OF BAYESIAN INFERENCE AND MODEL COMPARISON

Bayesian analysis considers any inversion problem, in terms of probabilistic inference, as the task of estimating the degree of belief on statements about parameter values or model evidence, conditional on observed data. It uses Bayes' rule (Bayes and Price, 1763),

$$p(\theta|d,M) = \frac{p(d|\theta,M)\,p(\theta|M)}{\int p(d|\theta,M)\,p(\theta|M)\,d\theta}, \tag{1}$$

which says that our state of knowledge about a set of parameters θ of a given model M, conditional on the observed data d, is a combination of what we know independently of the data, the prior $p(\theta|M)$, and the likelihood of obtaining a given data realisation as a function of the parameter vector, the likelihood function $p(d|\theta, M)$. Their combination gives the posterior distribution, $p(\theta|d, M)$, that encloses all the information about the set of parameters conditional on the observed data and the assumed model. The prior and the likelihood function need to be directly assigned in order to compute the posterior. Bayes' rule offers a tool to perform rational inference based on the combination of conditional probability distributions. The tool can be applied at three different levels.

In parameter inference the global posterior is computed for the full set of N parameters, $\theta = \{\theta_1, \ldots, \theta_i, \ldots, \theta_N\}$, and is then marginalised to obtain information about the one of interest. This is achieved by integration of the full posterior with respect to the remaining parameters,

$$p(\theta_i|d) = \int p(\theta|d)d\theta_1 \ldots, d\theta_{i-1}d\theta_{i+1} \ldots, d\theta_N. \tag{2}$$

This is the so-called marginal posterior for model parameter θ_i, which contains all the information available in the priors and the data. The uncertainty of the rest of parameters to the one of interest is correctly propagated by this procedure. To summarise the result one can then provide the mean, the mode, the median, etc. It is common to provide the maximum a posteriori estimate of the inferred parameter, θ_i^{MAP}, the value of θ_i that makes the posterior the largest together with credible regions containing a particular fraction of the mass of the distribution. A simple way of computing such credible region is to sort the probability values $p(\theta_i|d)$ in descending order. Then, starting with the largest one, add successively smaller values of $p(\theta_i|d)$ until the next value would exceed the desired value of e.g., 68%. At each step, one needs to keep track of the corresponding θ_i values. The credible region is then the range in parameter space that includes all the θ_i values corresponding to the $p(\theta_i|d)$ values that were added. The boundaries of the credible region give the lower and upper errors. They are the smallest and largest values of θ_i obtained by this procedure. The process of marginalisation can also be applied to the so-called nuisance parameters, those that must be incorporated in the modelling but are not of immediate interest.

The denominator in **Eq. 1** is the so-called marginal likelihood or evidence,

$$p(d|M) = \int_\theta p(\theta, d|M)\,d\theta = \int_\theta p(d|\theta,M)\,p(\theta|M)\,d\theta, \tag{3}$$

an integral of the joint distribution of parameters and data over the full parameter space that normalises the likelihood function to turn the result into a probability. It plays a crucial role in model comparison because it is a measure of relational evidence. The measure of evidence is relational because it examines a relation between the predictions by model M and the observed data d. The marginal likelihood quantifies the evidence for a model in relation to the data that it predicts. The general aim of model comparison is to assess the relative evidence between alternative models in explaining the same data. Given two models, M_1 and M_2, this is achieved with the calculation of the posterior ratio $p(M_1|d)/p(M_2|d)$. If the two models are equally probable a priori, $p(M_1) = p(M_2)$, and the posterior ratio is equal to the ratio of marginal likelihoods of the two models

$$B_{12} = 2\log\frac{p(M_1|d)}{p(M_2|d)} = 2\log\frac{p(d|M_1)}{p(d|M_2)} = -B_{21}, \quad (4)$$

where the logarithmic scale is used to translate Bayes factors into levels of evidence. The Bayes factors B_{12} and B_{21} defined in **Eq. 4** measure relative evidence. They quantify the relative plausibility of each of the two models to explain the same data. To evaluate the levels of evidence an empirical table, such as the one by Kass and Raftery (1995), is employed. For values of B_{12} from 0 to 2, the evidence in favour of model M_1 in front of model M_2 is inconclusive; for values from 2 to 6, positive; for values from 6 to 10, strong; and for values above 10, very strong. A similar tabulation applies to B_{21}.

After a model comparison procedure has been performed, it may be the case that the evidence in favour of any of the models under consideration is not large enough to deem positive evidence. A convenient solution is then to consider the third level of Bayesian inference, model averaging. This is a procedure that combines the posteriors inferred with each model to calculate a model-averaged posterior,

$$p(\theta_i|d) = \sum_k p(\theta_i|d, M_k)\, p(M_k|d), \quad (5)$$

weighted with the evidence for each model. In this manner, parameter constraints that account for the uncertainty about the models are obtained. Such a calculation makes use of all the available information in the data and models in a fully consistent manner. The resulting marginal posteriors are the best inference one can obtain with the available information.

3 RECENT APPLICATIONS TO THE SOLAR CORONA

After the first application of Bayesian methods to coronal seismology by Arregui and Asensio Ramos (2011), most of the initial studies made use of simple forward models for the prediction of oscillation properties of magnetic structures, such as periods and damping times, and integration over a grid of points in low-dimensional parameter and data spaces (see e.g., Arregui et al., 2013a,b; Asensio Ramos and Arregui, 2013; Arregui and Asensio Ramos, 2014; Arregui and Soler, 2015; Arregui et al., 2015; Arregui and Goossens, 2019). Other studies considered the analysis of the time series of displacement amplitude of oscillations to infer equilibrium properties of coronal loops (see e.g., Pascoe et al., 2017a,b; Pascoe et al., 2018; Goddard et al., 2018). A review summarising those initial applications can be found in Arregui (2018). Additional developments were possible by the creation and application of data analysis tools based on Markov Chain Monte Carlo sampling of posterior distributions (see e.g., Goddard et al., 2017; Pascoe et al., 2017c, 2019; Duckenfield et al., 2019; Pascoe et al., 2020b,a). Details about these methods and their use as diagnostic tools for coronal seismology can be found in Anfinogentov et al. (2021b,a). In the following, we discuss some recent results, focusing on the inference of physical parameters in coronal loops and prominence plasmas and on the damping of transverse oscillations in coronal loops and extended regions of the solar corona.

3.1 Inferring the Magnetic Field Strength and Plasma Density in Coronal Loops

The first modern application of coronal seismology was presented by Nakariakov and Ofman (2001). By interpreting the transverse coronal loop oscillations observed with the Transition Region and Coronal Explorer (TRACE) as the fundamental kink mode of a magnetic flux tube in the long wavelength limit, they showed how the combination of observed period (P) and loop length (L) can enable to constrain the magnetic field strength. The procedure starts with the assumption of a simple expression, model M_1, for the phase speed of the kink mode as a function of the Alfvén speed in the interior of the loop, $v_{\rm Ai} = B_0/\sqrt{\mu_0\rho_{\rm i}}$, and the density contrast, $\zeta = \rho_{\rm i}/\rho_{\rm e}$,

$$v_{\rm ph} = v_{\rm Ai}\left(\frac{2\zeta}{1+\zeta}\right)^{1/2}, \quad (6)$$

with μ_0 the magnetic permeability and $\rho_{\rm i,\ e}$ the internal and external densities. This expression is valid assuming coronal loops can be modelled as one-dimensional density enhancements in cylindrical coordinates with the magnetic field pointing along the axis of the tube and under the long wavelength approximation. Adopting a given value for the density contrast, ζ, the observationally estimated phase speed ($v_{\rm ph} \sim 2L/P$) enables to obtain the Alfvén speed $v_{\rm Ai}$. By further assuming values of loop density on a given range, a range of magnetic field strength values is obtained.

In their Bayesian analysis, Arregui et al. (2019) showed that the problem can be formulated in terms of the inference of a three-dimensional posterior from one observable with the use of Bayes' rule as the product of likelihood and prior,

$$p(\{\rho_{\rm i}, \zeta, B_0\}|v_{\rm ph}, M_1\} \sim p(v_{\rm ph}|\{\rho_{\rm i}, \zeta, B_0\}, M_1)\, p(\{\rho_{\rm i}, \zeta, B_0\}|M_1).$$

Considering a particular observed event, a Gaussian likelihood function and uniform prior distributions for the three unknown parameters, $\theta = \{\rho_{\rm i}, \zeta, B_0\}$, over plausible ranges leads to the marginal posterior distribution for the magnetic field strength shown in **Figure 2A**. The result shows that not all values in the range found by Nakariakov and Ofman (2001) are equally probable. A well constrained marginal posterior is obtained which specifies the particular plausibility for each value of the magnetic field strength in the range. From this, estimates with asymmetric error bars can be obtained. Regarding the other two parameters, the density contrast and the loop density, their distribution does not permit to obtain constrained information on their most probable values. The marginal posterior for the magnetic field strength incorporates the uncertainty on these two parameters and can still be properly inferred, even if the values of plasma density inside and outside the coronal loops are highly uncertain.

One advantage of the Bayesian approach is that it offers a self-consistent way to update the posteriors when additional information is available. Spectroscopic measurements enable

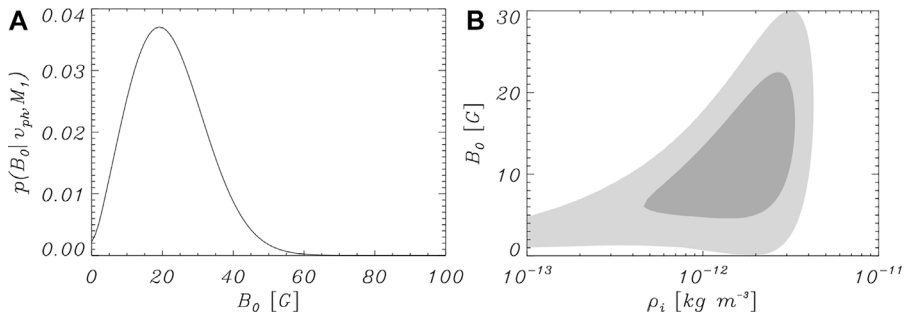

FIGURE 2 | (A) Posterior probability distribution for the magnetic field strength for a loop oscillation event with observed phase speed v_{ph} = 1,030 ± 410 km s^{-1} under model M_1, given by **Eq. 6**. The inferred median value for the magnetic field strength is $B_0 = 21^{+12}_{-9}$ G, with uncertainties given at the 68% credible interval. **(B)** Joint two-dimensional posterior distribution for the internal density of the waveguide and the magnetic field strength obtained for the inference with a Gaussian prior for the internal density centred at $\mu_{\rho_i} = 1.9 \times 10^{-12}$ kg m^{-3} and with $\sigma_{\rho_i} = 0.5\mu_{\rho_{mathrmi}}$. The inference with the more informative prior on density leads to $B_0 = 13^{+7}_{-6}$ G and $\rho_i = (2.2^{+0.9}_{-0.9}) \times 10^{-12}$ kg m^{-3}. The outer boundaries of the light grey and dark grey shaded regions indicate the 95% and 68% credible regions, respectively. From Arregui et al. (2019).

us to obtain information about physical properties of the coronal plasma, such as the density. Consider we have some estimate for the density inside the oscillating loop. This additional information can be added to the inference in the form of a Gaussian prior for the density, centred in the measured value. **Figure 2B** shows the joint posterior for plasma density and magnetic field strength for such an inference, with grey-shaded areas indicating the 68% and 95% credible regions, respectively. This example shows that the inclusion of additional information enables us to better constrain our estimates for the magnetic field strength and plasma density.

Observations show that transverse coronal loop oscillations are quickly damped, with characteristic damping times of a few oscillatory periods. Arregui et al. (2019) evaluated the influence of this observable on the inference of the magnetic field strength. The simplest available and more commonly accepted model is damping by resonant absorption due to the inhomogeneity of the plasma in the cross-field direction (Goossens et al., 2002; Ruderman and Roberts, 2002). Under the thin tube and thin boundary approximations, with a non-uniform layer of width l much shorter than the tube radius R ($l \ll R$), the damping time is given by

$$\tau_d\left(\rho_i, \zeta, B_0, l/R\right) = \frac{2}{\pi}\left(\frac{\zeta+1}{\zeta-1}\right)\left(\frac{1}{l/R}\right)\left(\frac{2L}{v_{ph}}\right). \quad (7)$$

The forward predictions of this new model M_2, given by **Eqs 6**, **7**, are coupled, hence some degree of influence is expected in the inference of the magnetic field, due to the consideration of wave damping. Now the problem can be formulated in terms of the inference of a four-dimensional posterior from three observables with the use of Bayes' rule as the product of likelihood and prior,

$$p\left(\{\rho_i, \zeta, B_0, l/R\}|\{v_{ph}, \tau_d, L\}, M_2\right) \sim p\left(\{v_{ph}, \tau_d, L\}|\{\rho_i, \zeta, B_0, l/R\}, M_2\right) p\left(\{\rho_i, \zeta, B_0, l/R\}|M_2\right).$$

Considering the same observed event as before, a Gaussian likelihood function and uniform prior distributions for the four unknown parameters, $\theta = \{\rho_i, \zeta, B_0.L\}$, over plausible ranges leads to the results displayed in **Figure 3**. The resulting marginal posteriors for the magnetic field strength for different damping times show little differences. The advantage of including the damping into the inference is that it enables to infer information on the transverse inhomogeneity length scale of the density at the boundary of the waveguide. This parameter is relevant in the context of wave dissipation processes (Arregui, 2015).

3.2 Inferring the Magnetic Field Strength and Thread Length in Prominences

Bayesian methods are also being applied in prominence seismology. Estimates of periods and phase speeds of propagating waves were obtained by Lin et al. (2009) for a number of threads in a prominence. A fundamental difference in the solution to the inverse problem, in comparison to the case with coronal loops, is that the internal prominence density is two orders of magnitude larger than the external coronal density. This makes the kink speed independent of the density contrast and simplifies **Eq. 6** to the approximate expression $v_{ph} \sim \sqrt{2}v_{Ai}$. By using this fact, Lin et al. (2009) were able to provide estimates for the magnetic field strength in the threads, upon assuming a given value for their plasma density.

Figure 4A gives ranges of variation for the magnetic field strength in 10 selected threads as a function of the prominence density computed by Montes-Solís and Arregui (2019) from data in Table 1 of Lin et al. (2009). From the Bayesian perspective, as in the case of coronal loops, all those values within the obtained ranges are not equally probable. The Bayesian solutions computed by Montes-Solís and Arregui (2019) in the form of marginal posteriors for each of the 10 threads are shown in **Figure 4B**. For each thread, the magnetic field strength can be properly inferred (see Table 2 in Montes-Solís and Arregui 2019). The distributions spread over a range of values from 1 to 20 G and seem to point to a highly inhomogeneous nature of the studied prominence area. Montes-Solís and Arregui (2019) continue their analysis with the computation of the joint two-dimensional

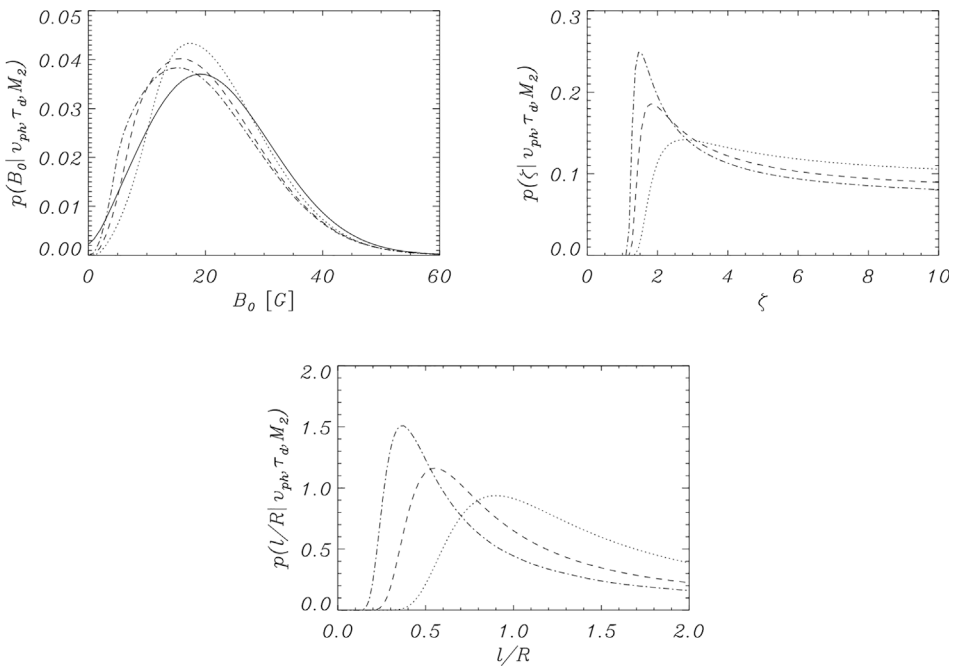

FIGURE 3 | Marginal posterior distributions for magnetic field strength, density contrast, and transverse inhomogeneity length scale for the inversion of the problem with forward model given by **Eqs 6**, **7**, a transverse oscillation with $v_{\rm ph} = 1{,}030 \pm 410$ km s^{-1} and different values for the damping time: no damping (solid line), $\tau_{\rm d} = 500$ s (dotted line), $\tau_{\rm d} = 800$ s (dashed line), and $\tau_{\rm d} = 1{,}200$ s (dash-dotted line) with an associated uncertainty of 50 s in all cases. The inferred medians with errors at the 68% credible interval are $B_0 = 21^{+12}_{-9}$ G for the undamped case, $B_0 = 20^{+11}_{-8}$ G for $\tau_{\rm d} = 500$ s, $B_0 = 19^{+11}_{-8}$ G for $\tau_{\rm d} = 800$ s, and $B_0 = 18^{+11}_{-9}$ G for $\tau_{\rm d} = 1{,}200$ s. A fixed value for the loop length $L = 1.9 \times 10^{10}$ cm was considered in all computations. Adapted from Arregui et al. (2019).

posterior for magnetic field strength and prominence density which, in the case of a Gaussian prior for the density, is well constrained (see **Figure 4C**).

In contrast to the case of coronal loops, prominence threads do not occupy the entire length of the magnetic flux tube. We only observe the cold and dense plasma occupying a fraction of a longer but unobservable structure. Soler et al. (2010) constructed a model that provides us with an approximate analytical expression for the phase speed of kink modes in partially filled tubes

$$v_{\rm ph}^{\rm par} = \frac{2}{\pi\sqrt{\frac{L_{\rm p}}{L}\left(1-\frac{L_{\rm p}}{L}\right)}}\, v_{\rm ph}^{\rm tot} \tag{8}$$

in terms of the phase speed in a totally filled tube, $v_{\rm ph}^{\rm tot} = \sqrt{2}v_{\rm Ai}$, with $L_{\rm p}$ the length of the thread and L the length of the flux tube. **Figure 4D** shows results for the inference of the magnetic field strength performed for different models for the density along the thread considering: a fully filled tube, a partially filled tube with a uniform prior distribution for $L_{\rm p}/L$, and a partially filled tube with a Gaussian prior distribution for $L_{\rm p}/L$. The results indicate the importance of having an approximate idea about the ratio $L_{\rm p}/L$ in order to obtain an accurate inference.

Even in the case of a fully filled tube, $L_{\rm p} = L$, as in the case with coronal loops, the inferred posterior for the magnetic field strength is dependent on the amount of information we have on the value of plasma density. **Figure 5** shows marginal posteriors for the magnetic field strength corresponding to thread # 5 in Table 2 of Montes-Solís and Arregui (2019). They were calculated with three different priors for the density. One considers a uniform prior. The other two Gaussian distributions centred at two different density values. The results indicate that the obtained posteriors clearly differ.

One of the reasons why prominence seismology is in a less developed stage than coronal loop seismology is because there are fewer observations of transverse oscillations in these structures, but also because of the complexity in their modelling. As in prominence threads we only observe the cold and dense part of a longer but unobservable structure, the length of the flux tube cannot be directly estimated. However, using seismology diagnostics with multiple periods one can obtain posterior probability distributions for the ratio of the length of the thread to the length of the flux tube, $L_{\rm p}/L$. Also, a number of observations show that threads oscillate and flow simultaneously. This affects the oscillation period which changes in time. By measuring the period at two different moments and using theoretical developments by Soler and Goossens (2011) a number of parameters, such as the flow speed, the length of the thread and the length of the flux tube can be inferred. Applications of these principles can be found in the study by Montes-Solís and Arregui (2019).

3.3 Assessing Damping Mechanisms for Coronal Loop Oscillations

The damping of magnetohydrodynamic waves has been a matter of interest since the first imaging observations of transverse coronal loop oscillations (Aschwanden et al., 1999; Nakariakov

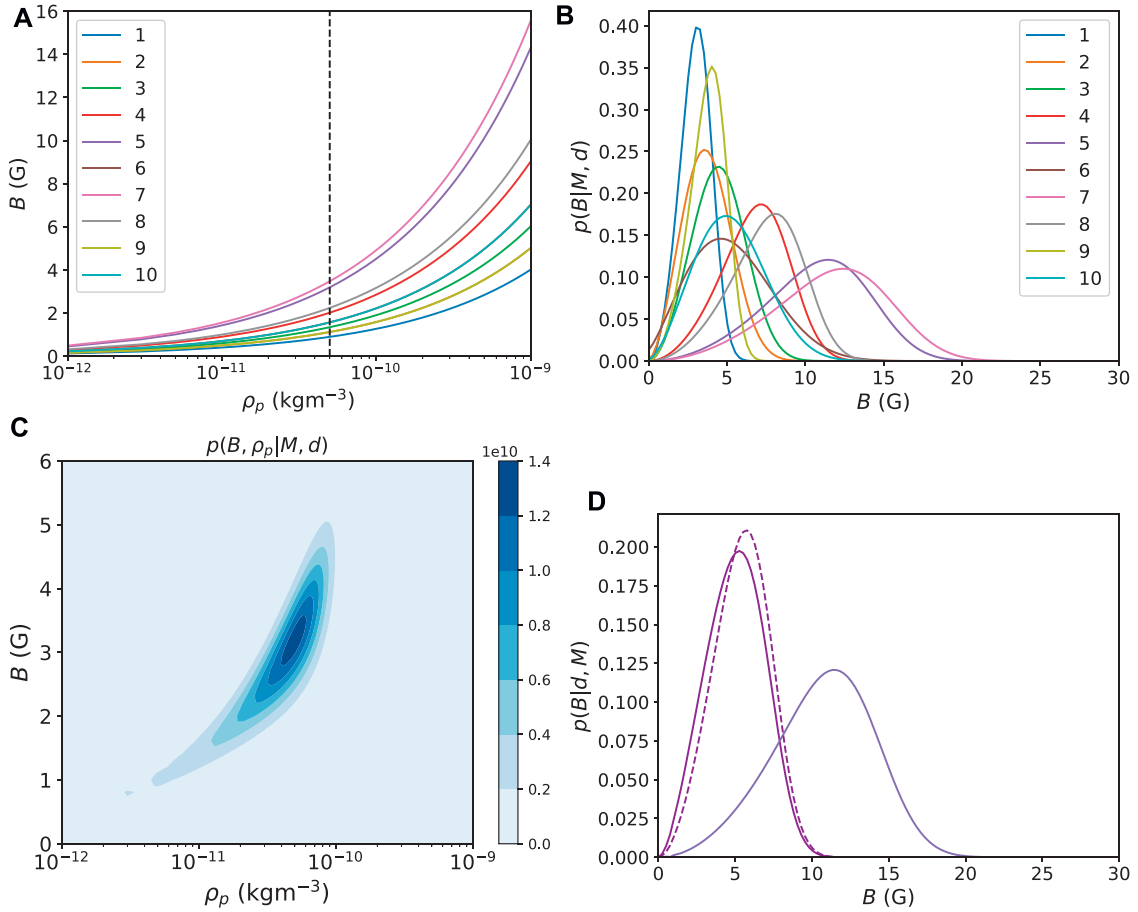

FIGURE 4 | (A) Curves obtained for each thread observed by Lin et al. (2009). **(B)** Posterior distributions of magnetic field strength (B) for each considered thread obtained with Bayesian methods. **(C)** Global posterior computed for the fifth thread, considering a Gaussian prior of the internal density (ρ_i) centred in a value equal to $\rho_p = 5 \times 10^{-11}$ kg m^{-3} and given an uncertainty of 50%. **(D)** Comparison of posterior distributions of the magnetic field strength in a totally filled tube (purple line), partially filled tube (magenta line), and partially filled tube considering a Gaussian prior for the proportion of thread length centred in $L_p/L = 0.5$ with 50% of uncertainty (dashed magenta line). From Montes-Solís and Arregui (2019).

et al., 1999). Of particular interest in explaining the observations are the mechanisms based on the cross-field inhomogeneity of the waveguides, such as phase mixing and resonant absorption (Heyvaerts and Priest, 1983; Goossens et al., 2002; Ruderman and Roberts, 2002; Goossens et al., 2006), or those involving lateral or foot-point leakage of wave energy (Spruit, 1982; Roberts, 2000; De Pontieu et al., 2001; Cally, 2003).

Attempts to discriminate between alternative mechanisms were initially focused on the computation of the damping time scales predicted by each mechanism for plausible values of the unknown relevant physical parameters. This approach enables for instance to discard viscous or resistive processes because of the too long timescales they predict. Ofman and Aschwanden (2002) proposed a method based on the comparison between theoretically predicted and fitted power-law indexes between periods and damping times to assess the plausibility of alternative damping mechanisms. This suggestion is based on the assumption that each mechanism is characterised by a particular power-law index, a premise that was shown to be questionable by Arregui et al. (2008). For instance, resonant absorption is able to generate data realisations leading to different scaling laws with different power-law indexes.

A Bayesian approach to comparing the relative plausibility among several proposed damping mechanisms for coronal loop oscillations was followed by Montes-Solís and Arregui (2017). They considered the mechanisms of resonant absorption in the Alfvén continuum (Goossens et al., 2002), phase mixing of Alfvén waves (Heyvaerts and Priest, 1983), and wave leakage of the principal leaky mode (Cally, 2003).

For resonant damping, the theoretically predicted damping time τ_d over the period P, under the thin tube and thin boundary approximations, reads (Goossens et al., 2002; Ruderman and Roberts, 2002)

$$\frac{\tau_d}{P} = \frac{2}{\pi}\frac{R}{l}\frac{\zeta+1}{\zeta-1}, \tag{9}$$

with l the thickness of the non-uniform layer at the boundary of the loop and $\zeta = \rho_i/\rho_e$ the density contrast between the internal and external densities. Plausible ranges of variation for the

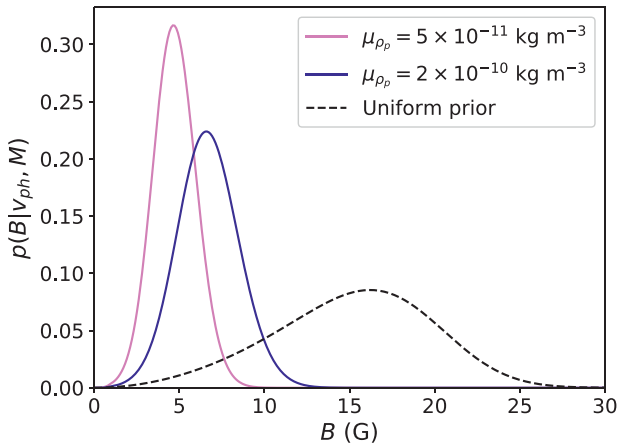

FIGURE 5 | Posterior distributions of the magnetic field strength (B) in a fully filled magnetic thread considering a Gaussian prior for the internal density (ρ_i) with mean $\mu_{\rho_p} = 5 \times 10^{-11}$ kg m^{-3} (pink), a Gaussian prior with mean $\mu_{\rho_p} = 2 \times 10^{-10}$ kg m^{-3} (blue), and a uniform prior (black dashed line). Uncertainties in Gaussian priors are considered to be of 50%. From Montes-Solís and Arregui (2019).

unknown parameters are $\zeta \in (1, 10]$ and $l/R \in (0, 2]$. They are capable of producing the observed fast damping.

For phase mixing, an analytical expression for the damping ratio was derived by Roberts (2000).

$$\frac{\tau_d}{P} = \left(\frac{3}{\pi^2 \nu}\right)^{1/3} w^{2/3} P^{-1/3}. \tag{10}$$

Here $\nu = 4 \times 10^3$ km^2 s^{-1} is the coronal kinematic shear viscosity coefficient and w the transverse inhomogeneity length scale. Considering values of the unknown parameter in the range $w \in [0.5, 6]$, the required observed damping times scales can be well reproduced.

The third considered mechanism, wave leakage, consist of the presence of a wave that radiates part of its energy to the background medium while oscillating with the kink mode frequency. An analytical expression for the damping ratio was derived by Cally (2003),

$$\frac{\tau_d}{P} = \frac{4}{\pi^4}\left(\frac{R}{L}\right)^{-2}, \tag{11}$$

with R and L the radius and length of the loop, respectively. A plausible range for their ratio is $R/L \in [10^{-4}, 0.3]$, which leads to predicted damping ratio values as low as 0.5 or as high as 10^5.

In Montes-Solís and Arregui (2017), the three damping mechanisms were compared by considering how well they are able to reproduce the observed period and damping timescales, taking into account the observations and their associated uncertainty. **Figure 6** shows the results from the computation of Bayes factors for the one-to-one comparison between damping mechanisms in the plane of observables damping time vs oscillation period. The subscripts 0, 1, and 2, are used to identify resonant absorption, phase mixing, and wave leakage, respectively. The first apparent result is that the evidence distribution in the plane of observables in favour of any of the models in comparison to another depends on the combination of observed periods and damping times. For instance, in the comparison between resonant absorption and phase mixing, **Figure 6A** shows strong and very strong evidence for resonant damping in the upper-left corner of the plane of observables. For low damping ratios, at the lower-right corner, the evidence supports the phase-mixing model. In the area in between, differently coloured bands denote different levels of evidence. **Figure 6B** shows the comparison between resonant absorption and wave leakage. In most of the observable plane, there is a lack of evidence supporting either of the two mechanisms. Only for the lowest damping ratio values there is evidence in favour of resonant damping. Finally, **Figure 6C** shows the results from the comparison between phase mixing and wave leakage. For combinations of period and damping time leading to large damping ratios, the evidence in favour of wave leakage is larger. For low damping ratios, the evidence strongly supports the mechanism of phase mixing.

The results discussed so far were obtained by application of Bayesian model comparison methods to synthetic hypothetical data in the plane of observables of period and damping time. Montes-Solís and Arregui (2017) also considered the computation of Bayes factors for a selection of 89 loop oscillation events listed in the databases by Verwichte et al. (2013) and Goddard et al. (2016). The results are displayed in **Figure 7**. The colours indicate the level of evidence, based on the magnitude of the corresponding Bayes factor. It is clear that the events in blue colour, which correspond to evidence that is not worth a bare mention, dominate in all three panels. In the comparison between resonant absorption and phase mixing (left panel), in approximately 78% of the events the evidence is not strong enough to favour one model or the other. The evidence is positive for resonant absorption in 8% of the events and for phase mixing in about 14% of the events. In the middle panel, the comparison between resonant absorption and wave leakage is shown. The evidence is not large enough to support any of the two mechanisms. The panel in the right shows the evidence assessment between phase mixing and wave leakage. In this case, the evidence is inconclusive for 79% of the events. There is positive evidence in favour of wave leakage in 3% of the events, those corresponding to oscillations with very strong damping. For the remaining 18%, the evidence is positive in favour of phase mixing.

The results presented by Montes-Solís and Arregui (2017) do not allow us to identify a unique mechanism as responsible for the quick damping of coronal loop oscillations. However, the method makes use of all the available information in the models, observed data with their uncertainty, and prior information in a consistent manner.

3.4 Evidence for Resonant Damping of Coronal Waves With Foot-point Wave Power Asymmetry

Waves propagating in extended regions of the solar corona offer another opportunity to test our models for the damping of waves and oscillations. Their existence was first demonstrated by

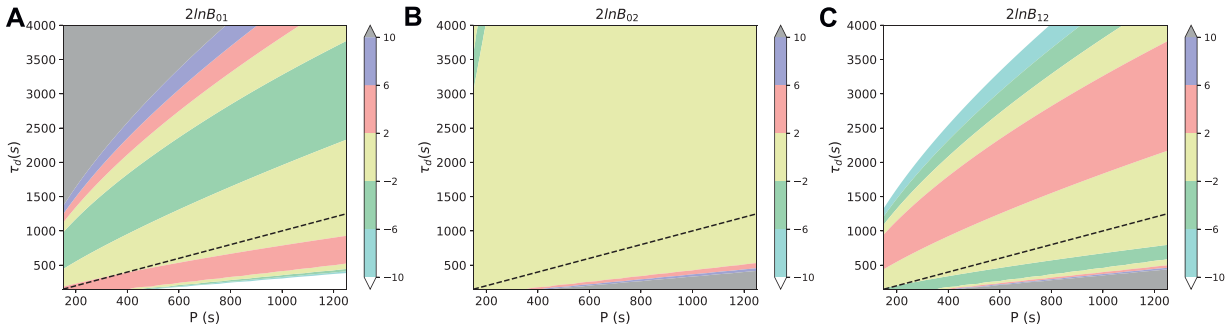

FIGURE 6 | Bayes factors in the one-to-one comparison between resonant absorption, phase mixing, and wave leakage mechanisms as a function of the observables period and damping time with uncertainties of 10% for each. The dashed lines indicate $\tau_d = P$. The different levels of evidence are indicated in the colour bars. Not Worth a bare mention (NWM, yellow), Positive (PE, green/red), Strong (SE, blue/purple), Very strong (VSE, white/grey). Adapted from Montes-Solís and Arregui (2017).

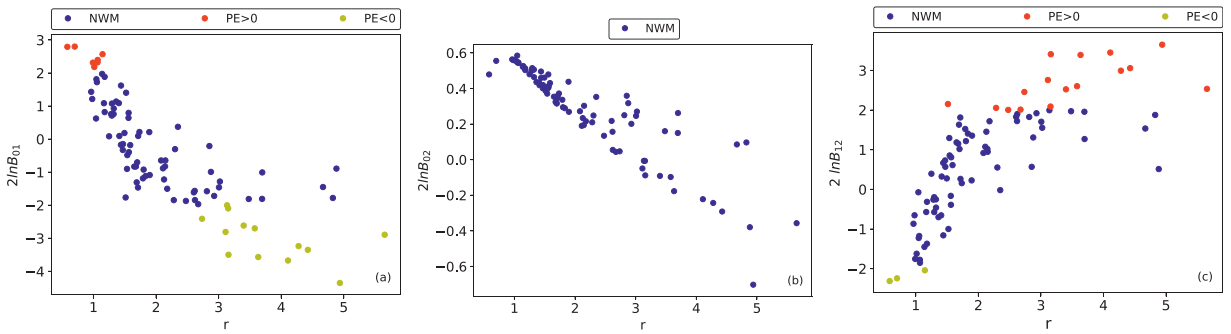

FIGURE 7 | Representation of the Bayes factors computed for the 89 events selected from Verwichte et al. (2013) and Goddard et al. (2016). The different panels correspond to the three one-to-one comparisons between resonant absorption, phase mixing, and wave leakage, here represented with the subscripts 0, 1, and 2 respectively. Adapted from Montes-Solís and Arregui (2017).

Tomczyk et al. (2007) and, although first interpreted as Alfvén waves, theoretical arguments by Goossens et al. (2012) showed that an interpretation in terms of kink waves damped by resonant absorption offers a more accurate description. The observed waves show signatures of *in situ* wave damping in the form of a discrepancy between the outward and the inward wave power. This led Verth et al. (2010) to produce a theoretical model connecting the average power ratio for inward and outward propagating waves with their damping rate. This expression is

$$\langle P(f)\rangle_{\rm ratio} = R_0 \exp\left(\frac{2L}{v_{\rm ph}\xi_{\rm E}} f\right), \qquad (12)$$

with $R_0 = P_{\rm out}(f)/P_{\rm in}(f)$ the ratio of powers generated at the two foot-points. The exponential factor contains wave propagation and damping properties: the wave travel time along the full wave path of length L, $2L/v_{\rm ph}$, the frequency f and the damping ratio, $\xi_{\rm E}$. In the absence of damping, $\xi_{\rm E} \to \infty$ and $\langle P(f)\rangle_{\rm ratio} = R_0$, thus the average power ratio equals the ratio of powers at the two foot-points.

The model by Verth et al. (2010) predicts an exponential dependence of the average power ratio with frequency. A least squares fit to a set of data from the Coronal Multi-channel Polarimeter (CoMP) performed by these authors shows a good qualitative agreement and enabled them to infer a value for the damping ratio $\xi_{\rm E}$. However, one must bear in mind that a fitting procedure consists of adopting a model M and obtaining the set of so-called best fit parameters θ. As explained above, Bayesian inference aims at obtaining a solution in terms of a probability distribution of the parameters conditional on the model and on the data, $p(\theta|M, D)$. There is no room for absolute statements concerning model evidence because the evidence in favour of a model is always relative to the evidence in favour of another.

Montes-Solís and Arregui (2020) performed a Bayesian analysis to quantify the evidence in favour of resonant damping using CoMP. Instead of considering an alternative damping mechanism the focus was on trying to quantify the evidence in favour of resonant damping in front of the other possible source of discrepancy between the inward and outward power ratio in the corona, namely, an asymmetry in the wave power ratio at the foot-points, i.e., $R_0 \neq 1$ in **Eq. 12**. We note that if $P_{\rm out}(f) > P_{\rm in}(f)$, foot-point driving asymmetry will increase the contribution of resonant damping. Conversely, for $P_{\rm out}(f) < P_{\rm in}(f)$, the asymmetry will decrease the contribution of resonant damping to the average power ratio.

In their analysis, Montes-Solís and Arregui (2020) first consider the inference of the two parameters of interest, $\xi_{\rm E}$

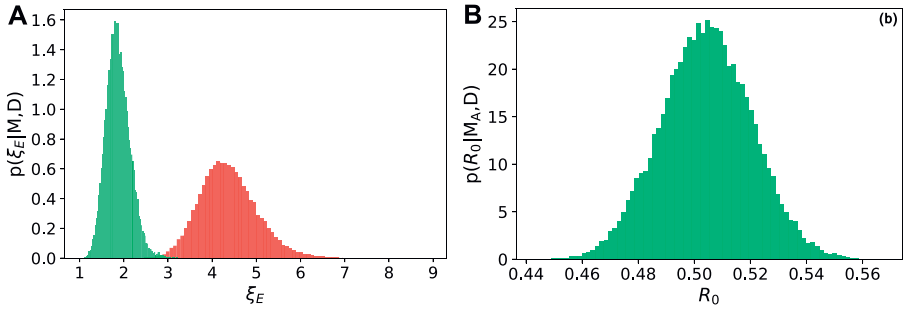

FIGURE 8 | (A) Marginal posterior distributions for ξ_E conditional on data D and models M_R (red) and M_A (green). **(B)** Marginal posterior distribution for R_0 conditional on data D and model M_A. The data D consist of the collective use of the CoMP data set analysed by Verth et al. (2010). The posterior summaries are $\hat{\xi}_E^{MAP} = 4.3^{+0.7}_{-0.6}$ for $p(\xi_E|M_R, D)$, $\hat{\xi}_E^{MAP} = 1.9^{+0.3}_{-0.2}$ for $p(\xi_E|M_A, d)$, and $\hat{R}_0^{MAP} = 0.5 \pm 0.02$ for $p(R_0|M_A, D)$, with uncertainty given at the 68% credible interval. The posteriors are computed by Markov Chain Monte Carlo (MCMC) sampling using the Python emcee package (Foreman-Mackey et al., 2013). Adapted from Montes-Solís and Arregui (2020).

and R_0, from a set of CoMP data points for average power ratio as a function of frequency in the range 0.05–4 mHz. The resulting marginal posteriors are shown in **Figure 8**. The set of CoMP observations is equally well explained by a reduced model, M_R, with parameter distribution $p(\xi_E|M_R, D)$ that considers resonant damping as the sole contributor to the average power ratio and by the larger model, M_A, with parameter distributions $p(\xi_E|M_A, D)$ and $p(R_0|M_A, D)$, which additionally considers foot-point asymmetry. The full posterior for R_0 is within the region below one. This means that $P_{out} < P_{in}$ and there is asymmetry in the power generated at both foot-points. The corresponding inference for ξ_E is shifted towards the region corresponding to stronger damping (smaller values of ξ_E) to counterbalance the decreasing factor due to the asymmetry at the foot-points.

Observations can therefore be equally well explained by two models, with or without foot-point asymmetry. To quantify the relative merit of the two explanations, Montes-Solís and Arregui (2020) perform model comparison using the Bayes factor

$$B_{RA} = \frac{p(\mathcal{D}|M_R)}{p(\mathcal{D}|M_A)}. \tag{13}$$

The Bayes factor is computed in the two-dimensional plane of synthetic data $\mathcal{D}$, covering the full ranges in frequency and average power ratio of CoMP observations, $\mathcal{D} = (f, \langle P(f)\rangle_{ratio})$. To this end, **Eq. 12** is used to generate theoretical predictions over a grid of points in f and $\langle P(f)\rangle_{ratio}$.

Figure 9 shows the distribution of Bayes factor values over the two-dimensional synthetic data space. It is clear that the evidence distribution is inhomogeneous and three different regions, can be identified. They are delimited by the boundaries where the marginal likelihoods are equal and therefore the Bayes factor is zero. In the central region, within the solid boundary lines, model M_R is in principle more plausible than model M_A, because the marginal likelihood for this model is larger. The level of relative plausibility depends on the Bayes factor value, B_{RA}. In the white area, the evidence in favour of M_R is inconclusive and then varies from positive to very strong in the blue to green areas. Above and below the solid lines, model M_A is more plausible, because the marginal likelihood for this model is larger. However, based on the numerical value for the Bayes factor B_{AR}, the evidence is inconclusive in the white areas and varies from positive to very strong as we move further towards the upper and lower areas in the plane of observables.

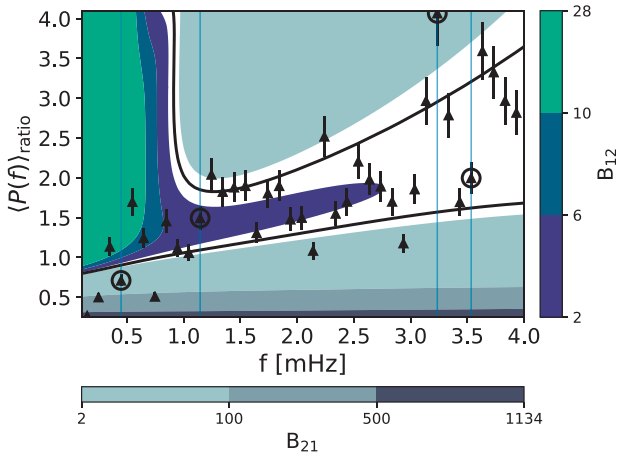

FIGURE 9 | Filled contour with the distribution of Bayes factors, B_{RA} and B_{AR}, over the two-dimensional data space $\mathcal{D}$. Solid lines connect points with $p(\mathcal{D}|M_R) = p(\mathcal{D}|M_A)$ (Bayes factor zero). The computations are performed over a grid of points ($N_f = 80$, $N_{\langle P(f)\rangle_{ratio}} = 155$) over the ranges $f \in [0.05, 4]$ and $\langle P(f)\rangle_{ratio} \in [0.25, 4.1]$. The priors are $p(\xi_E) \sim \mathcal{G}(1.9, 0.3)$ for M_R and $p(\xi_E) \sim \mathcal{G}(4.3, 0.6)$; $p(R_0) \sim \mathcal{G}(0.5, 0.02)$ for M_A. Triangles represent CoMP data. Following Kass and Raftery (1995), the evidence in favour of a model i in front of an alternative j is inconclusive for values of $2\log(B_{ij})$ from 0 to 2; positive from 2 to 6; strong from 6 to 10; and very strong for values above 10. Adapted from Montes-Solís and Arregui (2020).

Superimposed over the distribution of Bayes factors in **Figure 9** are observed CoMP data with assumed error bars. We can see that in most of the cases, data fall over regions where the marginal likelihood for model M_R is larger. A fraction of them are located over areas where the evidence in favour of model M_R is conclusive. Interestingly, some of them fall into the two regions where the marginal likelihood for model M_A is larger. In some cases, they are over areas where the evidence supports model M_A in front of model M_R.

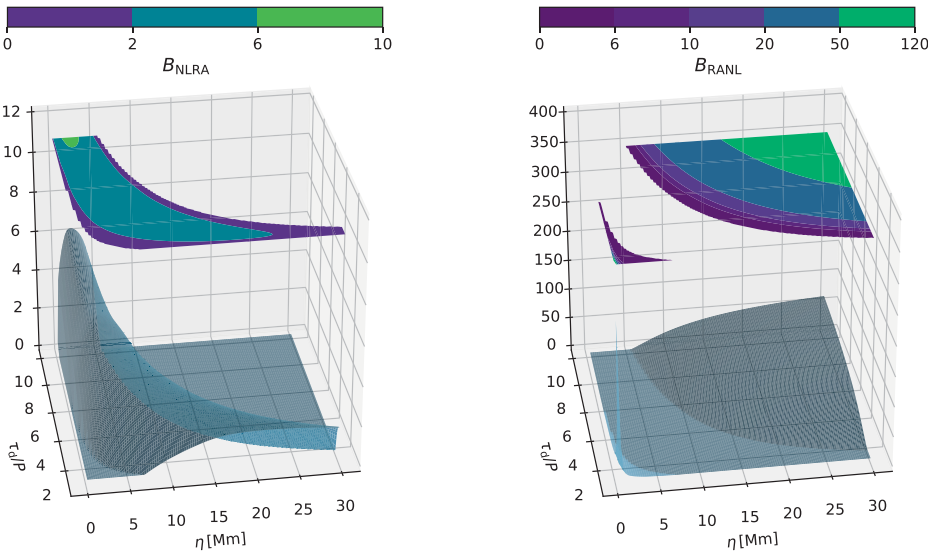

FIGURE 10 | Surface and filled contour representations of the Bayes factors B_{NLRA} (left panel) and B_{RANL} (right panel) in the synthetic data space $\mathcal{D} = (\eta, \tau_d/P)$. The Bayes factors were computed using the marginal likelihood calculations for each model. Adapted from Arregui (2021).

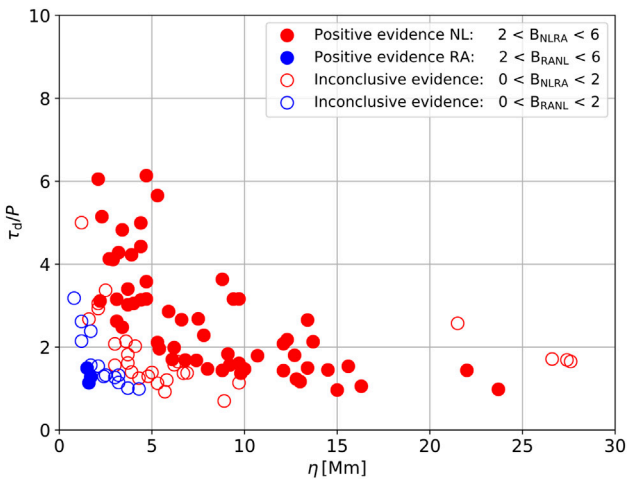

FIGURE 11 | Scatter plot of the 101 loop oscillation events in the Nechaeva et al. (2019) catalog with information about both the oscillation amplitude and the damping ratio. In the Bayesian evidence analysis by Arregui (2021), colour filled circles represent cases with conclusive evidence. Edge coloured circles represent cases with either $p(\mathcal{D}|M_{NL}) > p(\mathcal{D}|M_{RA})$ or vice versa, yet inconclusive evidence. Red is for M_{NL} and blue for M_{RA}. Adapted from Arregui (2021).

These results indicate that CoMP measurements of integrated average power ratio for propagating coronal waves cannot exclude an explanation in terms of asymmetry in the wave power generated at the foot-points. Some observations are equally or even better explained by larger models with foot-point wave power asymmetry than by the reduced models with identical power at the two foot-points and resonant damping as the only contributor to the observed average power ratio.

3.5 Evidence for a Nonlinear Damping Model for Waves in the Corona

Recent observational and theoretical studies have shown that the damping of transverse loop oscillations depends on the oscillation amplitude (Goddard et al., 2016; Magyar and Van Doorsselaere, 2016). The increase in the number and quality of observations has led to the creation of catalogs with a large number of events (Anfinogentov et al., 2015; Goddard et al., 2016; Nechaeva et al., 2019; Tiwari et al., 2021). When the damping time over the period is plotted against the oscillation amplitude, the data are scattered forming a cloud with a triangular shape (see e.g., Figure 6 in Nechaeva et al., 2019). In general, larger amplitudes correspond to smaller damping ratio values and vice versa. In a recent study, Arregui (2021) considered the mechanisms of linear resonant absorption, in the formulation given by Ruderman and Roberts (2002) and Goossens et al. (2002) and of nonlinear damping, in the formulation given by Van Doorsselaere et al. (2021). Their analytical developments provide us with two analytical expressions for the damping ratio in cylindrically symmetric waveguides.

For linear resonant absorption, model M_{RA}, the damping ratio is given by

$$\frac{\tau_d}{P}\|_{M_{RA}} = \mathcal{F}\,\frac{\zeta+1}{\zeta-1}\,\frac{R}{l}, \tag{14}$$

with $\zeta = \rho_i/\rho_e$ the ratio of internal to external density, l/R the length of the non-uniform layer at the boundary of the waveguide with radius R, and $\mathcal{F} = 2/\pi$ for a sinusoidal variation of density over the non-uniform layer. The predictions from the damping model M_{RA} given by **Eq. 14** for the observable damping ratio are determined by the parameter vector $\theta_{RA} = \{\zeta, l/R\}$.

For the nonlinear damping of standing kink waves, due to the energy transfer to small scales in the radial and azimuthal directions, the damping ratio is given by

$$\frac{\tau_{\rm d}}{P}\|_{M_{\rm NL}} = 40\sqrt{\pi}\,\frac{1}{2\pi a}\,\frac{1+\zeta}{\sqrt{\zeta^2-2\zeta+97}}, \tag{15}$$

with $a = \eta/R$ the ratio of the displacement η to the loop radius. The predictions from the damping model $M_{\rm NL}$ given by **Eq. 15** for the observable damping ratio, for known oscillation amplitude, are determined by the parameter vector $\theta_{\rm NL} = \{R, \zeta\}$.

Theoretical predictions from these two models can be confronted by computing the marginal likelihood of the data in the plane of observables defined by the damping ratio and the oscillation amplitude, $\mathcal{D} = \{\eta, \tau_{\rm d}/P\}$. The ratio of marginal likelihoods leads to the Bayes factor distributions over $\mathcal{D}$-space shown in **Figure 10**. The two panels show that there is a clear separation between the regions over synthetic data space over which evidence in favour of one or the other model dominates. The evidence supports the nonlinear damping model in a particular region corresponding to combinations with small amplitude and large damping ratio values in the upper-left region of the plane and extending towards the lower-right region corresponding to combinations with smaller damping ratio and larger oscillation amplitude values in a broader range. On the other hand, the evidence supports resonant damping in two regions. The first one extends towards the right-hand side of the domain. The second consists of a small region corresponding to combinations of very small amplitude and strong damping. Overall, the observed data fall within the regions with the largest Bayes factor values for the nonlinear damping model.

The analysis using synthetic data over prescribed ranges for the observable amplitude and damping ratio offers a birds-eye view of the distribution of the evidence. The application to observed data offers a better informed result on the level of evidence for or against each damping model. The catalog by Nechaeva et al. (2019) contains 223 loop oscillating loops observed with SDO/AIA in the period 2010–2018. In 101 cases, they contain information about the damping and the oscillation amplitude. Arregui (2021) applied a Bayesian evidence analysis to these data to assess the strength of the evidence for nonlinear damping relative to that for resonant absorption.

Figure 11 displays the results obtained for all 101 cases, regardless of the conclusive or inconclusive nature of the evidence. The red colour indicates evidence in favour of nonlinear damping. The full red dots indicate positive evidence. The edge coloured circles are cases with marginal likelihood for nonlinear damping larger than the marginal likelihood for resonant damping, but inconclusive evidence because the Bayes factor is below 2. The blue colour indicates evidence in favour of resonant damping. The full blue dots indicate positive evidence. The edge coloured circles are cases with marginal likelihood for resonant damping larger than the marginal likelihood for nonlinear damping, but inconclusive evidence because the Bayes factor is below 2. The marginal likelihood in favour of nonlinear damping is larger in the majority of cases. The events with conclusive evidence for nonlinear damping largely outnumber those in favour of linear resonant absorption. The evidence for the nonlinear damping model relative to linear resonant absorption is therefore appreciable to a reasonable degree of Bayesian certainty.

4 SUMMARY

Bayesian analysis tools are increasingly being used in seismology of the solar corona. In parameter inference, they led to the inference of relevant information on the structure of coronal loops or prominence plasmas, such as the magnetic field strength or the plasma density. Model comparison techniques have been used to assess the damping mechanism operating in coronal loop oscillations. In a comparison between a particular linear and a particular nonlinear damping mechanism, the latter seems to be more plausible in explaining observations. Note that we might have left out important alternative physical processes that could be more plausible instead. This could be assessed by performing additional one-to-one comparisons. Because of our inability to directly measure the physical conditions in the structures of interest, the Bayesian approach offers the best solution to inference problems under uncertain and incomplete information. It uses principled ways to combine the information from data, theoretical models and previous knowledge. The grow in the number of dedicated computing tools to sample multidimensional posterior and marginal likelihood spaces will enable us to apply these methods to additional phenomena related to the structure, dynamics and heating of the solar atmosphere.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

Funding provided under project PGC2018-102108-B-I00 from Ministerio de Ciencia, Innovación y Universidades (Spain) and FEDER funds.

ACKNOWLEDGMENTS

I am grateful to Andrés Asensio Ramos, Marcel Goossens, and María Montes-Solís for years of fruitful collaboration. Calculations and figures were implemented with `numpy` (Harris et al., 2020) and `matplotlib` (Hunter, 2007). This research has made use of NASA's Astrophysics Data System Bibliographic Services.

REFERENCES

Anfinogentov, S. A., Antolin, P., Inglis, A. R., Kolotkov, D., Kupriyanova, E. G., McLaughlin, J. A., et al. (2021a). Novel Data Analysis Techniques in Coronal Seismology. *arXiv e-prints*. arXiv:2112.13577.

Anfinogentov, S. A., Nakariakov, V. M., and Nisticò, G. (2015). Decayless Low-Amplitude Kink Oscillations: a Common Phenomenon in the Solar corona? *Astron. Astrophysics* 583, A136. doi:10.1051/0004-6361/201526195

Anfinogentov, S. A., Nakariakov, V. M., Pascoe, D. J., and Goddard, C. R. (2021b). Solar Bayesian Analysis Toolkit-A New Markov Chain Monte Carlo IDL Code for Bayesian Parameter Inference. *ApJS* 252, 11. doi:10.3847/1538-4365/abc5c1

Arregui, I., and Asensio Ramos, A. (2011). Bayesian Magnetohydrodynamic Seismology of Coronal Loops. *Astrophysical J.* 740, 44. doi:10.1088/0004-637x/740/1/44

Arregui, I., and Asensio Ramos, A. (2014). Determination of the Cross-Field Density Structuring in Coronal Waveguides Using the Damping of Transverse Waves. *Astron. Astrophysics* 565, A78. doi:10.1051/0004-6361/201423536

Arregui, I., Asensio Ramos, A., and Díaz, A. J. (2013a). Bayesian Analysis of Multiple Harmonic Oscillations in the Solar Corona. *Astrophysical J.* 765, L23. doi:10.1088/2041-8205/765/1/l23

Arregui, I., Asensio Ramos, A., and Pascoe, D. J. (2013b). Determination of Transverse Density Structuring from Propagating Magnetohydrodynamic Waves in the Solar Atmosphere. *Astrophysical J.* 769, L34. doi:10.1088/2041-8205/769/2/l34

Arregui, I., Ballester, J. L., and Goossens, M. (2008). On the Scaling of the Damping Time for Resonantly Damped Oscillations in Coronal Loops. *Astrophysical J.* 676, L77–L80. doi:10.1086/587098

Arregui, I. (2018). Bayesian Coronal Seismology. *Adv. Space Res.* 61, 655–672. doi:10.1016/j.asr.2017.09.031

Arregui, I. (2021). Bayesian Evidence for a Nonlinear Damping Model for Coronal Loop Oscillations. *ApJL* 915, L25. doi:10.3847/2041-8213/ac0d53

Arregui, I., and Goossens, M. (2019). No Unique Solution to the Seismological Problem of Standing Kink Magnetohydrodynamic Waves. *Astron. Astrophysics* 622, A44. doi:10.1051/0004-6361/201833813

Arregui, I., Montes-Solís, M., and Asensio Ramos, A. (2019). Inference of Magnetic Field Strength and Density from Damped Transverse Coronal Waves. *Astron. Astrophysics* 625, A35. doi:10.1051/0004-6361/201834324

Arregui, I., Soler, R., and Asensio Ramos, A. (2015). Model Comparison for the Density Structure across Solar Coronal Waveguides. *Astrophysical J.* 811, 104. doi:10.1088/0004-637X/811/2/104

Arregui, I., and Soler, R. (2015). Model Comparison for the Density Structure along Solar Prominence Threads. *Astron. Astrophysics* 578, A130. doi:10.1051/0004-6361/201525720

Arregui, I. (2015). Wave Heating of the Solar Atmosphere. *Phil. Trans. R. Soc. A.* 373, 20140261. doi:10.1098/rsta.2014.0261

Aschwanden, M. J., Fletcher, L., Schrijver, C. J., and Alexander, D. (1999). Coronal Loop Oscillations Observed with theTransition Region and Coronal Explorer. *Astrophysical J.* 520, 880–894. doi:10.1086/307502

Asensio Ramos, A., and Arregui, I. (2013). Coronal Loop Physical Parameters from the Analysis of Multiple Observed Transverse Oscillations. *Astron. Astrophysics* 554, A7. doi:10.1051/0004-6361/201321428

Asensio Ramos, A., Martínez González, M. J., and Rubiño-Martín, J. A. (2007). Bayesian Inversion of Stokes Profiles. *Astron. Astrophysics* 476, 959–970. doi:10.1051/0004-6361:20078107

Bayes, M., and Price, M. (1763). An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *R. Soc. Lond. Philos. Trans. Ser.* 53, 370–418.

Cally, P. S. (2003). Coronal Leaky Tube Waves and Oscillations Observed with Trace. *Solar Phys.* 217, 95–108. doi:10.1023/A:1027326916984

De Pontieu, B., Martens, P. C. H., and Hudson, H. S. (2001). Chromospheric Damping of Alfven Waves. *Astrophysical J.* 558, 859–871. doi:10.1086/322408

Duckenfield, T. J., Goddard, C. R., Pascoe, D. J., and Nakariakov, V. M. (2019). Observational Signatures of the Third Harmonic in a Decaying Kink Oscillation of a Coronal Loop. *Astron. Astrophysics* 632, A64. doi:10.1051/0004-6361/201936822

Foreman-Mackey, D., Hogg, D. W., Lang, D., and Goodman, J. (2013). Emcee: The MCMC Hammer. *Publications Astronomical Soc. Pac.* 125, 306–312. doi:10.1086/670067

Gates, E., Krauss, L. M., and White, M. (1995). Treating Solar Model Uncertainties: A Consistent Statistical Analysis of Solar Neutrino Models and Data. *Phys. Rev. D* 51, 2631–2643. doi:10.1103/PhysRevD.51.2631

Goddard, C. R., Antolin, P., and Pascoe, D. J. (2018). Evolution of the Transverse Density Structure of Oscillating Coronal Loops Inferred by Forward Modeling of EUV Intensity. *Astrophysical J.* 863, 167. doi:10.3847/1538-4357/aad3cc

Goddard, C. R., Nisticò, G., Nakariakov, V. M., and Zimovets, I. V. (2016). A Statistical Study of Decaying Kink Oscillations Detected Using SDO/AIA. *Astron. Astrophysics* 585, A137. doi:10.1051/0004-6361/201527341

Goddard, C. R., Pascoe, D. J., Anfinogentov, S., and Nakariakov, V. M. (2017). A Statistical Study of the Inferred Transverse Density Profile of Coronal Loop Threads Observed with Sdo/aia. *Astron. Astrophysics* 605, A65. doi:10.1051/0004-6361/201731023

Goossens, M., Andries, J., and Arregui, I. (2006). Damping of Magnetohydrodynamic Waves by Resonant Absorption in the Solar Atmosphere. *Phil. Trans. R. Soc. A.* 364, 433–446. doi:10.1098/rsta.2005.1708

Goossens, M., Andries, J., and Aschwanden, M. J. (2002). Coronal Loop Oscillations. *Astron. Astrophysics* 394, L39–L42. doi:10.1051/0004-6361:20021378

Goossens, M., Andries, J., Soler, R., Van Doorsselaere, T., Arregui, I., and Terradas, J. (2012). Surface Alfvén Waves in Solar Flux Tubes. *Astrophysical J.* 753, 111. doi:10.1088/0004-637X/753/2/111

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array Programming with NumPy. *Nature* 585, 357–362. doi:10.1038/s41586-020-2649-2

Heyvaerts, J., and Priest, E. R. (1983). Coronal Heating by Phase-Mixed Shear Alfven Waves. *Astron. Astrophys.* 117, 220.

Hunter, J. D. (2007). Matplotlib: A 2d Graphics Environment. *Comput. Sci. Eng.* 9, 90–95. doi:10.1109/MCSE.2007.55

Kass, R. E., and Raftery, A. E. (1995). Bayes Factors. *J. Am. Stat. Assoc.* 90, 773–795. doi:10.1080/01621459.1995.10476572

Lin, Y., Soler, R., Engvold, O., Ballester, J. L., Langangen, Ø., Oliver, R., et al. (2009). Swaying Threads of a Solar Filament. *Astrophysical J.* 704, 870–876. doi:10.1088/0004-637x/704/1/870

Magyar, N., and Van Doorsselaere, T. (2016). Damping of Nonlinear Standing Kink Oscillations: a Numerical Study. *Astron. Astrophysics* 595, A81. doi:10.1051/0004-6361/201629010

Marsh, M. S., Ireland, J., and Kucera, T. (2008). Bayesian Analysis of Solar Oscillations. *Astrophysical J.* 681, 672–679. doi:10.1086/588751

Montes-Solís, M., and Arregui, I. (2017). Comparison of Damping Mechanisms for Transverse Waves in Solar Coronal Loops. *Astrophysical J.* 846, 89. doi:10.3847/1538-4357/aa84b7

Montes-Solís, M., and Arregui, I. (2019). Inferring Physical Parameters in Solar Prominence Threads. *Astron. Astrophysics* 622, A88. doi:10.1051/0004-6361/201834406

Montes-Solís, M., and Arregui, I. (2020). Quantifying the Evidence for Resonant Damping of Coronal Waves with Foot-point Wave Power Asymmetry. *Astron. Astrophysics* 640, L17. doi:10.1051/0004-6361/201937237

Nakariakov, V. M., Ofman, L., DeLuca, E. E., Roberts, B., and Davila, J. M. (1999). TRACE Observation of Damped Coronal Loop Oscillations: Implications for Coronal Heating. *Science* 285, 862–864. doi:10.1126/science.285.5429.862

Nakariakov, V. M., and Ofman, L. (2001). Determination of the Coronal Magnetic Field by Coronal Loop Oscillations. *Astron. Astrophysics* 372, L53–L56. doi:10.1051/0004-6361:20010607

Nechaeva, A., Zimovets, I. V., Nakariakov, V. M., and Goddard, C. R. (2019). Catalog of Decaying Kink Oscillations of Coronal Loops in the 24th Solar Cycle. *ApJS* 241, 31. doi:10.3847/1538-4365/ab0e86

Ofman, L., and Aschwanden, M. J. (2002). Damping Time Scaling of Coronal Loop Oscillations Deduced from [ITAL]Transition Region and Coronal Explorer [/ITAL] Observations. *Astrophys. J. Lett.* 576, L153–L156. doi:10.1086/343886

Pascoe, D. J., Anfinogentov, S. A., Goddard, C. R., and Nakariakov, V. M. (2018). Spatiotemporal Analysis of Coronal Loops Using Seismology of Damped Kink Oscillations and Forward Modeling of EUV Intensity Profiles. *Astrophysical J.* 860, 31. doi:10.3847/1538-4357/aac2bc

Pascoe, D. J., Anfinogentov, S., Nisticò, G., Goddard, C. R., and Nakariakov, V. M. (2017a). Coronal Loop Seismology Using Damping of Standing Kink Oscillations by Mode Coupling. *Astron. Astrophysics* 600, A78. doi:10.1051/0004-6361/201629702

Pascoe, D. J., Goddard, C. R., Anfinogentov, S., and Nakariakov, V. M. (2017b). Coronal Loop Density Profile Estimated by Forward Modelling of EUV Intensity. *Astron. Astrophysics* 600, L7. doi:10.1051/0004-6361/201730458

Pascoe, D. J., Goddard, C. R., and Van Doorsselaere, T. (2020a). Oscillation and Evolution of Coronal Loops in a Dynamical Solar corona. *Front. Astron. Space Sci.* 7, 61. doi:10.3389/fspas.2020.00061

Pascoe, D. J., Hood, A. W., and Van Doorsselaere, T. (2019). Coronal Loop Seismology Using Standing Kink Oscillations with a Lookup Table. *Front. Astron. Space Sci.* 6, 22. doi:10.3389/fspas.2019.00022

Pascoe, D. J., Russell, A. J. B., Anfinogentov, S. A., Simões, P. J. A., Goddard, C. R., Nakariakov, V. M., et al. (2017c). Seismology of Contracting and Expanding Coronal Loops Using Damping of Kink Oscillations by Mode Coupling. *Astron. Astrophysics* 607, A8. doi:10.1051/0004-6361/201730915

Pascoe, D. J., Smyrli, A., and Van Doorsselaere, T. (2020b). Tracking and Seismological Analysis of Multiple Coronal Loops in an Active Region. *Astrophysical J.* 898, 126. doi:10.3847/1538-4357/aba0a6

Richardson, W. H. (1972). Bayesian-Based Iterative Method of Image Restoration*. *J. Opt. Soc. Am.* 62, 55. doi:10.1364/josa.62.000055

Roberts, B., Edwin, P. M., and Benz, A. O. (1984). On Coronal Oscillations. *Astrophysical J.* 279, 857. doi:10.1086/161956

Roberts, B. (2000). Waves and Oscillations in the corona - (Invited Review). *Solar Phys.* 193, 139–152. doi:10.1023/a:1005237109398

Ruderman, M. S., and Roberts, B. (2002). The Damping of Coronal Loop Oscillations. *Astrophysical J.* 577, 475–486. doi:10.1086/342130

Soler, R., Arregui, I., Oliver, R., and Ballester, J. L. (2010). Seismology of Standing Kink Oscillations of Solar Prominence fine Structures. *Astrophysical J.* 722, 1778–1792. doi:10.1088/0004-637x/722/2/1778

Soler, R., and Goossens, M. (2011). Kink Oscillations of Flowing Threads in Solar Prominences. *Astron. Astrophysics* 531, A167. doi:10.1051/0004-6361/201116536

Spruit, H. C. (1982). Propagation Speeds and Acoustic Damping of Waves in Magnetic Flux Tubes. *Sol. Phys.* 75, 3–17. doi:10.1007/BF00153456

Sturrock, P. A. (1973). Evaluation of Astrophysical Hypotheses. *Astrophysical J.* 182, 569–580. doi:10.1086/152165

Tiwari, A. K., Morton, R. J., and McLaughlin, J. A. (2021). A Statistical Study of Propagating MHD Kink Waves in the Quiescent Corona. *Astrophysical J.* 919, 74. doi:10.3847/1538-4357/ac10c4

Tomczyk, S., McIntosh, S. W., Keil, S. L., Judge, P. G., Schad, T., Seeley, D. H., et al. (2007). Alfve´n Waves in the Solar Corona. *Science* 317, 1192–1196. doi:10.1126/science.1143304

Uchida, Y. (1970). Diagnosis of Coronal Magnetic Structure by Flare-Associated Hydromagnetic Disturbances. *Pub. Astron. Soc. Jpn.* 22, 341.

Van Doorsselaere, T., Goossens, M., Magyar, N., Ruderman, M. S., and Ismayilli, R. (2021). Nonlinear Damping of Standing Kink Waves Computed with Elsässer Variables. *Astrophysical J.* 910, 58. doi:10.3847/1538-4357/abe630

Verth, G., Terradas, J., and Goossens, M. (2010). Observational Evidence of Resonantly Damped Propagating Kink Waves in the Solar Corona. *Astrophysical J.* 718, L102–L105. doi:10.1088/2041-8205/718/2/L102

Verwichte, E., Van Doorsselaere, T., White, R. S., and Antolin, P. (2013). Statistical Seismology of Transverse Waves in the Solar corona. *Astron. Astrophysics* 552, A138. doi:10.1051/0004-6361/201220456

Wheatland, M. S. (2004). A Bayesian Approach to Solar Flare Prediction. *Astrophysical J.* 609, 1134–1139. doi:10.1086/421261

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Astronomy and Space Sciences

METHODS
published: 24 March 2022
doi: 10.3389/fspas.2022.858990

Identification and Classification of Relativistic Electron Precipitation at Earth Using Supervised Deep Learning

Luisa Capannolo, Wen Li and Sheng Huang*

Center for Space Physics, Boston University, Boston, MA, United States

We show an application of supervised deep learning in space sciences. We focus on the relativistic electron precipitation into Earth's atmosphere that occurs when magnetospheric processes (wave-particle interactions or current sheet scattering, CSS) violate the first adiabatic invariant of trapped radiation belt electrons leading to electron loss. Electron precipitation is a key mechanism of radiation belt loss and can lead to several space weather effects due to its interaction with the Earth's atmosphere. However, the detailed properties and drivers of electron precipitation are currently not fully understood yet. Here, we aim to build a deep learning model that identifies relativistic precipitation events and their associated driver (waves or CSS). We use a list of precipitation events visually categorized into wave-driven events (REPs, showing spatially isolated precipitation) and CSS-driven events (CSSs, showing an energy-dependent precipitation pattern). We elaborate the ensemble of events to obtain a dataset of randomly stacked events made of a fixed window of data points that includes the precipitation interval. We assign a label to each data point: 0 is for no-events, 1 is for REPs and 2 is for CSSs. Only the data points during the precipitation are labeled as 1 or 2. By adopting a long short-term memory (LSTM) deep learning architecture, we developed a model that acceptably identifies the events and appropriately categorizes them into REPs or CSSs. The advantage of using deep learning for this task is meaningful given that classifying precipitation events by its drivers is rather time-expensive and typically must involve a human. After post-processing, this model is helpful to obtain statistically large datasets of REP and CSS events that will reveal the location and properties of the precipitation driven by these two processes at all *L* shells and MLT sectors as well as their relative role, thus is useful to improve radiation belt models. Additionally, the datasets of REPs and CSSs can provide a quantification of the energy input into the atmosphere due to relativistic electron precipitation, thus offering valuable information to space weather and atmospheric communities.

Keywords: electron precipitation, wave-particle interactions, current sheet scattering, space sciences, supervised classification, LSTM, deep learning, radiation belts

OPEN ACCESS

Edited by:
Olga Verkhoglyadova, NASA Jet Propulsion Laboratory (JPL), United States

Reviewed by:
Alexei V. Dmitriev, Lomonosov Moscow State University, Russia
Emilia Kilpua, University of Helsinki, Finland

****Correspondence:***
Luisa Capannolo
luisacap@bu.edu

Specialty section:
This article was submitted to Space Physics, a section of the journal Frontiers in Astronomy and Space Sciences

Received: *20 January 2022*
Accepted: *01 March 2022*
Published: *24 March 2022*

Citation:
Capannolo L, Li W and Huang S (2022) Identification and Classification of Relativistic Electron Precipitation at Earth Using Supervised Deep Learning. Front. Astron. Space Sci. 9:858990. doi: 10.3389/fspas.2022.858990

1 INTRODUCTION

The radiation belt environment is highly dynamic and it is governed by acceleration, transport and loss processes (e.g., Li and Hudson, 2019; Reeves et al., 2003). One of the loss mechanisms is electron precipitation (EP), which occurs when the conservation of the first adiabatic invariant is violated (e.g., Schulz and Lanzerotti, 1974; Horne and Thorne, 1998): electrons are no longer trapped by the Earth's magnetic field and fall into the upper atmosphere. Not only electron depletion is important in the radiation belt evolution in time and flux, but electron precipitation is also known to drive many atmospheric effects related to space weather. Multiple studies have indeed associated conductivity variations and atmospheric chemistry changes (potentially leading to ozone reduction) with electron precipitation (Robinson et al., 1987; Fytterer et al., 2015; Mironova et al., 2015; Tyssøy et al., 2016; Khazanov et al., 2018; Meraner and Shmidt, 2018; Yu et al., 2018; Duderstadt et al., 2021; Sinnhuber et al., 2021).

It is well understood that electron precipitation can occur as a result of interactions between plasma waves existing in the magnetosphere and the trapped electron population in the radiation belts (e.g., Millan and Thorne, 2007; Thorne, 2010). Electrons can also be lost if the magnetic field line around which they gyrate is stretched away from Earth or undergoes a significant geometry variation such that the curvature radius of the field line is comparable to the gyroradius of the electrons (e.g., Büchner and Zelenyi, 1989; Dubyagin et al., 2021; Sergeev et al., 1983, 1993). This process is called field line curvature scattering or current sheet scattering (CSS). Under these conditions, the field line no longer traps the electrons, and these electrons can precipitate into the atmosphere. The location where precipitation occurs (called isotropic boundary, IB) depends on electron energy (Capannolo et al., 2022; Yahnin et al., 2016; 2017). This phenomenon has also been widely studied for protons (Ganushkina et al., 2005; Gilson et al., 2012; Liang et al., 2014; Dubyagin et al., 2018).

A comprehensive understanding of which mechanism (waves or CSS) dominates the electron precipitation and thus the energy input into the Earth's atmosphere is still under active research. Given the Earth's magnetic field geometry, one would expect that on the dayside and at low L shells CSS does not contribute much, but more quantitative studies are still needed. Overall, while wave-driven precipitation can occur at all MLT (magnetic local time) sectors, CSS-driven precipitation is indeed primarily observed over 20–04 MLT (Yahnin et al., 2016; 2017), and overlaps with precipitation driven by waves (for the most part, electromagnetic ion cyclotron waves, EMIC) in the midnight sector (Capannolo et al., 2022).

These studies use data from the constellation of satellites called POES (Polar Orbiting Environmental Satellites) and MetOp (Meteorological Operational), described in **Section 2**. An example of a wave-driven (REP, relativistic electron precipitation) event is shown in **Figure 1A**, together with an example of a CSS-driven (CSS) event (**Figure 1B**). REP events show enhancements in the relativistic (>700 keV) precipitating electron flux (solid red line) and the precipitation is rather isolated (gray region) in space (L shell) with little/no precipitation around the main event. This region generally matches the location where the wave-particle interaction is efficient to violate the first adiabatic invariant. CSS events, instead, show an energy-dependent precipitation with higher energy electrons precipitating at lower L shells than lower energy electrons (**Figure 1B**; green, black, and blue solid lines). This is a direct result from the fact that the electron gyroradius depends on electron energy: higher energy electrons have a larger gyroradius, thus are lost by a stretched magnetic field line at distances closer to Earth (smaller L shells) than lower energy electrons. Given such a distinct pattern of precipitation, we can distinguish the precipitation drivers.

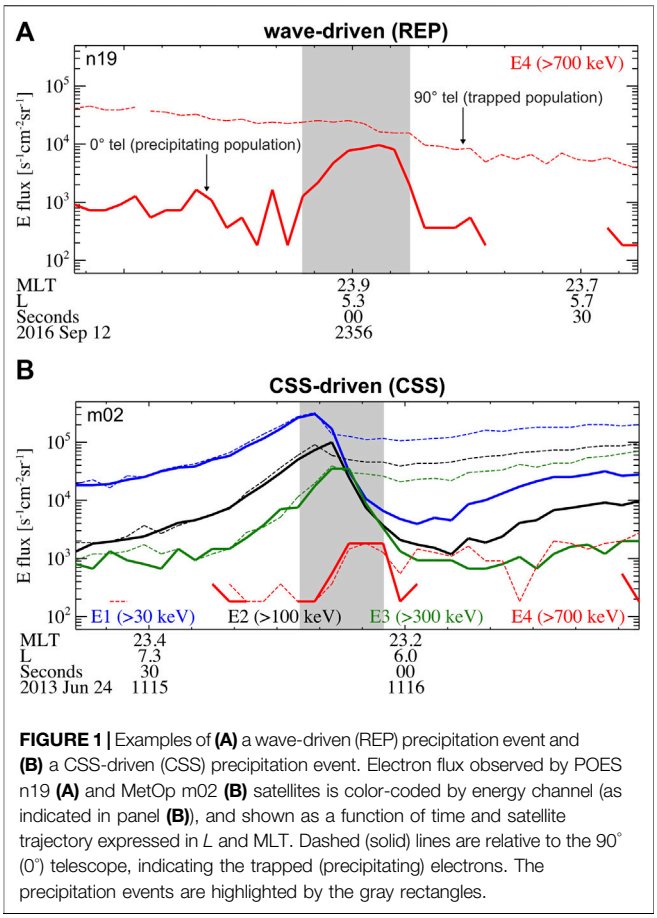

FIGURE 1 | Examples of **(A)** a wave-driven (REP) precipitation event and **(B)** a CSS-driven (CSS) precipitation event. Electron flux observed by POES n19 **(A)** and MetOp m02 **(B)** satellites is color-coded by energy channel (as indicated in panel **(B)**), and shown as a function of time and satellite trajectory expressed in L and MLT. Dashed (solid) lines are relative to the 90° (0°) telescope, indicating the trapped (precipitating) electrons. The precipitation events are highlighted by the gray rectangles.

So far, existing analyses aiming to distinguish the precipitation drivers have either focused on a limited time span (Yahnin et al., 2016; 2017) or on a limited MLT sector (Capannolo et al., 2022). Identifying precipitation events and visually inspecting their precipitation patterns to categorize their driver (waves or CSS) is a rather time-expensive task. Algorithms that find relativistic electron precipitation events (based on count rate or flux thresholds) exist in literature (e.g., Shekhar et al., 2017; Gasque et al., 2021; Capannolo et al., 2022), but they do not include the distinction between wave-driven precipitation and CSS-driven precipitation, which is a much more complex task to perform using algorithms. The goal of this work is to take

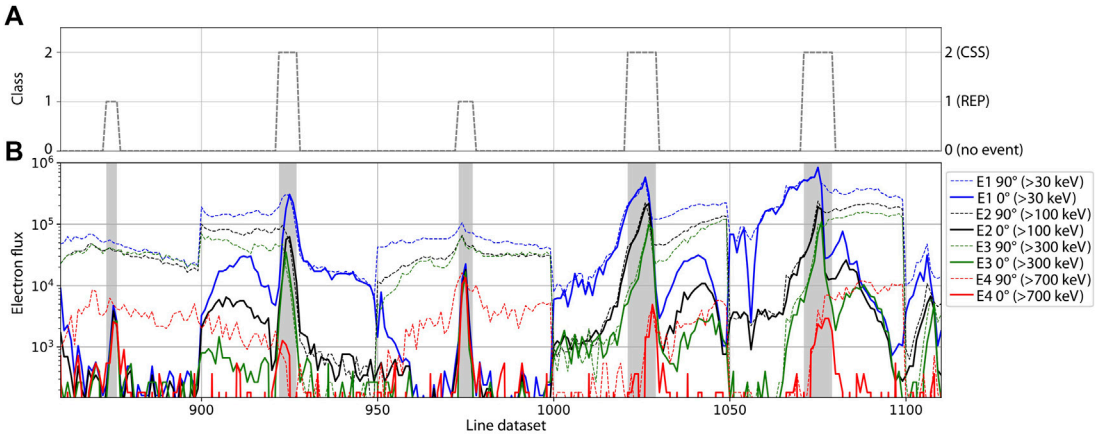

FIGURE 2 | Portion of the training dataset: **(A)** class of each data point and b) electron flux for different energies. Dashed and solid lines in panel **(B)** indicate the 90° and 0° telescope observations, respectively, as in **Figure 1**. Precipitation events are highlighted in gray in panel **(B)** and their relative class is shown in panel **(A)**, where class 0 indicates "no event", class 1 indicates "REP event" and class 2 indicates "CSS event".

advantage of deep learning techniques not only to find precipitation events, but also to categorize them into wave-driven (REP) and CSS-driven (CSS) events. We use the dataset of precipitation events analyzed in Capannolo et al. (2022), which were visually classified between wave-driven (REPs) and CSS-driven (CSSs) precipitation events (details in Capannolo et al., 2022). This work is an example of an application of supervised deep learning classification in space sciences that is able to provide a large dataset of precipitation events classified by driver (waves or CSS) after an initial manual classification of events.

2 SATELLITE DATA DESCRIPTION

We use data from the POES and MetOp network of sun-synchronous satellites in polar orbits at ~800–850 km of altitude (Evans and Greer, 2004). The Medium Energy Proton and Electron Detector (MEPED) provides electron (and proton) flux in three integral channels with cutoff energies of >30 keV (E1), >100 keV (E2), and >300 keV (E3) (Rodger et al., 2010). The P6 proton channel is designed to measure >6.9 MeV protons, however, it is also sensitive to electrons at >700 keV (Yando et al., 2011) in absence of high energy protons. Thus, we use the P6 channel as a fourth virtual electron channel, E4 (Green, 2013). Additionally, each satellite is equipped with two telescopes: one oriented along zenith (0° telescope) and one perpendicular to it (90° telescope), both with full field-of-view angle of 30°. At mid-to-high latitudes, the 0° telescope provides measurements of electrons precipitating deep into the loss cone and the 90° telescope provides observations of trapped electrons. Strong precipitation typically occurs when the flux observed by the 0° telescope approaches the flux observed by the 90° telescope, indicating that a large percentage of trapped electrons are precipitating. Precipitation events are marked in gray in **Figures 1–3**, and highlighted in brown (REP) and blue (CSS) in **Figure 4**. The resolution of the electron flux is 2 s, and the constellation of satellite covers a rather broad *L*-shell range and MLT sectors. Typical observations of POES/MetOp are shown in the **Supplementary Figure S1**. Each panel shows ¼ orbit of a POES/MetOp satellites (one pass through the radiation belts) and highlights the significant variability of flux during the satellite trajectory.

3 METHODS

In this section, we describe how we prepared the dataset of precipitation events in order to obtain a well-performing model. We also describe the model architecture and how it was decided, as well as how we trained the deep learning model.

3.1 Dataset Preparation

Capannolo et al. (2022) analyzed relativistic electron precipitation events observed by POES/MetOp from 2012 to 2020 over 22–02 MLT and classified these events between those driven by waves (called REP events in this work) from those driven by CSS (CSSs hereafter) using their characteristic precipitation profile (**Figure 1**). Note that this dataset was obtained after careful event classification: only events that clearly belonged to either category (REP or CSS) were considered, while ambiguous precipitation events were carefully discarded. More details on the classification are provided in Capannolo et al. (2022). In this work, we use this dataset of precipitation events classified over 22–02 MLT with additional preprocessing to improve the model performance as explained below.

Our goal is to build a dataset of precipitation events randomly stacked one after the other. We consider all four POES/MetOp electron channels and the two look directions (0° and 90°) for a total of eight inputs at a given time. The model output (or target)

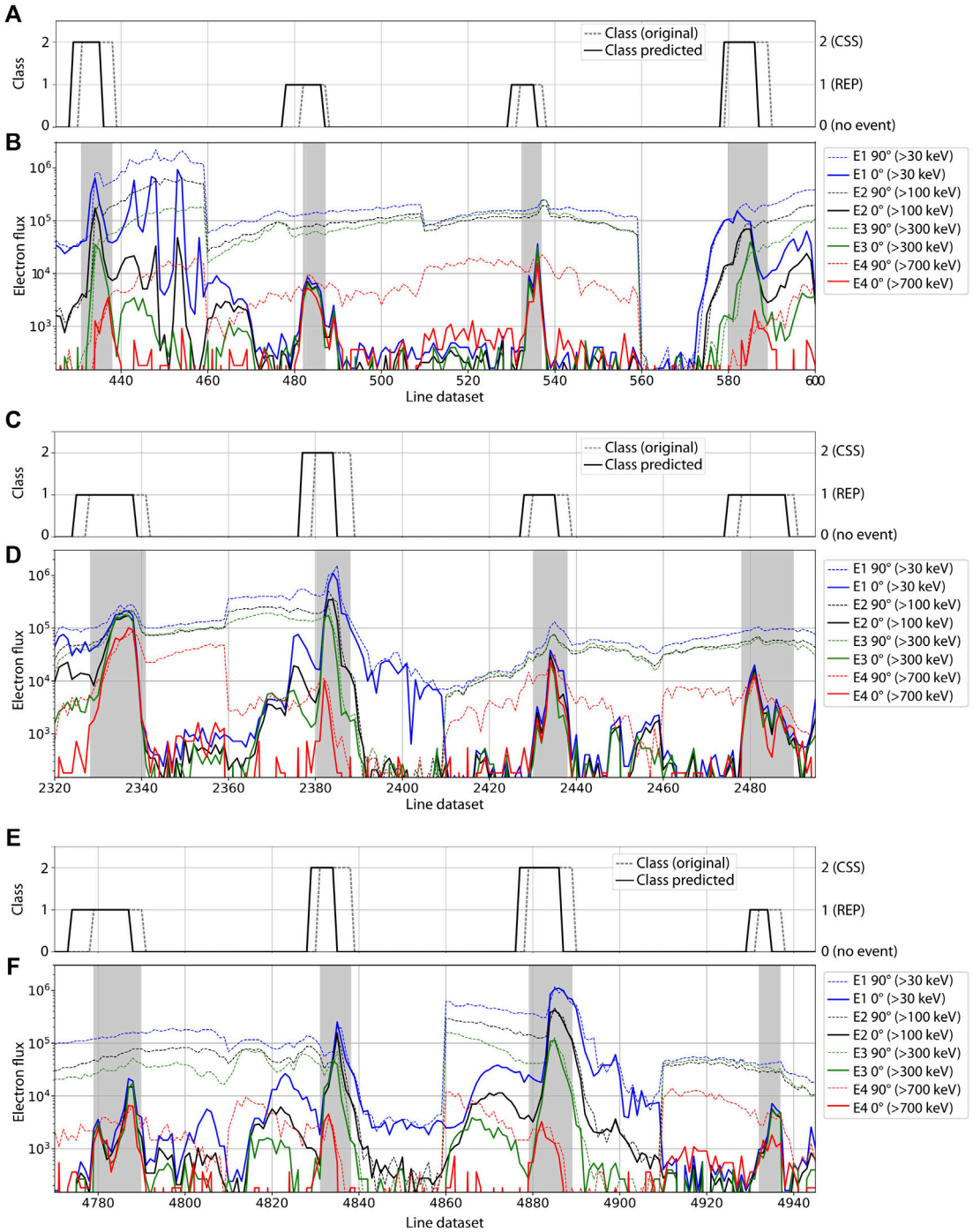

FIGURE 3 | Three different portions of the test dataset in a similar format as **Figure 2**. Panels **(A)**, **(C)** and **(E)** show the original class of each event in the dashed gray line and the class of each event predicted by the model in solid black. Panels **(B)**, **(D)** and **(F)** show the electron flux in a similar format as **Figure 2B**, where each event (originally identified) is highlighted in gray.

is the data point class (or label, used interchangeably hereafter): 0 is for no-event, 1 is for REP, and 2 is for CSS. Given one event, the data points are labeled as 1 or 2 during the precipitation (gray regions of **Figure 1**) and the adjacent data points (to the left and right of the event) are labeled with 0. Fluxes ≤0 for all channels are set to 0.01 (100) $s^{-1}cm^{-2}sr^{-1}$ for the 0° (90°) telescope measurements (negative values in POES/MetOp data indicate unreliable flux measurements). We apply the natural logarithm to the fluxes and normalize the whole dataset using the normalization parameters of the train dataset.

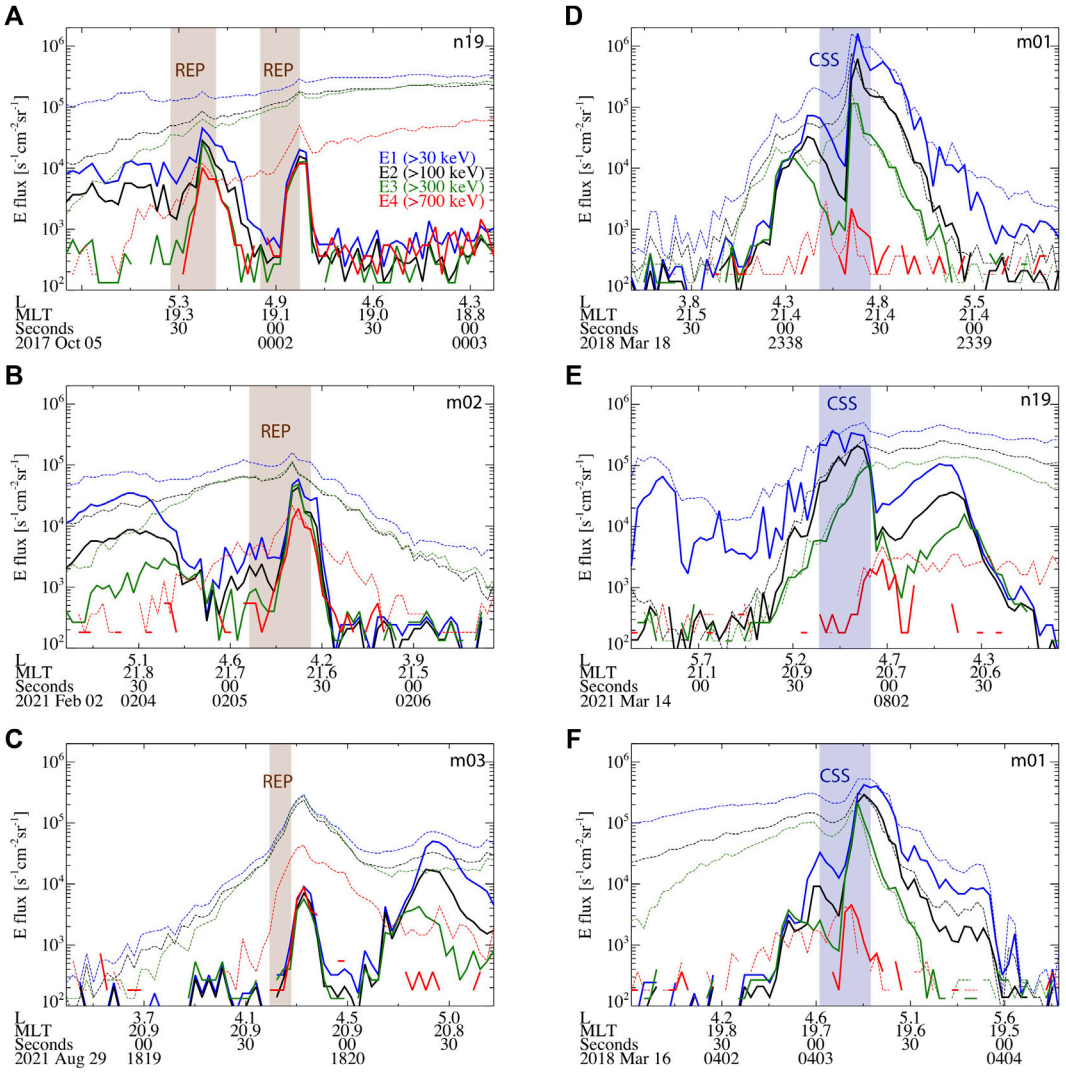

FIGURE 4 | Identification and classification of precipitation events on 6 days of POES/MetOp data. Each panel shows the electron flux color-coded in energy (legend in panel **(A)**) as a function of *L*, MLT, and time. Dashed (solid) lines indicate observations of trapped (precipitating) electrons from the 90° (0°) telescope. REP events identified by the model are highlighted in brown, while CSS events identified by the model are marked in blue.

As shown in **Supplementary Figure S1**, each pass through the radiation belts highlights a significant flux variability observed by POES/MetOp, while the precipitation events are rather short-lived (<30–60 s). As a result, if we use the full day of data when a given REP/CSS event occurs, we will obtain a label of mostly zeroes (no-event) and only a few data points at 1 or 2 (indicating the REP/CSS). This would make the full dataset of stacked events extremely imbalanced, where only a few percent of the labels are non-zero. With such dataset, the deep learning model is unable to perform well and it identifies only the no-events correctly. In order to overcome this obstacle, we consider a much shorter window of data for each event: given one event, we label the data points during precipitation with 1 or 2, but label with 0 only the data points adjacent to the left and right of the event such that the total number of data points is 50. In this way, we have windows of 50-point-long for each event which we stack one after the other in a random order. Additionally, we ensure that no other nearby events were occurring within the 50-point-long window such that in this window there is only one type of non-zero label (either 1 or 2). Note that if two events of different classes are adjacent to each other, we rule out both. Instead, if two REP events are adjacent to each other within the 50-point-long window, we widen the label of one to include both to ensure that in each 50-point-long window, there is only one continuous non-zero label. For the CSS events, we also manually extended the boundary of the precipitation events to include the full energy dispersion observed by POES/MetOp because we do not limit ourselves to the E4 precipitation alone (as done in Capannolo et al., 2022). This ensures that the full precipitation pattern (from low to high electron energy) is

identified as a CSS event and used to train the model. Using the boundaries as in Capannolo et al. (2022) worsens the model performance because the full extent of the energy-dependent pattern is not correctly learned by the model. We show a portion of the dataset in **Figure 2**: panel A) indicates the label and panel B) shows the electron flux for all energy channels and look directions, where the precipitation events are highlighted in gray.

In order to augment our dataset and provide the model with a wider variety of precipitation patterns, we also mirror each precipitation event about its main axis. This does not introduce data redundancy since each precipitation event (either mirrored or not) carries a meaningful information. In other words, a REP/CSS event can be directly observed by a POES/MetOp satellite following its actual trajectory (e.g., from low to high L shells), but the precipitation pattern would still be observed (though symmetrically) if the same POES/MetOp satellite was travelling along its opposite orbit (e.g., from high to low L shells) through the precipitation region at the same time. Note that this is possible since we are only interested in the profile of the precipitation (i.e., flux evolution as a function of dataset index) and not its temporal evolution. By using this methodology, we obtain a dataset of 460 REPs and 348 CSSs for a total dataset length of 40,400 data points. Although only ~20% of the data points are labeled with 1 or 2 (making this dataset still imbalanced with respect to the 0 class), the REP and CSS classes are approximately balanced (~10% data points are REPs and ~8% data points are CSSs) and the model is able to identify correctly no-events, REPs and CSSs as we show in the following sub-sections.

3.2 Model Structure and Training

We adapt a long short-term memory (LSTM; Hochreiter and Schmidhuber, 1997) architecture (a type of artificial recurrent neural network, RNN; Rumelhart et al., 1986) for the deep learning model because it retains input information at much earlier time steps, making it more efficiently than RNNs for problems that treat time series. As a matter of fact, the problem of our work is a time series classification. Although the time variable is not explicitly used, it is instead intrinsically represented by the shape of the precipitation. It is indeed the evolution of the precipitation pattern (isolated *vs* energy-dependent) that differentiates between the two drivers of precipitation, as mentioned in **Section 1**.

The input format required by LSTM is a tensor, which is composed of a stack of snapshots of the dataset identified by a sliding window with stride one and length 7. The label in each snapshot is assigned as the most probable one (*i.e.,* if the majority of data points have label of 0, the label assigned to that snapshot is also 0) and is *one-hot* encoded. The length of seven is set after trying different sliding window lengths and choosing the one that provided the best model performance.

The metrics we use are those of a standard classification problem and we focus on the F1 score (calculated as the weighted average of the precision and recall; it expresses how many events the classifier identifies correctly quantifying also how many are missed or mislabeled), the AUC (area under the ROC (Receiver Operating Characteristic) recall *vs* false-positive-rate curve) and the AUPRC (area under the precision *vs*. recall curve). We perform a k-fold cross validation with k = 10: the whole dataset is split into 10 portions of which one is used as a test set and the remaining nine are used as training set. We also consider a validation set that is 15% of the training set in each k-fold. The k-fold cross validation consists in training the model on k different datasets (described above) and estimating the model performance for each of the k iterations. The final model performance is the average of the k performances and the final model weights are obtained by training the model on the whole dataset (with the exception of 15% of the dataset used for testing purposes). During training, we use early stopping (with patience of 10 epochs) on the AUC calculated for the validation dataset.

4 MODEL PERFORMANCE

We tried different model configurations, all made of a LSTM layer followed by a fully connected (*i.e.,* dense) layer, ending with a dense layer of three neurons that outputs one predicted class. There are two dropout layers (with 0.5 dropout rate) after the LSTM layer and after the first dense layer. We validated each model configuration using the k-fold cross-validation (mentioned above) and we selected the model configuration with the highest F1 score, AUC and AUPRC. Out of all the configurations we tried (64 LSTM cells + 256 dense cells; 128 LSTM cells + 128 dense cells; 128 LSTM cells + 256 dense cells; 64 bidirectional LSTM cells + 256 dense cells; 64 bidirectional LSTM cells + 64 bidirectional LSTM cells + 128 dense cells) the model with the best performance is the one with a layer of 64 bidirectional LSTM cells followed by a fully connected layer of 256 cells (total number of free parameters is 71,171). The metrics resulting from the k-fold cross-validation for this model are: F1~0.948, AUC~0.995, and AUPRC~0.990. Note that the performance among the different model configurations is similar and differs only on the second or third decimal figure. **Supplementary Table S1** shows the performance scores (F1, AUC, AUPRC) resulting from the k-fold cross-validation for each architecture tested. As an example, **Supplementary Figure S2** (panels a–e) shows the metrics as a function of epoch for the k = 3 fold. Panel f) shows the confusion matrix averaged from all the confusion matrices of each k-fold: the highest values are focused along the diagonal, indicating that the model performs well in assigning the correct class to each snapshot.

To highlight that the model appropriately identifies and classifies precipitation events, we show in **Figure 3** three examples of how the model performs on three portions of the test dataset. Panels A), C), and E) present the model (solid) and original (dashed) labels and panels B), D), F) show the electron fluxes in a similar format as **Figure 2**. The precipitation events (originally assigned) are highlighted in gray and their associated class is reported in the panels A), C), E). Not only the model identifies all precipitation events, but each event is categorized in the class originally assigned. Note

that the indices where the labels are non-zero only indicate that nearby that region the probability of finding an event is higher than the probability of a no-event, but these indices do not necessarily represent the exact precipitation event boundaries (as the original class does). Nevertheless, the labels predicted by the model are in good agreement with the original location and class of the events highlighted in gray. The model labels seem to be shifted to the left by a few data points compared to the original classes, due to the fact that we assign a class to each snapshot of length 7 (described in Section 3.1). In other words, the very first snapshot is classified with the most probable label in the first seven data points. As the sliding window progresses with stride 1, each label is associated with the following seven data points resulting in anticipating the snapshot classification.

4.1 Model Application on Several Days of POES/MetOp Data: Preliminary Results

As we showed in **Section 3.1**, the dataset used for training has been significantly shrunk to only 50 data points for each precipitation event observed by POES/MetOp. In this section, we explore the model performance on longer time periods (full day of POES/MetOp data, the significant flux variability of which is shown in **Supplementary Figure S1** to test its generalization ability.

We apply the model to several POES/MetOp days and show the results in **Figure 4** and **Supplementary Figure S3**. Each panel in these figures is from a different date and none of the events shown belong to the dataset prepared in **Section 3.1** (they are all out-of-sample). Here, we are only considering events occurring in the outer radiation belt, thus we filter out any events occurring at $L < 2.5$ or $L > 8.5$ (L is expressed using the International Geomagnetic Reference Field, IGRF, model in POES/MetOp data). The panels on the left column of **Figure 4** show REP events (highlighted in brown), whereas the events on the right column are CSSs (highlighted in blue). This classification is accurate because the classified REPs indeed show isolated E4 precipitation, while the classified CSSs display an energy-dependent precipitation. During REP events (**Figures 2–4**), although the low-energy electrons (E1, E2 and E3 channels) appear to precipitate as well, their flux is likely the result of proton contamination, which is known to affect the electron measurements onboard POES/MetOp satellites (e.g., Evans and Greer, 2004; Yando et al., 2011; Capannolo et al. 2019, 2021). Note again that the location where these events are identified by the model differs from the exact event location by a few data points. This is not a major concern as this shift appears to be systematic and can be corrected in the post-processing by shifting the predicted model class by a few data points.

On the contrary, **Supplementary Figure S3** shows examples when the model does not perform very well and identifies two adjacent precipitation events belonging to different classes (panels a and b), mislabeled events (panel c) or false positive events (panel d). The cases in panel a) only last one data point and could be potentially disregarded since the model does not identify a long enough non-zero label. The event in panel d) shows a precipitating E4 flux that is higher than the others, which could indicate a potential issue in the recorded POES/MetOp data. Events in panels b) and c) instead must be appropriately ruled out or inspected further (e.g., what is the probability of each class? Is the probability of the CSS class comparable to that of the REP?). Handling false positives is beyond the scope of this work and we are aware that post-processing on the model output is needed before using these results for scientific research. The post-processing should rule out events lasting only one data point, adjacent events belonging to different non-zero classes, and events in the South Atlantic Anomaly, as well as improving the L shell calculation for each event (using Tsyganenko models such as the T89 (Tsyganenko, 1989) or T05 (Tsyganenko and Sitnov, 2005)) used to consider events occurring only in the outer radiation belt.

5 CONCLUSIONS AND DISCUSSION

In this work, we showed an example of an application of supervised deep learning to space sciences. Understanding when, where and why relativistic electrons precipitate into the Earth's atmosphere has a longstanding relevance for a variety of reasons (from improving our knowledge on plasma dynamics to study the space weather impacts of electron precipitation). In this work, we focused specifically on relativistic electron precipitation. Our goal was to classify the relativistic electron precipitation events depending on their spatial precipitation pattern, which in turn corresponds to their magnetospheric driver (waves or current sheet scattering). We used data from the POES/MetOp constellation of low-Earth-orbit satellites. Our task was supervised because we used the list of events studied by Capannolo et al. (2022), which were visually classified. Note that these events were classified only in a limited MLT sector (22–02); however, their MLT value was not used as input in the model, and in fact, our model is able to identify precipitation events at any MLT.

The dataset preparation was key to obtain a satisfying model performance. By considering only a short time window around each event instead of the full day of POES/MetOp data, using non-zero labels to indicate REPs (class of 1) or CSSs (class of 2) and labels at 0 to indicate the no-event, and including electron fluxes observed at different energies and look directions, we were able to obtain an appropriate dataset to use for training. We found that the LSTM architecture is suitable for identifying precipitation events and classifying them by precipitation pattern given its ability to consider the data history (in our case the precipitation pattern profile evolution along the satellite trajectory).

Our model is composed of one layer of 64 bidirectional LSTM cells, one layer of 256 fully connected neurons, and one layer of three dense cells. The inputs are the electron fluxes at different energies and look directions, and the output is the class of each data point. We obtained the model metrics (F1~0.948, AUC~0.995, and AUPRC~0.990) by conducting a k-fold cross-validation (k = 10). Our model is able to learn the

dataset properties correctly. The model is not only able to identify the electron precipitation events, but it also appropriately classifies them by their drivers.

Since the dataset used for training and testing purposes has been specifically designed to obtain a good model performance, it shows less variability than that typically observed by POES/MetOp over an entire orbit. Nevertheless, our model is still able to identify and classify the precipitation events when applied to a full day of data (**Figure 4**), though some false positives might still be identified (**Supplementary Figure S3**). Post-processing of these results is needed before being able to use the model outputs for scientific research; however, this is beyond the scope of this paper and left for future investigation. Once the post-processing routine is developed, this model could be easily used as a tool to produce lists of relativistic electron precipitation events in a very short amount of time, overcoming the complex task of developing deterministic algorithms based on flux thresholds to delineate the precipitation patterns and the time-expensive task of visually classifying these events by driver. In this way, we would be able to extend the study conducted in Capannolo et al. (2022) to the whole MLT range and statistically investigate on where the CSS effects should be considered for radiation belt and precipitation modeling, as well as compare them with the precipitation driven by waves. Such event dataset would also potentially open additional avenues of machine learning applications to space sciences; for example, from a space weather point of view, we could investigate if the electron precipitation events can be predicted by using solar images, solar wind data and/or geomagnetic indices.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://satdat.ngdc.noaa.gov/sem/poes/data/processed/ngdc/uncorrected/full/. The dataset preparation and model training are done on a Linux OS (version 3.10.0-1160.49.1.el7.x86_64) machine (Shared Computer Cluster at Boston University) in Python (version 3.8.6), using the TensorFlow library (version 2.5.0, https://www.tensorflow.org) and the Python packages: Matplotlib (https://matplotlib.org), Scikit Learn (https://scikit-learn.org/stable/), Xarray (https://xarray.pydata.org/en/stable/), Joblib (https://joblib.readthedocs.io/en/latest/), Seaborn (https://seaborn.pydata.org/), Numpy (https://numpy.org), and Pandas (https://pandas.pydata.org). **Figures 1**,**4** and **Supplementary Figures S1**,**S3** have been produced in IDL (version 8.6.0). The trained model and a sample Python script to apply the model on any POES/MetOp date can be found in the GitHub repository here: https://github.com/luisacap/REPs_classifier_codes_for_paper.git.

AUTHOR CONTRIBUTIONS

LC conducted the core of this work (dataset preparation, model development, model training, etc.). WL and SH contributed equally by offering feedback during the preparation of this work, and they share the last authorship.

FUNDING

This research is supported by the NSF grants AGS-1723588 and AGS-2019950, the NASA grants 80NSSC20K0698 and 80NSSC20K1270, and the Alfred P. Sloan Research Fellowship FG-2018-10936. SH would like to acknowledge the NASA FINESST Award 80NSSC21K1385.

ACKNOWLEDGMENTS

LC would like to acknowledge M. Capannolo for the insightful conversations on machine and deep learning. LC also acknowledges her teammates in the 2020 Heliophysics hackweek (B. Tremblay, A. K. Tiwari, A. Hu, B. Thompson, S. Shekhar, M. Shumko, S. Forsyth, J. Li, and D. Linko).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fspas.2022.858990/full#supplementary-material.

REFERENCES

Büchner, J., and Zelenyi, L. M. (1989). Regular and Chaotic Charged Particle Motion in Magnetotaillike Field Reversals: 1. Basic Theory of Trapped Motion. *J. Geophys. Res.* 94 (A9), 11821–11842. doi:10.1029/JA094iA09p11821

Capannolo, L., Li, W., Ma, Q., Shen, X. C., Zhang, X. J., Redmon, R. J., et al. (2019). Energetic Electron Precipitation: Multievent Analysis of its Spatial Extent during EMIC Wave Activity. *J. Geophys. Res. Space Phys.* 124, 2466–2483. doi:10.1029/2018ja026291

Capannolo, L., Li, W., Millan, R., Smith, D., Sivadas, N., Sample, J., et al. (2022). Relativistic Electron Precipitation Near Midnight: Drivers, Distribution, and Properties. *J. Geophys. Res. Space Phys.* 127, e2021JA030111. doi:10.1029/2021ja030111

Capannolo, L., Li, W., Spence, H., Johnson, A. T., Shumko, M., Sample, J., et al. (2021). Energetic Electron Precipitation Observed by FIREBIRD-II Potentially Driven by EMIC Waves: Location, Extent, and Energy Range from a Multievent Analysis. *Geophys. Res. Lett.* 48, e2020GL091564. doi:10.1029/2020gl091564

Dubyagin, S., Apatenkov, S., Gordeev, E., Ganushkina, N., and Zheng, Y. (2021). Conditions of Loss Cone Filling by Scattering on the Curved Field Lines for 30 keV Protons during Geomagnetic Storm as Inferred from Numerical Trajectory Tracing. *J. Geophys. Res. Space Phys.* 126, e2020JA028490. doi:10.1029/2020ja028490

Dubyagin, S., Ganushkina, N. Y., and Sergeev, V. (2018). Formation of 30 keV Proton Isotropic Boundaries during Geomagnetic Storms. *J. Geophys. Res. Space Phys.* 123, 3436–3459. doi:10.1002/2017ja024587

Duderstadt, K. A., Huang, C.-L., Spence, H. E., Smith, S., Blake, J. B., Crew, A. B., et al. (2021). Estimating the Impacts of Radiation belt Electrons on Atmospheric Chemistry Using FIREBIRD II and Van Allen Probes

Observations. *J. Geophys. Res. Atmospheres* 126, e2020JD033098. doi:10.1029/2020jd033098

Evans, D. S., and Greer, M. S. (2004). *Polar Orbiting Environmental Satellite Space Environment Monitor-2: Instrument Descriptions and Archive Data Documentation, NOAA Tech. Mem.*, 93. Boulder, Colo: Space Weather Predict. Cent.

Fytterer, T., Mlynczak, M. G., Nieder, H., Pérot, K., Sinnhuber, M., Stiller, G., et al. (2015). Energetic Particle Induced Intra-seasonal Variability of Ozone inside the Antarctic Polar Vortex Observed in Satellite Data. *Atmos. Chem. Phys.* 15, 3327–3338. doi:10.5194/acp-15-3327-2015

Ganushkina, N. Y., Pulkkinen, T. I., Kubyshkina, M. V., Sergeev, V. A., Lvova, E. A., Yahnina, T. A., et al. (2005). Proton Isotropy Boundaries as Measured on Mid- and Low-Altitude Satellites. *Ann. Geophys.* 23, 1839–1847. doi:10.5194/angeo-23-1839-2005

Gasque, L. C., Millan, R. M., and Shekhar, S. (2021). Statistically Determining the Spatial Extent of Relativistic Electron Precipitation Events Using 2-s Polar-Orbiting Satellite Data. *J. Geophys. Res. Space Phys.* 126, e2020JA028675. doi:10.1029/2020ja028675

Gilson, M. L., Raeder, J., Donovan, E., Ge, Y. S., and Kepko, L. (2012). Global Simulation of Proton Precipitation Due to Field Line Curvature during Substorms. *J. Geophys. Res.* 117, a–n. doi:10.1029/2012JA017562

Green, J. C. (2013). *MEPED Telescope Data Processing Algorithm Theoretical Basis Document, Natl. Oceanic and Atmos. Admin*. Boulder, Colorado: National Geophysical Data Center.

Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.* 9 (8), 1735–1780. doi:10.1162/neco.1997.9.8.1735

Horne, R. B., and Thorne, R. M. (1998). Potential Waves for Relativistic Electron Scattering and Stochastic Acceleration during Magnetic Storms. *Geophys. Res. Lett.* 25 (15), 3011–3014. doi:10.1029/98gl01002

Khazanov, G. V., Robinson, R. M., Zesta, E., Sibeck, D. G., Chu, M., and Grubbs, G. A. (2018). Impact of Precipitating Electrons and Magnetosphere-Ionosphere Coupling Processes on Ionospheric Conductance. *Space Weather* 16, 829–837. doi:10.1029/2018sw001837

Li, W., and Hudson, M. K. (2019). Earth's Van Allen Radiation Belts: From Discovery to the Van Allen Probes Era. *J. Geophys. Res. Space Phys.* 124, 8319–8351. doi:10.1029/2018JA025940

Liang, J., Donovan, E., Ni, B., Yue, C., Jiang, F., and Angelopoulos, V. (2014). On an Energy-Latitude Dispersion Pattern of Ion Precipitation Potentially Associated with Magnetospheric EMIC Waves. *J. Geophys. Res. Space Phys.* 119, 8137–8160. doi:10.1002/2014JA020226

Meraner, K., and Schmidt, H. (2018). Climate Impact of Idealized winter Polar Mesospheric and Stratospheric Ozone Losses as Caused by Energetic Particle Precipitation. *Atmos. Chem. Phys.* 18, 1079–1089. doi:10.5194/acp-18-1079-2018

Millan, R. M., and Thorne, R. M. (2007). Review of Radiation belt Relativistic Electron Losses. *J. Atmos. Solar-Terrestrial Phys.* 69, 362–377. doi:10.1016/j.jastp.2006.06.019

Mironova, I. A., Aplin, K. L., Arnold, F., Bazilevskaya, G. A., Harrison, R. G., Krivolutsky, A. A., et al. (2015). Energetic Particle Influence on the Earth's Atmosphere. *Space Sci. Rev.* 194 (1-4), 1–96. doi:10.1007/s11214-015-0185-4

Reeves, G. D., McAdams, K. L., Friedel, R. H. W., and O'Brien, T. P. (2003). Acceleration and Loss of Relativistic Electrons During Geomagnetic Storms. *Geophys. Res. Lett.* 30 (10), 1529. doi:10.1029/2002GL016513

Robinson, R. M., Vondrak, R. R., Miller, K., Dabbs, T., and Hardy, D. (1987). On Calculating Ionospheric Conductances from the Flux and Energy of Precipitating Electrons. *J. Geophys. Res.* 92 (A3), 2565–2569. doi:10.1029/JA092iA03p02565

Rodger, C. J., Clilverd, M. A., Green, J. C., and Lam, M. M. (2010). Use of POES SEM-2 Observations to Examine Radiation belt Dynamics and Energetic Electron Precipitation into the Atmosphere. *J. Geophys. Res.* 115, a–n. doi:10.1029/2008JA014023

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning Representations by Back-Propagating Errors. *Nature* 323, 533–536. doi:10.1038/323533a0

Schulz, M., and Lanzerotti, L. J. (1974). "Particle Diffusion in the Radiation Belts," in *Physics and Chemistry in Space* (Berlin: Springer), 7. doi:10.1007/978-3-642-65675-0

Sergeev, V. A., Malkov, M., and Mursula, K. (1993). Testing the Isotropic Boundary Algorithm Method to Evaluate the Magnetic Field Configuration in the Tail. *J. Geophys. Res.* 98 (A5), 7609–7620. doi:10.1029/92JA02587

Sergeev, V. A., Sazhina, E. M., Tsyganenko, N. A., Lundblad, J. Å., and Søraas, F. (1983). Pitch-angle Scattering of Energetic Protons in the Magnetotail Current Sheet as the Dominant Source of Their Isotropic Precipitation into the Nightside Ionosphere. *Planet. Space Sci.* 31 (Issue 10), 1147–1155. doi:10.1016/0032-0633(83)90103-4

Shekhar, S., Millan, R., and Smith, D. (2017). A Statistical Study of the Spatial Extent of Relativistic Electron Precipitation with Polar Orbiting Environmental Satellites. *J. Geophys. Res. Space Phys.* 122 (11), 284. doi:10.1002/2017JA024716

Sinnhuber, M., Tyssoy, H. N., Asikainen, T., Bender, S., Funke, B., Hendrickx, K., et al. (2021). Heppa III Intercomparison experiment on Electron Precipitation Impacts, Part II: Model-Measurement Intercomparison of Nitric Oxide (NO) during a Geomagnetic Storm in April 2010. *J. Geophys. Res. Space Phys.* 126, e2021JA029466.

Thorne, R. M. (2010). Radiation Belt Dynamics: The Importance of Wave-Particle Interactions. *Geophys. Res. Lett.* 37, L22107. doi:10.1029/2010GL044990

Tsyganenko, N. A. (1989). A Solution of the Chapman-Ferraro Problem for an Ellipsoidal Magnetopause. *Planet. Space Sci.* 37 (9), 1037–1046. doi:10.1016/0032063389900767

Tsyganenko, N. A., and Sitnov, M. I. (2005). Modeling the Dynamics of the Inner Magnetosphere during strong Geomagnetic Storms. *J. Geophys. Res.* 110, A03208. doi:10.1029/2004JA010798

Tyssøy, H. N., Sandanger, M. I., Degaard, L. K. G., Stadsnes, S., Aasnes, A., and Zawedde, A. E. (2016). Energetic Electron Precipitation into the Middle Atmosphere—Constructing the Loss Cone Fluxes from MEPED POES. *J. Geophys. Res. Space Phys.* 121, 5693–5707. doi:10.1002/2016JA022752

Yahnin, A. G., Yahnina, T. A., Raita, T., and Manninen, J. (2017). Ground Pulsation Magnetometer Observations Conjugated with Relativistic Electron Precipitation. *J. Geophys. Res. Space Phys.* 122, 9169–9182. doi:10.1002/2017JA024249

Yahnin, A. G., Yahnina, T. A., Semenova, N. V., Gvozdevsky, B. B., and Pashin, A. B. (2016). Relativistic Electron Precipitation as Seen by NOAA POES. *J. Geophys. Res. Space Phys.* 121, 8286–8299. doi:10.1002/2016JA022765

Yando, K., Millan, R. M., Green, J. C., and Evans, D. S. (2011). A Monte Carlo Simulation of the NOAA POES Medium Energy Proton and Electron Detector Instrument. *J. Geophys. Res.* 116, A10231. doi:10.1029/2011ja016671

Yu, Y., Jordanova, V. K., McGranaghan, R. M., and Solomon, S. C. (2018). Self-Consistent Modeling of Electron Precipitation and Responses in the Ionosphere: Application to Low-Altitude Energization during Substorms. *Geophys. Res. Lett.* 45, 6371–6381. doi:10.1029/2018gl078828

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Physics

ORIGINAL RESEARCH
published: 08 April 2022
doi: 10.3389/fphy.2022.775792

Overshoot Structure Near the Earth's Subsolar Magnetopause Generated by Magnetopause Motions

Xiaojian Song[1], Pingbing Zuo[2], Zhenning Shen[2], Xueshang Feng[2], Xiaojun Xu[3], Yi Wang[2], Chaowei Jiang[2] and Xi Luo[1]*

[1]Shandong Institute of Advanced Technology, Jinan, China, [2]Laboratory for Space Weather Storms, Institute of Space Science and Applied Technology, Harbin Institute of Technology, Shenzhen, China, [3]State Key Laboratory of Lunar and Planetary Sciences, Macau University of Science and Technology, Macao, China

For magnetopause crossing events, the observed magnetospheric magnetic fields in the vicinity of the subsolar magnetopause frequently present an overshoot structure; that is, in small vicinity of the magnetopause, the closer to the magnetopause, the stronger the magnetospheric magnetic field is. In this investigation, an automatic identification algorithm is developed to rapidly and effectively search the magnetopause crossing events using THEMIS data from 2007 to 2021. Nearly 59% of magnetopause crossing events identified near the subsolar region appear an overshoot structure. The statistical result shows that, for overshoot cases, the normalized change rate of magnetospheric magnetic field near the magnetopause is linearly related to the normalized magnetopause velocity, which means that the overshoot structure may be caused by the redistribution of the magnetospheric magnetic field due to the rapid magnetopause motion.

Keywords: magnetopause motion, magnetospheric magnetic field, overshoot, solar wind, spacecraft data analysis

OPEN ACCESS

Edited by:
Bala Poduval,
University of New Hampshire,
United States

Reviewed by:
Elizaveta Antonova,
Lomonosov Moscow State University,
Russia
Anton Artemyev,
Space Research Institute (RAS),
Russia

****Correspondence:***
Pingbing Zuo
pbzuo@hit.edu.cn

Specialty section:
This article was submitted to Space Physics, a section of the journal Frontiers in Physics

Received: *14 September 2021*
Accepted: *25 February 2022*
Published: *08 April 2022*

Citation:
Song X, Zuo P, Shen Z, Feng X, Xu X, Wang Y, Jiang C and Luo X (2022) Overshoot Structure Near the Earth's Subsolar Magnetopause Generated by Magnetopause Motions. Front. Phys. 10:775792. doi: 10.3389/fphy.2022.775792

1 INTRODUCTION

The magnetospheric magnetic field (MMF) originates from the Earth's main field, current systems inside the magnetosphere, for example, ring current, tail current, ionospheric current, field-aligned current [[1], and references therein], Chapman–Ferraro current on the boundary [2], and the interconnection due to partial penetration of the interplanetary magnetic field into the magnetosphere [3, 4]. The MMF is totally confined in the magnetosphere when ignoring the magnetic reconnection process around the magnetopause. The motion and deformation of the magnetopause will lead to the redistribution of the MMF [4, 5]. The position of the magnetopause is determined by the pressure balance on both sides. As the dynamic pressure of the solar wind varies dramatically, the position of the magnetopause is extremely unstable with the subsolar point distributing from 5 to 22 R_E [6, 7], where R_E is the radius of the Earth. According to the statistical results of the work of Paschmann et al. [8], the maximum normal velocity of the magnetopause is 367 km/s and the mean value is 51 km/s. This result is consistent with previous investigations [9–12]. The period of fluctuation of the magnetopause is mostly less than 200 s [13, 14], which arises or grows due to the boundary-inherent Kelvin–Helmholtz instability, or external sources, for example, solar wind pressure pulses or waves and disturbances in the foreshock region [15, 16]. On the other hand, the magnetopause is not always a smooth surface. Some local distortions, driven by flux transfer events, Kelvin–Helmholtz waves, and magnetosheath jets, may appear on it [17, 18].

For spacecraft located near the magnetopause, a number of magnetopause crossing events (MCEs) are expected to be detected as the location of the magnetopause is very volatile along with the change of the solar wind conditions, especially when the dynamic pressure pulse structures imping on the magnetosphere. In this study, we analyze the MCEs detected by THEMIS when THEMIS's apogee was located near the subsolar point. It is found that a large fraction of cases interestingly appear an overshoot structure as observed in the bow shock region [19, 20]; that is, from the magnetosphere to the magnetosheath, the magnetic field intensity increases quickly right before the magnetopause ramp, so the MMF adjacent to the magnetopause is stronger than that further away from the magnetopause. This structure is not rarely observed, but it still has not been paid much attention yet in the community of magnetopause research. To further understand this phenomenon, we carry out a statistical research on the relationship between the change rate of MMF near the magnetopause and the instantaneous speed of the magnetopause motion, based on an overshoot-type MCE database constructed from nearly 15 years' THEMIS observations at the subsolar region. It is found that the normalized change rate of MMF during the overshoot interval depends linearly on the normalized magnetopause motion speed in the statistical sense.

In **Section 2**, we give a brief introduction to the THEMIS MCE dataset constructed by an automatic MCE identification algorithm, and then, some typical MCEs are shown to present the interesting overshoot structure in **Section 3**. A statistical analysis about the dependence of the variation of MMF near the subsolar magnetopause and the magnetopause motion is given in **Section 4**. In the last section, a brief summary and discussion are given.

2 MCE HUNTING ALGORITHM

The five THEMIS probes were placed in highly elliptical equatorial orbits on 17 February 2007 [6, 21]. Right after the launch, all probes were lined up in the same orbit with a 15.4 R_E apogee. Around 2008, the orbits began to separate, with the apogee of THB, THC, THD&E, and THA being 30 R_E, 20 R_E, 12 R_E, and 10 R_E, respectively. Since 2011, THB&C became ARTEMIS and orbited the moon, the remaining three Earth-orbiting probes had an apogee of approximately 12 R_E. The apogee rotated slowly around Earth to cover the dayside, dawnside, nightside, and duskside of the magnetosphere. In this study, the ion data from the electrostatic analyzer [22] and magnetic field measurements provided by the fluxgate magnetometer [23], both with the time resolution of ~ 3 s, are used to identify MCEs.

Manual identification of MCEs can be a labor intensive task, since for the spacecraft located near the magnetopause, a number of MCEs are expected to be detected as the location of the magnetopause is dynamically controlled by the change of the solar wind conditions and inherent waves. Especially in a long-term survey, with hundreds or thousands of potential MCEs, manual identification becomes impractical [24]. On the other hand, manual identification is bound to be biased in some way. An observation classified as a MCE by one observer will not necessarily be classified as such by another observer [12]. To improve the identification efficiency, some automatic MCE identification routines were developed. MCEs have been automatically identified [7, 13, 24] in terms of the distinct difference between the disturbance level of the magnetic field in the magnetosphere and in the magnetosheath. However, some structures, such as current sheet in the magnetosheath, may also exhibit a large difference in the disturbance of the magnetic field with respect to the background magnetosheath. These structures may be mistaken for MCEs under this simple criterion. Suvorova [25] established two criteria for GOES and LANL to identify geosynchronous MCEs. For GOES (without particle data), their criterion is the correlation between the magnetic field observed by GOES and upstream monitor and the deviation of the observed magnetic field from the MMF. For LANL (without magnetic field data), their criterion is the difference of the ratio of density and temperature of high-energy ions in the magnetosheath and in the magnetosphere. These two criteria can only be used in geosynchronous MCE identification. Jelinek et al. [26] used the ratio of the parameters (magnetic field intensity and plasma density) observed by THEMIS and ACE at the same time to determine the most probable magnetopause locations in a statistical sense but failed to give the accurate magnetopause crossing time.

In this study, we develop a new algorithm to automatically identify MCEs and accurately determine the boundary layer between the magnetosheath and the magnetosphere using the *in situ* plasma and magnetic field data. The automatic identification of MCEs is designed in a four-step manner.

1) STEP 1: recognization of the region in which the probe is located (magnetosphere or magnetosheath).

In STEP 1, ion spectral energy flux density is used to distinguish the region in which the probe is located, but when the probe is located in the inner magnetosphere, the quality of the particle data measured by ESA is not good and missing data often occur. On the other hand, Park et al. [7] mentioned that the position of the subsolar magnetopause ranges from 5 to 22 R_E. Here, the probe is regarded to be located in the magnetosphere, if the radial distance of the probe from the Earth, R, is less than 5 R_E.

Figure 1 shows the ion spectral energy flux measured by THD in the magnetosphere (left) and in the magnetosheath (right). The two energy spectral curves are distinctly different: in the magnetosheath, the flux of middle energy is high and the fluxes of low and high energy are low; contrarily, the high energy flux in the magnetosphere is high. To describe the characteristics of the energy spectral curve, some parameters are defined. $e_{\max}$ is the logarithmic value of energy (unit is eV) corresponding to the maximum flux of ion spectral flux density (see **Figure 1**). e_{left} and e_{right} are the logarithmic value of energies on both sides of $e_{\max}$ corresponding to the flux one-tenth lower than the maximum flux.

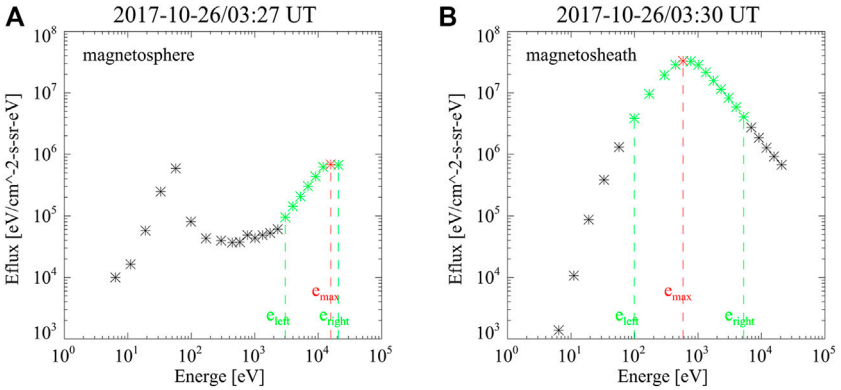

FIGURE 1 | (A,B) Ion spectral energy flux densities observed by THD in the magnetosphere and magnetosheath, respectively. The observation times are marked in the title of each subfigure. The red asterisks denote the maximum flux, and the green asterisks indicate the point at which the flux is one-tenth lower than the maximum flux.

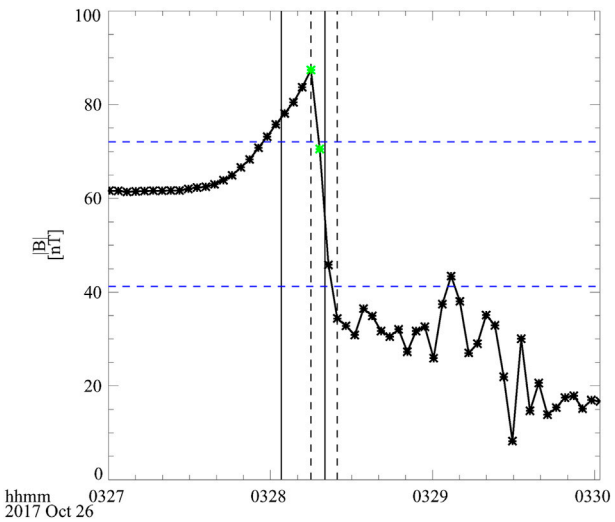

FIGURE 2 | Schematic picture to show how to get the accurate magnetopause crossing time in STEP 3. The solid (dashed) vertical lines denote the crossing time interval obtained in STEP 2 (STEP 3).

The probe is considered to be located in the magnetosphere, if the following conditions are satisfied:

- $R \leq 5\ R_E \bigcup e_{max} \notin [2,\ 3.5]$

If the following condition is satisfied, the probe is considered to be located in the magnetosheath:

- $R > 5\ R_E$
- $e_{max} \in (2, 3.5) \bigcap e_{right} - e_{max} > 0.5 \bigcap e_{max} - e_{left} > 0.5 \bigcap e_{right} - e_{left} > 1$

2) STEP 2: finding the candidate crossing time interval.

Based on the result of the region recognized in STEP 1, the candidate crossing time intervals, $[t_{sp,app},\ t_{sh,app}]$, are searched, where $t_{sp,app}$ denotes the start time (magnetospheric side) of crossing and $t_{sh,app}$ denotes the end time (magnetosheath side) of crossing. To avoid possible misjudgment, some more restrictions on the selection of MCEs are needed:

- Probe stays in the magnetosphere or magnetosheath region at least for 1 minute
- MCE completes in less than 1 minute

Note that these time constraints are mainly used to avoid possible misjudgment in STEP 1. It does not mean that the final result must meet the constraints in this step, as the next step slightly adjusts the start and end times of crossing to get the accurate one.

3) STEP 3: obtaining the accurate crossing time interval.

The third step is used to get the accurate crossing time based on the difference of the strength and disturbance level of B_z (the Z component of the magnetic field) in the magnetosphere and in the magnetosheath. As shown in **Figure 2**, the accurate start time of crossing is searched from $t_{sp,app} - 60$ to $t_{sh,app} + 60$ point by point, which satisfies the following conditions:

- $B_z\ [t_i] - B_z\ [t_{i+1}] > 3\sigma(B_{z,sp})$ (green asterisks in **Figure 2**)
- $\min(B_z[t_i:\ t_{i+5}]) < B_{z,max} - 0.25\,(B_{z,max} - B_{z,min})$ (upper blue dashed horizontal line)

The end time of crossing is the first time point that satisfies the following:

- $B_z\ [t_i] < B_{z,\,max} - 0.75(B_{z,\,max} - B_{z,\,min})$ (lower blue dashed horizontal line)

Here, $B_z\ [t_i]$ is the Z component of the magnetic field in the GSM coordinate system at time t_i; $\sigma(B_{z,sp})$ is the standard deviation of B_z observed within 1 minute just inside the magnetopause; $B_{z,\,\max}$ is the maximum value of B_z observed within 1 min from the magnetopause crossing time; and $B_{z,\,\min}$ is its minimum value.

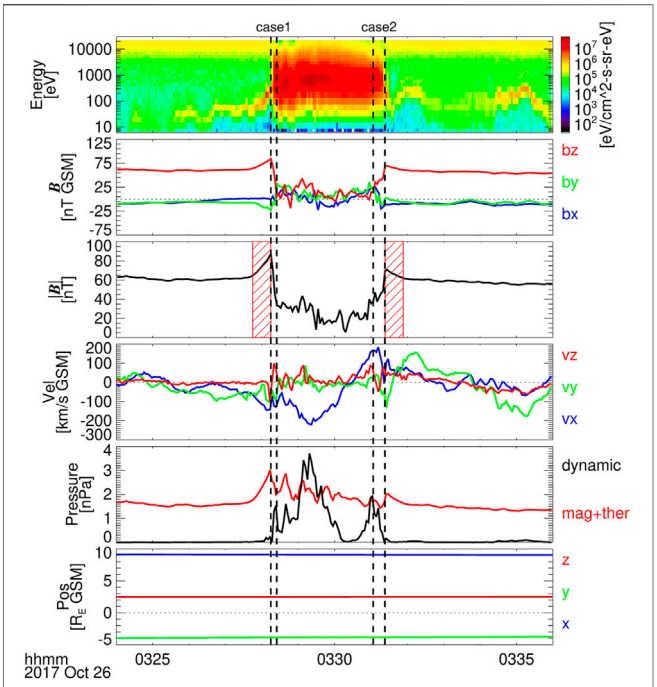

FIGURE 3 | Typical subsolar MCEs with an overshoot structure observed by THD between 03:24 UT and 03:36 UT on 26 October 2017. The parameters from top to bottom are ion spectral energy flux density, the three components of the magnetic field, the magnetic field strength, the bulk velocity of ion, various pressure (the red line means the total of magnetic and thermal pressure), and the position of THD. The time intervals marked by two vertical dashed lines are the magnetopause crossing time identified by our MCE automatic identification algorithm. The change rate of MMF is calculated by data measured during the time interval marked by oblique stripes.

4) STEP 4: confirming the crossing time interval.

The last step is to confirm the crossing time interval based on the criteria of Ivchenko et al. [13]:

- MCE should be completed within 30 s
- The standard deviation of magnetic field in the magnetospheric side is required to be less than 40% of that in the magnetosheath side
- The northward component of the MMF is required to exceed 10 nT
- The northward component of the MMF is required to be at least a factor of 1.3 greater than the corresponding magnetosheath component

According to the used criteria, our method is more suitable to identify MCEs when the magnetic field in the magnetosheath is southward. It can also obtain a good result when the magnetic field in the magnetosheath is northward, but it requires that B_z in the magnetosphere is 1.3 times bigger than that in the magnetosheath. Exactly speaking, this method may lose some cases, especially when the magnetic field strengths on the magnetospheric and magnetosheath sides are nearly equal to each other. These cases usually cannot give a clear magnetopause crossing time even by

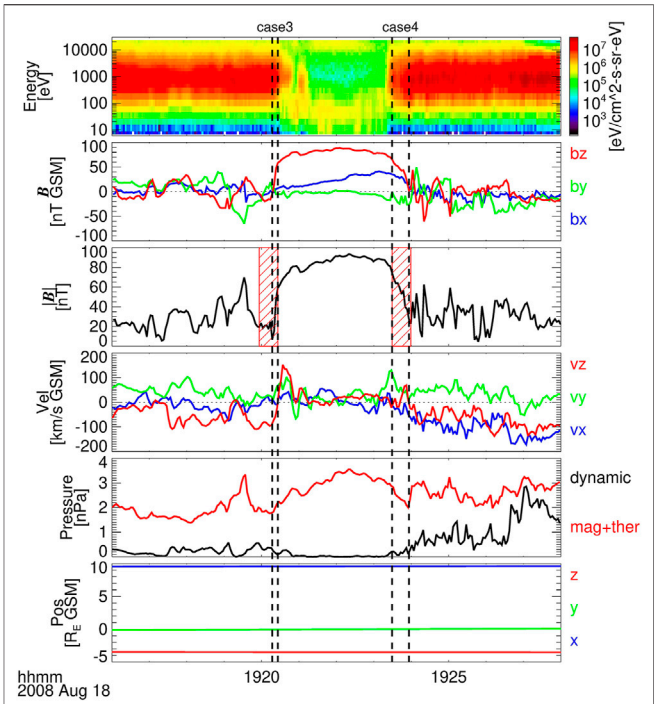

FIGURE 4 | Typical subsolar MCEs without an overshoot structure observed by THA between 19:16 UT and 19:28 UT on 18 August 2008. The format is the same as in **Figure 3**.

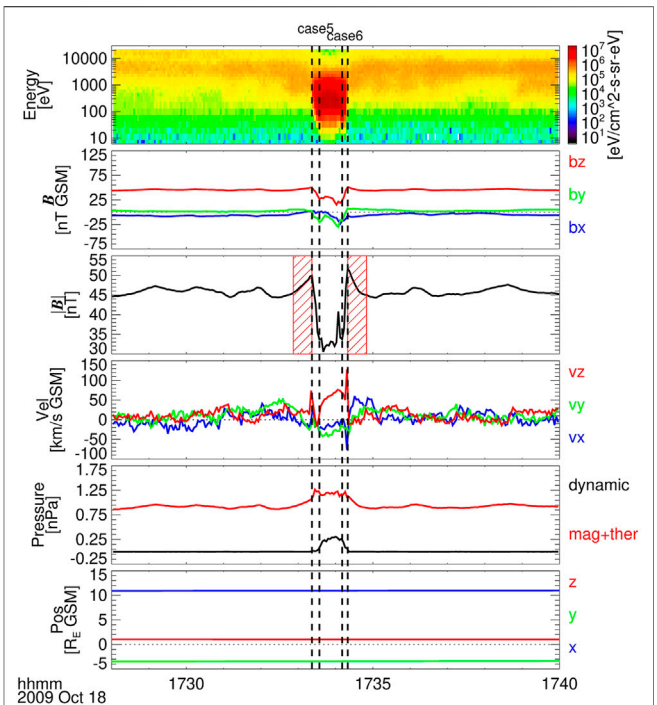

FIGURE 5 | Typical subsolar MCEs with overshoot structure observed by THE between 17:28 UT and 17:40 UT on 18 October 2009. The format is the same as in **Figure 3**.

manual inspection. So, it is rational to omit them. Our procedure has been applied to the observations of five THEMIS probes between 2007 and 2021 to search for MCEs. As we focus on the MCEs near the subsolar region, the MCEs observed only within 30^o between the Sun–Earth line are suitable for further research. As the position of the magnetopause and the magnetic field just inside the magnetopause can be affected by the dipole tilt angle [27, 28], the constraint that the MCEs should be within 20^o from the equator in the sum of the latitude of the magnetopause and the dipole tilt angle is added to the selection criteria. Eventually, 10,462 events of magnetopause crossing have been successfully identified by our method, which constructs an MCE database for further statistical study on the large-scale magnetopause structures and some important scientific problems related to small-scale structures of the magnetopause. In this study, we focus on the variation features of the MMF just inside the subsolar magnetopause.

3 OVERSHOOT STRUCTURE ADJACENT TO THE MAGNETOPAUSE

Figures 3–5 show six typical subsolar MCEs identified by our automatic identification algorithm. The parameters in each figure from top to bottom are ion spectral energy flux density, three components of the magnetic field, magnetic field intensity, bulk velocity of ions, and position of the probe. As the inspected time interval is short and the velocity of the probe is very small relative to the speed of the magnetopause, the probes are regarded as being located at fixed points during the inspected interval.

Figure 3 presents two consecutive MCEs detected by THD between 03:24 UT and 03:36 UT on 26 October 2017. At the beginning of this time interval, THD was located in the magnetosphere where the high-energy ions and strong magnetic fields were dominated. Around 03:28:14 UT, an abrupt decrease in the magnetic field strength and increase in the particle flux were observed, indicating that the magnetopause was moving inward and crossed THD. The crossing direction, *Dir*, is defined as 1 when the probe crosses the magnetopause from the magnetosphere to the magnetosheath and equal to -1 when the probe crosses in the opposite direction. The regions between two vertical dotted lines are procedure-given ramps of the magnetopause crossing, which denote the sharpest field change between the magnetosphere and the magnetosheath. It can be seen that, from the magnetosphere to the magnetosheath, the magnetic field in the vicinity of the ramp first increased gradually from a relatively stable state and then decreased sharply, which resembles a magnetic overshoot structure that is frequently observed at planetary bow shocks. The MMF strength observed by THD within 30 s adjacent to the magnetopause crossing time increased quickly and arrived at its peak just inside the magnetopause, B_0 = 83.71 nT. The variation of MMF can be fitted by a straight line. The slope of the fitted line, S_B, is 0.73 nT/s, and the mean absolute deviation from the observation, M_D, is 0.33 nT. At 03:28:14 UT, the magnetopause was located at (9.1, -3.8, 2.5) R_E in the GSM coordinate system, and the magnetopause standoff distance, R_0, is 10.19 R_E. Subsequently, the magnetopause moved outward and crossed THD again at 03:31:23 UT. After the second MCE, MMF decreased quickly from its peak value (B_0 is 69.15 nT). The change of MMF can also be fitted by a straight line with $S_B = -0.28\ nT/s$ and $M_D = 0.38\ nT$.

Figure 4 shows the observations of THA between 19:16 UT and 19:28 UT on 18 August 2008 when THA was located at (9.6, -0.01, -3.5) R_E. THA was located in the magnetosheath at the start time, and it crossed the magnetopause at 19:20:26 UT. Subsequently, the magnetopause moved inward and crossed THA at 19:23:56 UT. It can be seen from case 3 that, unlike case 2, the MMF increased gradually, and it can also be fitted by a straight line with $S_B = 0.82\ nT/s$ and $M_D = 0.87\ nT$. On the other hand, in case 4, unlike case 1, the MMF changed irregularly (M_D = 1.56 nT), although the overall trend was decreasing.

Figure 5 presents two consecutive MCEs detected by THE during the interval between 17:28 UT and 17:40 UT on 18 October 2009. During this time interval, THE was located at (10.9, -3.4, 1.0) R_E. At the beginning of this time interval, THE was located in the magnetosphere, and it crossed the magnetopause around 17:33:22 UT. The magnetic field just inside the magnetopause increased linearly with $S_B = 0.15\ nT/s$ and $M_D = 0.09\ nT$. The magnetopause moved outward and crossed THE again at 17:34:10 UT. After the second MCE, the MMF adjacent to the magnetopause also showed an overshoot structure with $S_B = -0.27\ nT/s$ and $M_D = 0.44\ nT$. For the two events, although the MMF had some oscillations in 17:28–17:32 and 17:35–17:40, which were possibly triggered by magnetopause motion or other small structures appearing on the magnetopause, the overshoot can be easily identified.

Thousands of MCEs have been detected by the five THEMIS probes. After visual inspection of the variations of MMFs just inside the magnetopause, it is found that, like case 1, case 2, case 5, and case 6, an overshoot structure, that is, from the magnetosphere to the magnetosheath, the magnetic field adjacent to the magnetopause plane increases quickly in a short interval from a relatively stable MMF state and then decreases sharply at the crossing ramp, is very common. Here, the criteria to judge an overshoot structure are as follows: M_D is smaller than 1 nT and $S_B{*}Dir > 0$. Totally, we got 6,170 (~ 59%) cases with overshoot in all 10,462 MCEs for further analysis.

4 RELATIONSHIP BETWEEN OVERSHOOT AND THE MAGNETOPAUSE MOTION

A statistical research on the relationship between the magnetic field intensity at the point fixed on the magnetopause, B_0, and the subsolar magnetopause standoff distance, R_0, was carried out by Shue et al. [29]. In their work, a simple equation was obtained to fit their dependence based on 614 subsolar MCEs with plateau magnetic fields in the magnetospheric side:

$$B_0 \propto R_0^D, \tag{1}$$

where the power law exponent, D, is used in contrast to the expected -3 for the pure dipole magnetic field. Is this equation still valid under an overshoot structure? Overshoot structure means that the magnetosphere is not in a steady state; this situation is most likely caused by the motion of magnetopause. So, we will

TABLE 1 | Criteria for selecting cases with reliable normal velocity.

No.	Criterion
1	$\lambda_2/\lambda_3 > 5$
2	$\Phi < 20^o$
3	$HT_{cc} > 0.8$

conduct a statistical research on the relationship between overshoot and the magnetopause motion in the following.

The normal speed of the magnetopause, V_{mp}, can be obtained by the de Hoffmann–Teller velocity [30], V_{HT}, and the magnetopause normal direction is obtained by constrained minimum variance analysis [31], n_{mvabc}. The ratio of the middle and the smallest eigenvalue obtained in the constrained minimum variance analysis procedure, λ_2/λ_3, marks the quality of the normal, and the larger the λ_2/λ_3, the more reliable the normal, and the threshold is often taken as 2 [32]. The angle between the magnetopause normal calculated by constrained minimum variance analysis and by Shue et al. [33], Φ, is also recorded to indicate the degree of magnetopause deformation from the normally smoothed magnetopause. The larger the Φ, the greater the deformation. On the other hand, the correlation coefficient of two electric fields calculated by $E_1 = -v \times B$ and $E_{HT} = -V_{HT} \times B$ (v, B are the ion bulk velocity and magnetic field observed by the probe adjacent to the magnetopause, respectively), HT_{cc}, denotes the quality of the de Hoffmann–Teller frame, and it ranges from 0 to 1. The larger the HT_{cc}, the more reliable the de Hoffmann–Teller frame. The reliability of the magnetopause normal velocity depends on the reliability of the normal direction, n_{mvabc}, and de Hoffmann–Teller velocity, V_{HT}. Three criteria with limits on λ_2/λ_3, HT_{cc}, and Φ are used to select cases with reliable normal velocity, which are shown in **Table 1**. $\lambda_2/\lambda_3 > 5$ guarantees the reliability of the calculated normal direction, $HT_{cc} > 0.8$ ensures that the calculated de Hoffmann–Teller frame is reliable, and $\Phi < 20^o$ denotes that the magnetopause is not greatly deformed. Among the identified MCEs with overshoot, 1,641 cases meet these requirements, and the corresponding normal velocities are calculated.

The interplanetary magnetic field direction may have a great influence on the state of the magnetosphere, but the uncertainty of the traveling time of solar wind from bow shock to the magnetopause is large, and the magnetic field direction may change when they travel to the magnetosheath. Figures 3 and 4 clearly show the existence of the turbulent fluctuations of the magnetic field and velocity in the magnetosheath. To date, a number of distinct case studies and a few statistical explorations at different parts of the magnetosheath show that the turbulence feature is highly related to the background and upstream conditions [34]. Magnetosheath turbulence will disconnect the Bz components of the magnetic field in the solar wind and near magnetopause. Pulinets et al. [35] show that the sign of the Bz near the magnetopause subsolar point does not coincide with the sign of interplanetary magnetic field Bz in ~ 30% cases, but it is the magnetosheath magnetic field that directly influences the state of the magnetosphere. So, the averaged Bz within 30 s just outside the magnetopause is used to study the direction effect. We select and divide these events into two groups: 923 cases with northward magnetosheath magnetic field ($Bz_{sh} > 2\ nT$) and 587 cases of southward magnetosheath magnetic field ($Bz_{sh} < -2\ nT$). The parameter $Bz_{sh} = \pm 2\ nT$ is chosen based on two principles: 1) enough samples (> 500) to provide meaningful statistical results, and 2) the data set in different groups should be distinguished significantly. The distribution of MCEs in the (x, $\sqrt{y^2 + z^2}$) plane are plotted in **Figure 6**. There is no obvious regional aggregation and no obvious difference under southward and northward magnetosheath magnetic field.

Figure 7 shows the statistical results based on these events. In **Figure 7A**, $\log_{10}(B_0)$ is plotted against $\log_{10}$ (R_0) for the northward magnetosheath magnetic field. **Figure 7B** shows the relation between S_B/B_0 and V_{mp}/R_0 for the northward magnetosheath magnetic field. **Figure 7C** and **Figure 7D** are drawn in the same format as **Figure 7A** and **Figure 7B**, respectively, except for the southward magnetosheath magnetic field. These data can be fitted by straight lines, and the fitting parameters are integrated into **Table 2**. It can be seen that $\log_{10}(B_0)$ and $\log_{10}$ (R_0) have a good linear relationship in **Figures 7A,C**, and their correlation coefficients, *cc*, are -0.89 and -0.87, respectively. An F test is performed to evaluate the confidence level of a fit [36]. The critical F value tabulated with 95% confidence and 921 (585) degrees of freedom is 3.86 (3.86). The calculated F values from the data are 3,483 and 1,797, which are much larger than the critical F value. This demonstrates the rationality of **Eq. 1**. The normalized change rate of MMF, S_B/B_0, and the normalized speed of the magnetopause, V_{mp}/R_0, shown in **Figures 7C,D**, all have a clear linear relationship with *cc* equal to -0.68 and -0.71, respectively, and the calculated F value (806 and 588) is also much larger than the critical F value.

5 SUMMARY AND DISCUSSION

In this study, the variation of the MMF just inside the subsolar magnetopause is studied, and we find that more than half of the MCEs show an overshoot structure. It is also found that the normalized change rate of the magnetic field intensity just inside the subsolar magnetopause is linearly related to the normalized velocity of the magnetopause in cases showing an overshoot structure.

It is reasonable to consider that the overshoot may be a certain kind of the magnetopause current layer itself. Generally, the magnetopause is made up of the magnetopause current (it may be composed of several current layers). Some other structures will be distributed on both sides, such as the depletion layer, magnetosheath boundary layer, and low latitude boundary layer. However, according to previous studies, no evidence indicated that these current sheets and structures can result in the formation of the overshoot magnetic structure near the magnetopause. In addition, some kinds of waves and local indentations may appear on the magnetopause, which have been reported in few case studies. These structures may be common (although few reported), but so far, there is no statistical study on this issue. Although they are possibly responsible for the formation of the overshoot in a statistical sense, it is difficult to explain the linear relationship between the variation of magnetic field and the magnetopause motion for this kind of

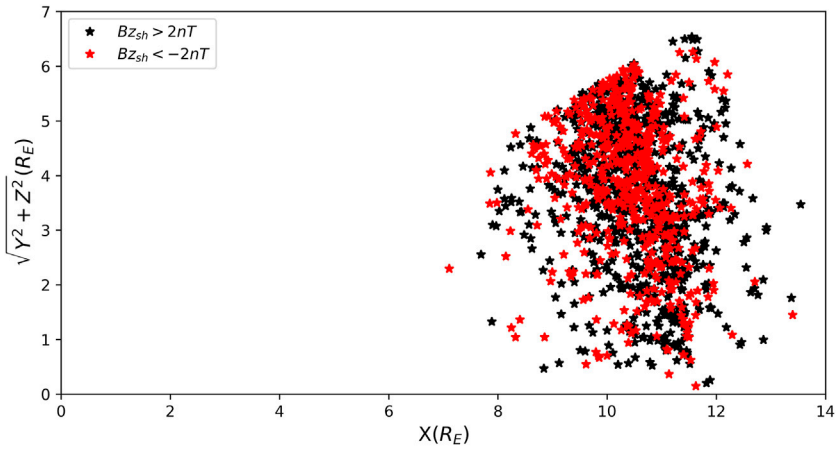

FIGURE 6 | Distribution of the location of magnetopause crossing events in the (x, $\sqrt{y^2 + z^2}$) plane. Black (red) color means the case is selected under the northward (southward) magnetosheath magnetic field.

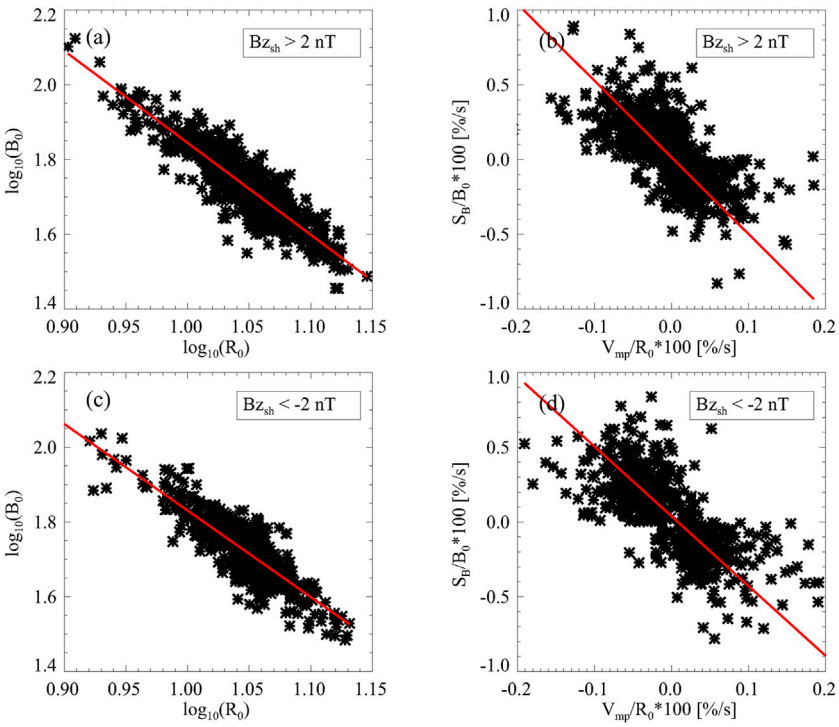

FIGURE 7 | Statistical results based on MCEs with an overshoot structure. **(A)** Here, $\log_{10}(B_0)$ is plotted against $\log_{10}(R_0)$ for $Bz_{sh} > 2nT$. **(B)** The relation of S_B/B_0 and V_{mp}/R_0 for $Bz_{sh} > 2nT$. **(C)** and **(D)** Plotted in the same format as in (A) and (B), respectively, except for $Bz_{sh} < -2nT$. These data are fitted by straight lines shown by red color.

overshoot. To study the wave propagation and the indentation at the magnetopause surface, multiple spacecraft data analyses are needed, and it is more suitable for a case study. For a single spacecraft, it is impossible to distinguish whether the magnetic field variation comes from spatial or temporal effect.

The fitting result between $\log_{10}(B_0)$ and $\log_{10}(R_0)$ shows that the magnetic field strength just inside the magnetopause with the northward magnetosheath magnetic field is usually larger than that with the southward magnetosheath magnetic field. This result is consistent with the results of the work of Shue et al. [29] and Wang et al. [37]. The fitting result between S_B/B_0 and V_{mp}/R_0 shows that the magnetic field strength just inside the magnetopause with the northward magnetosheath magnetic field may be slightly more compressed than that with the southward magnetosheath

TABLE 2 | Some parameters in **Figure 7**.

Bz_{sh}	Number	$\log_{10}(B_0)$ vs. $\log_{10}(R_0)$				S_B/B_0 vs. V_{mp}/R_0			
		Intercept	Slope	Cc	F (critical F)	Intercept	Slope	cc	F (critical F)
Northward	923	4.31	−2.46	−0.89	3,483 (3.86)	0.01	−5.13	−0.68	806 (3.86)
Southward	587	4.15	−2.32	−0.87	1797 (3.86)	0.04	−4.67	−0.71	588 (3.86)

Critical F is the value tabulated with 95% confidence and corresponding degrees of freedom.

magnetic field with the same inward magnetopause velocity, as the slope in **Figure 7B** is a little smaller than that in **Figure 7D**. This effect may be caused by the magnetic erosion under the southward magnetosheath magnetic field [38].

Considering the aforementioned information, we think the temporal change due to magnetosphere compression or decompression is very likely to be responsible for the gradual magnetic increase (overshoot under magnetopause inward motion) or decrease (overshoot under magnetopause outward motion). Here, we give a brief explanation to the overshoot structure. When the magnetopause moves inward or outward rapidly, the MMF will change dramatically resulting from the quick change of position and intensity of the magnetopause current system. One probe at a fixed position in the magnetosphere near the subsolar magnetopause will experience a very rapid increasing or decreasing magnetic field due to the reconfiguration of MMF, in response to the sudden compression or decompression of the magnetosphere. Therefore, the overshoot structure is expected to be formed. Likewise, when the magnetopause is stable or the motion of magnetopause is relatively slow, the variations of MMF at a fixed point can be negligible. In addition, sometimes when the magnetopause moves slowly, the influence of other processes (e.g., plasma wave and other current systems) may result in the irregular variation of MMF near the magnetopause.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

XS developed the MCE automatic hunting approach and wrote the draft. XS and PZ performed the data analysis and result analysis. All authors participated in discussions and revisions on the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant No. 41731067), the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2019A1515011067), and the Shenzhen Natural Science Fund (the Stable Support Plan Program GXWD20201230155427003-20200822192703001).

ACKNOWLEDGMENTS

The authors acknowledge NASA contract NAS5-02099 and V. Angelopoulos for the use of data from the THEMIS Mission. Specifically, they thank C. W. Carlson and J. P. McFadden for the use of ESA data and K. H. Glassmeier, U. Auster, and W. Baumjohann for the use of FGM data provided under the lead of the Technical University of Braunschweig and with financial support through the German Ministry for Economy and Technology and the German Center for Aviation and Space (DLR) under contract 50OC0302.

REFERENCES

1. Ganushkina NY, Liemohn MW, Dubyagin S, Daglis IA, Dandouras I, De Zeeuw DL, et al. Defining and Resolving Current Systems in Geospace. *Ann Geophys* (2015) 33:1369–402. doi:10.5194/angeo-33-1369-2015
2. Chapman S, Ferraro VCA. A New Theory of Magnetic Storms. *J Geophys Res* (1930) 38:79–96. doi:10.1038/126129a0
3. Pudovkin MI, Semenov VS. Peculiarities of the MHD-Flow by the Magnetopause and Generation of the Electric Field in the Magnetosphere. *Ann de Geophysique* (1977) 33:423–7.
4. Tsyganenko NA. Effects of the Solar Wind Conditions in the Global Magnetospheric Configurations as Deduced from Data-Based Field Models (Invited). In: EJ Rolfe B Kaldeich, editors. International Conference on Substorms, Versailles, France, 12-17, May, 1996. ESA Special Publication (1996). p. 181.
5. Song X, Zuo P, Feng X, Shue J-H, Wang Y, Jiang C, et al. Abnormal Magnetospheric Magnetic Gradient Direction Reverse Around the Indented Magnetopause. *Astrophys Space Sci* (2019) 364:146. doi:10.1007/s10509-019-3635-8
6. Rufenach CL, Martin RF, Jr, Sauer HH. A Study of Geosynchronous Magnetopause Crossings. *J Geophys Res* (1989) 94:15125–34. doi:10.1029/JA094iA11p15125
7. Park E, Moon Y-J, Lee K. Observational Test of Empirical Magnetopause Location Models Using Geosynchronous Satellite Data. *J Geophys Res Space Phys* (2016) 121:10,994–11,006. doi:10.1002/2015ja022271
8. Paschmann G, Haaland SE, Phan TD, Sonnerup BUÖ, Burch JL, Torbert RB, et al. Large-scale Survey of the Structure of the Dayside Magnetopause by Mms. *J Geophys Res Space Phys* (2018) 123:2018–33. doi:10.1002/2017ja025121
9. Berchem J, Russell CT. The Thickness of the Magnetopause Current Layer: ISEE 1 and 2 Observations. *J Geophys Res* (1982) 87:2108–14. doi:10.1029/JA087iA04p02108
10. Le G, Russell CT. The Thickness and Structure of High Beta Magnetopause Current Layer. *Geophys Res Lett* (1994) 21:2451–4. doi:10.1029/94gl02292
11. Phan TD, Paschmann G. Low-latitude Dayside Magnetopause and Boundary Layer for High Magnetic Shear: 1. Structure and Motion. *J Geophys Res* (1996) 101:7801–15. doi:10.1029/95ja03752

12. Haaland S, Reistad J, Tenfjord P, Gjerloev J, Maes L, DeKeyser J, et al. Characteristics of the Flank Magnetopause: Cluster Observations. *J Geophys Res Space Phys* (2014) 119:9019–37. doi:10.1002/2014ja020539
13. Ivchenko NV, Sibeck DG, Takahashi K, Kokubun S. A Statistical Study of the Magnetosphere Boundary Crossings by the Geotail Satellite. *Geophys Res Lett* (2000) 27:2881–4. doi:10.1029/2000gl000020
14. Plaschke F, Glassmeier K-H, Auster HU, Angelopoulos V, Constantinescu OD, Fornaçon K-H, et al. Statistical Study of the Magnetopause Motion: First Results from Themis. *J Geophys Res* (2009) 114:A00C10. doi:10.1029/2008ja013423
15. Yumoto K. Low-frequency Upstream Wave as a Probable Source of Low-Latitude Pc 3-4 Magnetic Pulsations. *Planet Space Sci* (1985) 33:239–49. doi:10.1016/0032-0633(85)90133-3
16. Plaschke F, Angelopoulos V, Glassmeier K-H. Magnetopause Surface Waves: Themis Observations Compared to Mhd Theory. *J Geophys Res Space Phys* (2013) 118:1483–99. doi:10.1002/jgra.50147
17. Dmitriev AV, Suvorova AV. Traveling Magnetopause Distortion Related to a Large-Scale Magnetosheath Plasma Jet: Themis and Ground-Based Observations. *J Geophys Research-Space Phys* (2012) 117:A08217. doi:10.1029/2011ja016861
18. Plaschke F, Hietala H, Archer H, Blanco-Cano M, Kajdič X, Karlsson P, et al. Jets Downstream of Collisionless Shocks. *Space Sci Rev* (2018) 214:81. doi:10.1007/s11214-018-0516-3
19. Heppner JP, Sugiura M, Skillman TL, Ledley BG, Campbell M. Ogo-a Magnetic Field Observations. *J Geophys Res (1967-1977)* (1967) 72:5417–71. doi:10.1029/JZ072i021p05417
20. Russell CT, Greenstadt EW. Initial ISEE Magnetometer Results: Shock Observation. *Space Sci Rev* (1979) 23:3–37. doi:10.1007/BF00174109
21. Angelopoulos V. The Themis mission. *Space Sci Rev* (2008) 141:5–34. doi:10.1007/s11214-008-9336-1
22. Auster HU, Glassmeier KH, Magnes W, Aydogar O, Baumjohann W, Constantinescu D, et al. The Themis Fluxgate Magnetometer. *Space Sci Rev* (2008) 141:235–64. doi:10.1007/s11214-008-9365-9
23. McFadden JP, Carlson CW, Larson D, Ludlam M, Abiad R, Elliott B, et al. The Themis Esa Plasma Instrument and In-Flight Calibration. *Space Sci Rev* (2008) 141:277–302. doi:10.1007/s11214-008-9440-2
24. Case NA, Wild JA. The Location of the Earth's Magnetopause: A Comparison of Modeled Position and *In Situ* Cluster Data. *J Geophys Res Space Phys* (2013) 118:6127–35. doi:10.1002/jgra.50572
25. Suvorova A. Necessary Conditions for Geosynchronous Magnetopause Crossings. *J Geophys Res* (2005) 110:A01206. doi:10.1029/2003ja010079
26. Jelínek K, Němeček Z, Šafránková J. A New Approach to Magnetopause and bow Shock Modeling Based on Automated Region Identification. *J Geophys Res* (2012) 117:A05208. doi:10.1029/2011ja017252
27. Spreiter JR, Briggs BR. Theoretical Determination of the Form of the Boundary of the Solar Corpuscular Stream Produced by Interaction with the Magnetic Dipole Field of the Earth. *J Geophys Res (1962-1977)* (1962) 67:37–51. doi:10.1029/JZ067i001p00037
28. Petrinec SM, Russell CT. An Examination of the Effect of Dipole Tilt Angle and Cusp Regions on the Shape of the Dayside Magnetopause. *J Geophys Res* (1995) 100:9559–66. doi:10.1029/94JA03315
29. Shue J-H, Chen Y-S, Hsieh W-C, Nowada M, Lee BS, Song P, et al. Uneven Compression Levels of Earth's Magnetic fields by Shocked Solar Wind. *J Geophys Res* (2011) 116:A02203. doi:10.1029/2010ja016149
30. De Hoffmann F, Teller E. Magneto-hydrodynamic Shocks. *Phys Rev* (1950) 80:692–703. doi:10.1103/physrev.80.692
31. Sonnerup BUO, Scheible M. Minimum and Maximum Variance Analysis. *ISSI Scientific Rep Ser* (1998) 1:185–220.
32. Neugebauer M, Alexander CJ. Shuffling Foot Points and Magnetohydrodynamic Discontinuities in the Solar Wind. *J Geophys Res* (1991) 96:9409. doi:10.1029/91JA00566
33. Shue J-H, Song P, Russell CT, Steinberg JT, Chao JK, Zastenker G, et al. Magnetopause Location under Extreme Solar Wind Conditions. *J Geophys Res* (1998) 103:17691–700. doi:10.1029/98ja01103
34. Rakhmanova L, Riazantseva M, Zastenker G. Plasma and Magnetic Field Turbulence in the Earth's Magnetosheath at Ion Scales. *Front Astron Space Sci* (2021) 7:616635. doi:10.3389/fspas.2020.616635
35. Pulinets MS, Antonova EE, Riazantseva MO, Znatkova SS, Kirpichev IP. Comparison of the Magnetic Field before the Subsolar Magnetopause with the Magnetic Field in the Solar Wind before the bow Shock. *Adv Space Res* (2014) 54:604–16. doi:10.1016/j.asr.2014.04.023
36. Bevington PR, Robinson DK. *Data Reduction and Error Analysis for the Physical Sciences*. New York: McGraw-Hill (2003).
37. Wang M, Lu J, Liu Z, Pei S. Dependence of Magnetic Field Just inside the Magnetopause on Subsolar Standoff Distance: Global Mhd Results. *Chin Sci Bull* (2012) 57:1438–42. doi:10.1007/s11434-011-4961-6
38. Aubry MP, Russell CT, Kivelson MG. Inward Motion of the Magnetopause before a Substorm. *J Geophys Res (1970-1977)* (1970) 75:7018–31. doi:10.1029/JA075i034p07018

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Astronomy and Space Sciences

PERSPECTIVE
published: 29 April 2022
doi: 10.3389/fspas.2022.852222

Understanding Large-Scale Structure in Global Ionospheric Maps With Visual and Statistical Analyses

Olga Verkhoglyadova, Xing Meng and Jacob Kosberg*

Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, United States

We applied two different techniques to identify high-density structures in global maps of height-integrated electron density of the Earth's ionosphere. We discuss benefits and limitations of these approaches to structure identification. We suggest that they are complementary and can aid our understanding of the properties of the global ionosphere. We stress out importance of a consistent definition of large-scale ionospheric structures.

Keywords: ionosphere, machine learning, statistics, electron density, geomagnetic activity

OPEN ACCESS

Edited by:
Philip J. Erickson, Massachusetts Institute of Technology, United States

Reviewed by:
Alexei V. Dmitriev, Lomonosov Moscow State University, Russia
Ercha Aa, Massachusetts Institute of Technology, United States

****Correspondence:***
Olga Verkhoglyadova
Olga.Verkhoglyadova@jpl.nasa.gov

Specialty section:
This article was submitted to Space Physics, a section of the journal Frontiers in Astronomy and Space Sciences

Received: *10 January 2022*
Accepted: *18 March 2022*
Published: *29 April 2022*

Citation:
Verkhoglyadova O, Meng X and Kosberg J (2022) Understanding Large-Scale Structure in Global Ionospheric Maps With Visual and Statistical Analyses. Front. Astron. Space Sci. 9:852222. doi: 10.3389/fspas.2022.852222

INTRODUCTION

Global ionospheric state of the Earth's upper atmosphere is frequently characterized by the total electron content (TEC) that is vertically integrated electron density. TEC distribution over the globe features prominent daytime equatorial ionization anomalies (EIAs), see for instance (Schunk and Nagy, 2009). Observations and follow-up modeling provide evidence of multiple regions with elevated TEC (Maruayama et al., 2016; Astafyeva et al., 2016; Astafyeva et al., 2017) that also include EIAs, i.e., high density regions (HDRs). Physical mechanisms responsible for HDR formations are not well understood. However, knowledge of HDRs, their occurrence, morphology, and evolution are important for space weather forecasting. We suggest that a robust methodology needs to be developed to identify TEC structures, i.e., HDRs, and create an extensive database of the structure occurrences. Such a database should contain information on locations of HDRs, TEC magnitude, time and local time of occurrences, and ancillary information on geomagnetic and solar activity. The data will be crucial for identifying physical mechanisms, testing physical hypotheses and validating modeling results.

To illustrate two different approaches to HDR identification, we use a timeseries of global ionospheric maps (GIMs), a gridded 2D data product for TEC that is commonly used to visualize global ionospheric state. We used two techniques, a mixture method approach and a computer vision approach, that can be utilized to address the following questions. How many anomalies and how many HDRs are present in a GIM? How does the number of the HDRs and their intensities depend on solar and geomagnetic activity? We used the GIM dataset (binned 1° by 1° for every 15 min, https://sideshow.jpl.nasa.gov/pub/iono_daily/gim_for_research/jpli/) produced by Jet Propulsion Laboratory, California Institute of Technology for over 20 years to demonstrate these two approaches. We would like to note that there are several GIM data products available (see the reviews by Hernandez-Pajares et al. (2017); Roma-Dollase et al. (2018)). We chose JPL GIM in this study as a representative GIM dataset with 15-min temporal resolution. **Figure 1** shows an example of JPL GIM with two EIAs over the South America and several other HDRs. Below we will briefly discuss our approaches. We focus on large-scale (thousands of km) structuring of the ionosphere.

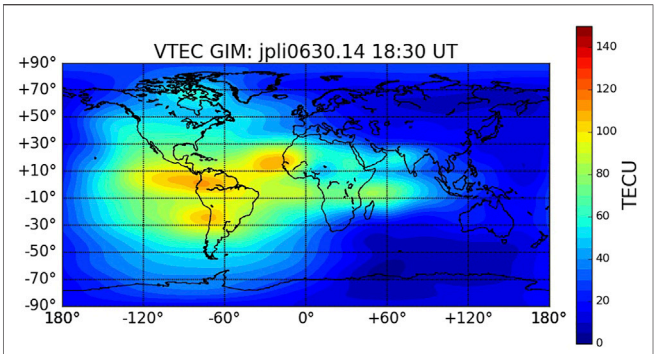

FIGURE 1 | Example of GIM on 4 March, 2014. TEC is shown in TEC units (TECU), where 1 TECU = 10^{16} electrons/m^2. Horizontal and vertical axes correspond to geographic longitude and latitude, correspondingly.

MIXTURE METHOD

An unsupervised Gaussian Mixture (GM) method, as implemented by scikit-learn (Pedregosa et al., 2011) is utilized to identify unique TEC sub-populations or HDRs (https://scikit-learn.org/stable/modules/mixture.html#mixture). This approach is informative and can be used to understand hierarchical layering of Gaussian clusters. We assume TEC data points arise from a mixture of a finite number of Gaussian distributions whose parameters are unknown. Due to the topology of GIMs, we had to extend scikit's GM implementation to account for periodic boundaries, i.e., a high-density region that "wraps around" the zero meridian will not be counted as two regions. For each possible number of expected clusters (one to ten, but no higher, to reduce computational requirements and match the order of magnitude of the Computer Vision Method), we compute GM parameters and identify the associated Bayesian Information Criteria (BIC) of each fit. The optimal cluster count was then selected by choosing the knee of the emerging plot of cluster count vs. BIC. This method is sometimes called the Elbow method in statistical clustering. **Figures 2A,B** show examples of identifications of 2 and 5 clusters, correspondingly. Here, horizontal axes correspond to geographic latitude and longitude. The vertical axis shows the cluster number. This approach is based on statistical properties of the TEC distribution and is sensitive to visually small changes in background density. Note that the GM method designates the background as one of the clusters. Thus, upon visual inspection there is one HDR on **Figure 2A**. However, it is difficult to determine visually which cluster or clusters correspond to the background density in **Figure 2B**. For our purpose of understanding large-scale structure, we appreciate that this method accounts for information contained in the data which has physical significance, whether or not that information is visually discernible. For this reason, we believe the following method based on image processing to be complementary.

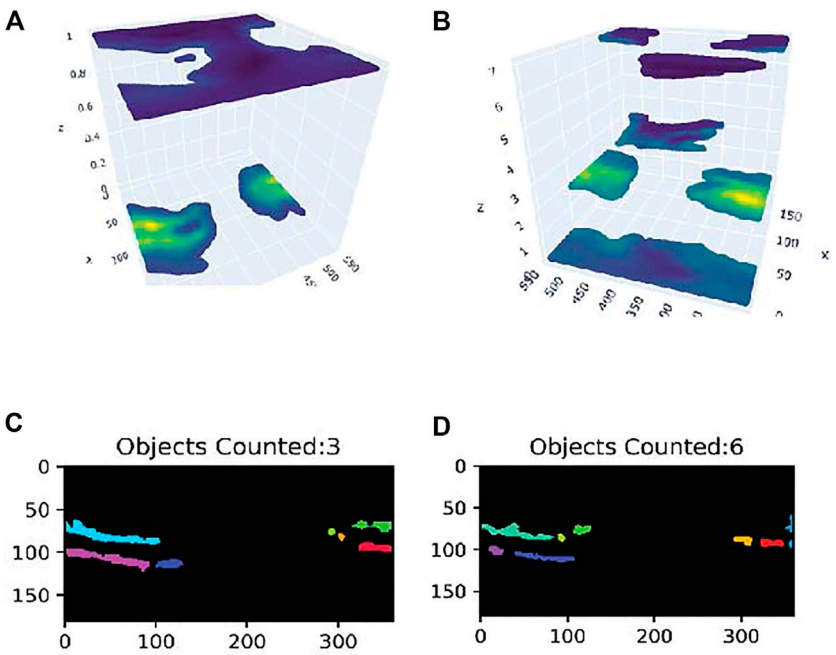

FIGURE 2 | Results of GM method application to HDR identification for a GIM on 27 October, 2002 **(A)** and on 19 July, 2001 **(B)**. Corresponding results for the image processing approach are shown in **(C,D)**. For 27 October, 2002, the GM method identifies 3 HDRs **(A)**, while the image processing approach identifies 3 HDRs **(C)**. For 19 July, 2001, the GM method identifies 5 HDRs **(B)**, while the image processing approach identifies 6 HDRs **(D)**.

COMPUTER VISION METHOD

Large scale TEC dynamics was analyzed by Dmitriev (2018) by applying visual analysis to individual GIMs. This technique allowed to consider detailed dynamics and provided insight into corresponding physical processes. We advocate building upon such visual approaches and develop ways for automated classification of large-scale TEC features that will be applicable for large datasets. Alternatively, the image processing library OpenCV for Python together with edge-enhancing technique was applied to identify HDRs in a selected GIM dataset with visual inspection. This is an improvement upon our image classification approach (Verkhoglyadova et al., 2021). First, for each TEC map, represented by gridded TEC values, we round the float TEC values to integer numbers and linearly scale the TEC values to numbers between 0 and 255. The TEC map is thus converted to a gray-scale image. Second, we apply the Laplacian operator often used to detect edges in an image, to the gray-scale TEC image brightness over the 2D map. Third, going back to the original TEC map, we neglect regions with TEC values smaller than the half of the TEC global maximum and regions with the Laplacian values greater than a threshold chosen after visual testing of a variety of values for the limited number of TEC maps, and then apply the Dilate, Erode, and medianBlur methods from OpenCV. Finally, HDRs on the TEC map are identified and counted by OpenCV's connectedComponents operator. A minimum absolute-value threshold for an "edge sharpness" in an image can be applied in order to focus on regions with significantly higher TEC than the surrounding area. Wide range of thresholds were tested and classification results were qualitatively compared to find an optimal value. We found out that our approach works successfully when there are visually identified sharp edges to a TEC brightening. We are fairly confident in selection of an HDR as relatively bright TEC region (by the TEC magnitude) compared to neighboring regions. Since success of the automated classification relies partially on sharpness of the main features of a TEC map, visual inspection is necessary to correctly identify HDRs. **Figures 2C,D** show examples of identifications of 3 and 6 HDRs on a latitude by longitude map, correspondingly. Introducing an additional procedure of image sharpening allowed to separate two EIAs and identify faint HDRs even if they are not well separated. However, there is a bias in adapting this algorithm to accommodate for visual perception. The identification results are not evident for everyone but for an expert in ionospheric physics and GIMs. Additional complication is encountered when neighboring bright regions are not well separated. We suggest that development of a quantitative criterion of a degree of separation based on statistical properties of TEC distribution is necessary to determine efficiency and applicability of the method.

DISCUSSION

The unexpected result of the study is a realization that different approaches to GIM classification and TEC feature extraction result in different HDR counts and provide different information, each with their own utility for identifying large-scale structure. The GM method is an advanced mixture method that identifies TEC clusters as sub-populations in a GIM by assuming Gaussian distribution of TEC within the clusters. Background TEC is also selected as a separate cluster. GM is a robust method that utilizes optimization tools to select the most common clustering result and account for periodic boundaries. However, it does not always identify visually bright structures inside an extended but less bright structure as separate clusters. Instead, the image classification approach allows for a threshold on TEC value to select the most intense HDRs and ignore the background. The results were validated by visual inspection. However, the latter approach is biased towards sharpening edges of bright features in a map and does not have a selection criterion based on strict statistical properties of TEC distribution. Inter-comparison between these two methods showed different clustering results for several GIMs. Thus, the algorithm based on visual perception of bright regions and the algorithm based on statistical properties of TEC sub-populations in a GIM can produce different outcomes. These results raise an important question of how to robustly define HDRs in GIMs and calls for further investigation. How physical is the definition of distinct HDRs and EIAs? Shall we rely on strict statistics-based approaches or the ones tailored to human eye? How to determine if bright TEC regions are well separated? It is likely that a realistic view of global ionosphere includes HDRs of varying density embedded into backgrounds of varying density that change with solar cycle phases in long term and eruptive solar events in short-term. Depending on the purpose of specific studies, HDRs could be identified using a universal criterion to allow for cross comparison among different background TECs, or using different criteria to make the HDRs outstanding in individual TEC maps. Sharp gradients between different large-scale ionospheric features may not typically occur. Addressing these questions and further research will provide important insights into large -scale ionospheric structure and are crucial for space weather forecast.

DATA AVAILABILITY STATEMENT

The datasets used in this study can be found in https://sideshow.jpl.nasa.gov/pub/iono_daily/gim_for_research/jpli/.

AUTHOR CONTRIBUTIONS

The authors equally contributed to the ideas presented in the manuscript. XM developed software for the image processing approach. JK developed software for mixture method approach. OV took a lead on writing the manuscript.

ACKNOWLEDGMENTS

Portions of work were performed at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with NASA. Sponsorship of the Heliophysics Division of the NASA Science Mission Directorate is gratefully acknowledged.

REFERENCES

Astafyeva, E., Zakharenkova, I., and Huba, J. D. (2017). "Three-Peak Ionospheric Equatorial Ionization Anomaly: Development, Drivers, Statistics," in Proceedings of the 15th ionospheric effects symposium, Alexandria, VA, May 9–11, 2017.

Astafyeva, E., Zakharenkova, I., and Pineau, Y. (2016). Occurrence of the Dayside Three-Peak Density Structure in the F2 and the Topside Ionosphere. *J. Geophys. Res.* 121 (6936), 6936–6949. doi:10.1002/2016ja022641

Dmitriev, A. (2018). Spatial Characteristics of Recurrent Ionospheric Storms at Low Latitudes during Solar Minimum. *J. Atmos. Solar-Terrestrial Phys.* 179, 553–561. doi:10.1016/j.jastp.2018.09.013

Hernández-Pajares, M., Roma-Dollase, D., Krankowski, A., García-Rigo, A., and Orús-Pérez, R. (2017). Methodology and Consistency of Slant and Vertical Assessments for Ionospheric Electron Content Models. *J. Geod* 91, 1405–1414. doi:10.1007/s00190-017-1032-z

Maruayama, N., Sun, Y.-Y., Richards, P. G., Middlecoff, J., Fang, T.-W., Fuller-Rowell, T. J., et al. (2016). A New Source of the Midlatitude Ionospheric Peak Density Structure Revealed by a New Ionosphere-Plasmasphere Model. *Geophys. Res. Lett.* 43 (2429), 2429–2435. doi:10.1002/2015GL067312

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-Learn: Machine Learning in Python. *J. Machine Learn. Res.* 12 (85), 2825–2830. Available at: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html.

Roma-Dollase, D., Hernández-Pajares, M., Krankowski, A., Kotulak, K., Ghoddousi-Fard, R., Yuan, Y., et al. (2018). Consistency of Seven Different GNSS Global Ionospheric Mapping Techniques during One Solar Cycle. *J. Geod* 92, 691–706. doi:10.1007/s00190-017-1088-9

Schunk, R. W., and Nagy, A. (2009). *Ionospheres: Physics, Plasma Physics, and Chemistry*. Cambridge: Cambridge University Press.

Verkhoglyadova, O., Maus, N., and Meng, X. (2021). Classification of High Density Regions in Global Ionospheric Maps with Neural Networks. *Earth Space Sci.* 8, e2021EA001639. doi:10.1029/2021ea001639

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Astronomy and Space Sciences

ORIGINAL RESEARCH
published: 02 May 2022
doi: 10.3389/fspas.2022.846291

Multi-Variate LSTM Prediction of Alaska Magnetometer Chain Utilizing a Coupled Model Approach

Matthew Blandin[1], Hyunju K. Connor[1,2], Doğacan S. Öztürk[1], Amy M. Keesee[3], Victor Pinto[3], Md Shaad Mahmud[4], Chigomezyo Ngwira[5] and Shishir Priyadarshi[6]*

[1]Department of Physics and Geophysical Institute, University of Alaska, Fairbanks, AK, United States, [2]NASA Goddard Space Flight Center, Greenbelt, MD, United States, [3]Department of Physics and Astronomy and Space Science Center, University of New Hampshire, Durham, NH, United States, [4]Department of Electrical and Computer Engineering, University of New Hampshire, Durham, NH, United States, [5]ASTRA, Boulder, CO, United States, [6]Department of Electronics and Electrical Engineering, University of Bath, Bath, United Kingdom

During periods of rapidly changing geomagnetic conditions electric fields form within the Earth's surface and induce currents known as geomagnetically induced currents (GICs), which interact with unprotected electrical systems our society relies on. In this study, we train multi-variate Long-Short Term Memory neural networks to predict magnitude of north-south component of the geomagnetic field ($|B_N|$) at multiple ground magnetometer stations across Alaska provided by the SuperMAG database with a future goal of predicting geomagnetic field disturbances. Each neural network is driven by solar wind and interplanetary magnetic field inputs from the NASA OMNI database spanning from 2000–2015 and is fine tuned for each station to maximize the effectiveness in predicting $|B_N|$. The neural networks are then compared against multivariate linear regression models driven with the same inputs at each station using Heidke skill scores with thresholds at the 50, 75, 85, and 99 percentiles for $|B_N|$. The neural network models show significant increases over the linear regression models for $|B_N|$ thresholds. We also calculate the Heidke skill scores for $d|B_N|/dt$ by deriving $d|B_N|/dt$ from $|B_N|$ predictions. However, neural network models do not show clear outperformance compared to the linear regression models. To retain the sign information and thus predict B_N instead of $|B_N|$, a secondary so-called polarity model is utilized. The polarity model is run in tandem with the neural networks predicting geomagnetic field in a coupled model approach and results in a high correlation between predicted and observed values for all stations. We find this model a promising starting point for a machine learned geomagnetic field model to be expanded upon through increased output time history and fast turnaround times.

Keywords: space weather, GIC, geomagnetic storms, ground geomagnetic field, machine learning, neural networks, LSTM

OPEN ACCESS

Edited by:
Peter Wintoft, Swedish Institute of Space Physics, Sweden

Reviewed by:
Andrew Smith, University College London, United Kingdom
Octav Marghitu, Space Science Institute, Romania

****Correspondence:***
Matthew Blandin
mjblandin@alaska.edu

Specialty section:
This article was submitted to Space Physics, a section of the journal Frontiers in Astronomy and Space Sciences

Received: *31 December 2021*
Accepted: *24 March 2022*
Published: *02 May 2022*

Citation:
Blandin M, Connor HK, Öztürk DS, Keesee AM, Pinto V, Mahmud MS, Ngwira C and Priyadarshi S (2022) Multi-Variate LSTM Prediction of Alaska Magnetometer Chain Utilizing a Coupled Model Approach. Front. Astron. Space Sci. 9:846291. doi: 10.3389/fspas.2022.846291

INTRODUCTION

Geomagnetically induced currents (GICs) are produced when the solar wind interacts with the Earth's magnetic field, driving disturbances that map to the Earth's surface (Oliveira and Ngwira, 2017). Electrically conductive materials, like the Earth's crust, in the presence of these

disturbances experience electric fields proportional to that of the changing geomagnetic field which are known as geomagnetically induced electric fields (Pirjola, 2000) and drive currents, GICs, as a response. These GICs, if strong enough, can disrupt and damage sensitive electrical devices on the ground that are not designed to handle these currents. GICs have been known to cause power outages, transformer damage, and pipeline corrosion on the ground which impacts our technology and fossil fuel dependent economy; such an event was the cause for a 9 h power grid blackout in Quebec, Canada on 13 March 1989, where strong GICs overloaded and damaged a transformer of the Hydro-Quebec electric company.

In response to the damage done by these events the science community has put focus on the prediction of GICs. While GICs can happen in any part of the globe, there is a higher occurrence of these events in higher magnetic latitude regions. Large geomagnetic field disturbances are often observed in these regions due to geomagnetic field lines at the surface connected to dynamic regions of the magnetosphere (e.g., polar cusps and the magnetotail). During times of high geomagnetic activity, strong geomagnetic field disturbances may propagate lower into middle magnetic latitudes, leading to a majority of GIC studies being focused on high and mid magnetic latitudes (Pirjola, 2005; Pulkkinen et al., 2005; Pirjol et al., 2007; Fiori et al., 2014; Blake et al., 2016). Some studies have also shown geomagnetic field disturbances in the low magnetic latitudes formed from oblique pressure shocks (Carter et al., 2015; Zhang et al., 2016; Oliveira et al., 2018). The widespread nature of geomagnetic field disturbances in response to fluctuating solar wind parameters can also be seen in **Supplementary Movie S1** of the supplemental material, where a global response to a geomagnetic storm is observed, with the strongest field fluctuations located at high magnetic latitudes in the midnight sector, reaffirming the focus on high latitude GICs.

Various mechanisms are known to cause large geomagnetic field disturbances on the ground. The study by Carter et al. (2015) has shown interplanetary shocks being a viable creation mechanism for large geomagnetic field disturbances at high magnetic latitudes and the magnetic equator. Recent studies from Heyns et al. (2021) and Rogers et al. (2020) analyzed GICs as a function of geomagnetic pulsations, indicating that ULF waves can drive GICs for extended periods at high and mid latitudes. The studies by Rodger et al. (2017) and Dimmock et al. (2019) identified extreme geomagnetic storm activity and sudden geomagnetic storm commencement as drivers of GICs in the New Zealand and Fennoscandia regions. However, the studies by Ngwira et al. (2015) and Dimmock et al. (2020) have shown the timing of GICs in relation to geomagnetic storm activity can vary based on location. Local dB/dt is a function of ionospheric and magnetospheric currents, generally associated with geomagnetic storm activity and local conductivity gradients within the Earth's crust.

To understand these localized peaks, high resolution physics-based models have been utilized to determine these fluctuations (Welling, 2019), however these models, while important in the progress of our understanding, are computationally expensive and time consuming, which make them inefficient for real-time predictions of GICs. The need for efficient and computationally inexpensive models has led to the utilization of machine learned neural networks, such as the ones done by Wintoft et al. (2015), Lotz and Cilliers (2015), and Keesee et al. (2020), to capture these disturbances from upstream drivers, mainly the solar wind parameters.

Machine learned algorithms are efficient due to their fast computation times after training and range in complexity defined by the user, allowing for versatile solutions. The bulk of computational requirements needed for most neural networks are during training, with the finished models being significantly lightweight at runtime. The lightweight models this study aims to produce are focused on the Alaska region, which is a high magnetic latitude area susceptible to pipeline corrosion from GICs (Gummow and Eng, 2002; Pirjola et al., 2003; Khanal et al., 2019; Liu et al., 2019). The models produced for this region make use of real-time magnetometer data that are available locally, allowing for an enhancement in model performance.

The first step in forecasting GICs in Alaska starts with a geomagnetic field prediction model utilizing 16 years of SuperMAG geomagnetic field observations across Alaska and NASA OMNIweb solar wind and interplanetary magnetic field (IMF) conditions. The data is used as input to three different model types: a multi-variate LSTM model, a multi-variate linear regression (MLR) model, and a coupled set of LSTM models. In the following section the data sources and different model types will be explained. **Sections 3, 4** will cover the model results and a discussion on the effectiveness of these models, their shortcomings, and routes of improvement, followed by a summary of the work presented.

DATA AND MODELS

SuperMAG and OMNIweb Data

The study presented utilizes the SuperMAG magnetometer database and the NASA OMNIweb solar wind database from 01/01/2000 to 12/31/2015. The data selected from OMNIweb database provides solar wind plasma and interplanetary magnetic field (IMF) propagated from solar wind monitors at Lagrangian point 1 to a subsolar bow shock location. The data provided from OMNIweb are solar wind density, flow speed, dynamic pressure, temperature, IMF magnitude, and IMF B_Z in geocentric solar magnetic (GSM) coordinates with a 1 min resolution. The parameters were selected from a combination of known indicators of geomagnetic storms and substorms utilized in similar studies such as Lotz and Cilliers (2015) and Keesee et al. (2020). The use of derived parameters (i.e., parameters calculated from other data values, such as dynamic pressure) was limited to avoid data redundancy within the neural network which can result in poor performance within these models. Due to a non-continuous data coverage, a linear interpolation was applied to fill in gaps of 10 min or less, increasing the coverage from 70 to 80% during the 16 years segment.

The SuperMAG database hosted by Johns Hopkins University (Gjerloev, 2012) collects data from world wide magnetometer stations and provides baseline removed geomagnetic field data with a consistent coordinate system. From SuperMAG we selected four Alaska stations at geographic coordinates: Fort Yukon (FYU) at 66.56° N 214.87° E, College (CMO) at 64.87° N 212.15° E, Poker Flat (PKR) at 65.12° N 212.57° E, and Kaktovik (KAV) at 70.14° N 216.35° E; and utilized the north-south component of the observed field with 1 min resolution. Geomagnetic field information from SuperMAG is in their local magnetic coordinate system where the Z-component is kept static and the other two components are rotated to maximize North-South and minimize East-West with respect to a slowly varying declination angle. The magnetic field datasets utilized have had standard yearly and daily baseline removal as detailed by Gjerloev (Gjerloev, 2012). The data from CMO was used as an initial testing set to determine the best performing model configuration, and then the configuration was re-applied to the other three stations. The stations were chosen for their locations, which are roughly on a magnetic meridional line, making them perpendicular to the typical auroral oval in Alaska. This property of the stations chosen makes them suitable for ionospheric current predictions above Alaska.

LSTM Model

A basic recurrent neural network (RNN) takes incoming data, computes the output, and sends the output as an input parameter to be used with the next incoming dataset (Brownlee, 2017). This means that each output is directly reliant on the previous known output and the newest data, making RNN able to predict time dependent sequences where the features act on a small time scale. The downside to RNN is that features acting on a long time scale are not properly accounted for, since the newest information is always the most relevant in the prediction process. These neural networks also rely on continuous datasets, treating a known gap in data as a standard time step in progression. Further, these networks are prone to vanishing or exploding error gradients during trainings which can result in excessive computation for minimal gains in performance.

For datasets where the input features have long or variable time scale implications RNN are not suitable, as the network will forget the information. To combat this, advanced RNNs, such as LSTM, are developed and utilized to retain information on a longer time scale (Hochreiter and Schmidhuber, 1997). LSTM achieves this through the use of two internal features known as gates and the cell state. Within the LSTM kernel three gates are utilized, these gates determine which information is added and removed from the cell state, and determine which parts of the incoming data are relevant to the current prediction. Unlike the standard RNN, where the output prediction is fed to the next iteration, the cell state is passed to the next LSTM kernel, and is used in unison with the incoming dataset to make a prediction. Due to the activation of the gates, a cell state may not be updated every sequence, allowing it to retain information from previous datasets. This model type has been successfully applied for predicting magnetometer data (Keesee et al., 2020).

Linear Regression

Linear regression is a widely used approach for many empirical models (Verbeek, 2017). This study developed a MLR model for each station for comparison against the state-of-the-art machine learning models. The MLR models developed utilize the same inputs and outputs as the LSTM and are applied with the same 16 years of SuperMAG and OMNIweb data to fit the following equation:

$$|B(t)| = \sum_{i} \sum_{n=1}^{30} \alpha_{in} x_i (t-n) + C \tag{1}$$

where $x_i(t-n)$ is an input variable from n minutes prior, C is an offset variable, and α_{in} is a scaling factor for variable $x_i(t-n)$. This model type is applied to each station and compared to their respective LSTM models. We chose MLR for comparison due to its ease of implementation and wide use for empirical studies.

MODEL DEVELOPMENT AND RESULTS

Geomagnetic Field Prediction Model Using CMO Dataset

The initial phase of models produced was aimed at determining the length of time history of each feature to provide to the LSTM model. Four models were trained off a smaller 2009–2014 dataset utilizing 1-, 30-, 60-, and 120- min time histories as input for predicting $|B_N|$ at the next minute and tested for performance against the 2015 test set. Each model takes the same basic inputs of IMF magnitude, its Z-component in GSM coordinates, solar wind density, speed, flow (ram) pressure, temperature, and magnetic local time (MLT) of the station. The MLT has been converted into sin (MLT) and cos (MLT) to maintain a cyclic dependence of this variable, which is important for preserving nighttime and daytime dependencies and transitions. While the LSTM kernel is adept at remembering via the implemented cell state, the previous 1-, 30-, 60-, and 120 min of each variable were used as input in a 2D array of shape [m x n], where m is the number of features and n is the amount of time history, as seen in **Figure 1**. For consistency, each model utilized a single LSTM layer with 32 hidden neurons, a rectified linear unit (ReLU) activation layer, a 68.75-25-6.25 training-validation-testing set split, adaptive moment estimation (ADAM) optimizer, a learning rate of 0.001, mean squared error loss, cube root normalization, 360 batches during training before updating weights, and a single unit dense layer to pass the final output. One may think to stack multiple LSTM networks to achieve better performance, however, in testing, the use of 2 or more LSTM layers degraded performance rather than enhancing it. Finally, each of the models was trained for up to 100 epochs, however the training implemented early stopping, occurring at around 20 epochs on average, based on the validation loss statistic and saved only the best model during the training to avoid overfitting the model to the dataset. From this testing we are able to determine the best amount of time history to train with, aiming to supply immediate information to LSTM with the known previous solar wind data. One minute time history was chosen as a starting

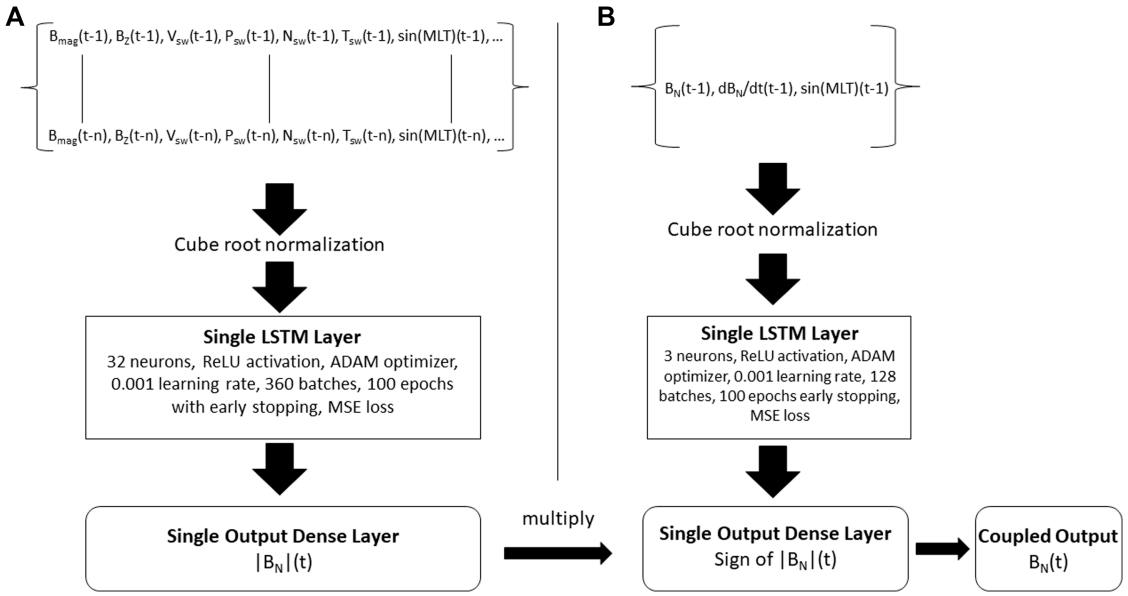

FIGURE 1 | Subfigure **(A)** showcases the setup of the LSTM $|B_N|$ model utilized in the study. The inputs are structured in a 2D array where each column is one feature of the solar wind and IMF inputs detailed in **Section 2.2** while the rows refer to each of the features at a previous time point from t–1 to t–n, where n is the amount of time history selected. The input array is provided directly to the LSTM model which sends the output to a single dense layer to provide the final output. In subfigure **(B)** the polarity model is displayed in a similar manner to the LSTM model of subfigure **(A)**. The input array is a 1D array sent to an LSTM model which sends the final prediction to a single dense layer. The output of the polarity model is decoded and multiplied against the LSTM $|B_N|$ model to obtain the coupled output.

choice because it is the standard LSTM setup. The other 30-, 60-, and 120-min time histories are selected in consideration of variable propagation of solar wind and IMF information from the subsolar bow shock to geomagnetic activity (Connor et al., 2014; Maggiolo et al., 2017). **Figures 2A–G** compares geomagnetic field predictions of the four LSTM models described for the 07-Sep-2015 geomagnetic storm. **Figures 2A–E** show IMF in GSM coordinates, solar wind speed, density (black) and flow pressure (red), temperature, and AE (black) and SYM/H (red) geomagnetic indices. **Figure 2F** shows the magnitude of north-south component of the geomagnetic field ($|B_N|$) as observed from CMO (black) and predicted by the LSTM models using different time histories as input (dashed lines). **Figure 2G** shows the time derivative of the north-south geomagnetic field component ($d|B_N|/dt$) observed from CMO (black) and the value derived from the LSTM $|B_N|$ models (dashed lines) with gaps in predictions occurring where one or more input variables are missing due to a gap larger than 10 min. **Figures 2H–K** show other modeling experiments and are discussed later in the paper. In **Figure 2F** we can see that the performance between the four models for the event are quite close in $|B_N|$ predictions. The 1 min time history model shows a drop in $|B_N|$ around the 1600 UT mark whereas the other 3 models show a drop occurring an hour later. When looking at the $d|B_N|/dt$ values we find the 30 min model shows the most activity, though vastly underestimated, where the other 3 models predict closer to 0.

Individual time history model performance is also seen in **Figures 3A–D**, where the predicted and observed value for the CMO station across the year of 2015 are plotted for correlation. In **Figure 3** the dashed red line indicates a perfect correlation between the prediction and observed values and the solid red line shows the line of best fit. The legend in the upper right corner shows the equation for the line of best fit and the Pearson correlation (r) between the predicted and observed data. In these plots we find the Pearson correlation (r) between the 4 models is relatively similar with small increases corresponding with time history supplied to the model. Looking at these values we see an 11% increase in correlation from the 1 min to the 30 min model. As the input data increases we find an 8% increase when doubling the time history from 30 to 60 min and a 2% increase when increasing time history from 60 to 120 min. Looking at the lines of best fit (solid red) we can see a slow shift towards the perfect correlation (dashed red) line as input data increases, from 1 min time history to 120 min. The 30-min time history model was selected as the suitable model due to a reduction in model complexity with nearly similar results as the more complex 120-min model. Further, we find the 30-min time history more suitable for our long term application of GIC predictions due to its performance in event time $d|B_N|/dt$ fluctuations as indicated by **Figure 2G**.

In **Figures 2H,I** the 30-min time history model was selected and trained for $|B_N|$ with (red line) and without (green line) previous known geomagnetic field time derivative ($d|B_N|/dt$) history as an input parameter. This is also found in **Figures 3E,F**, which show an increase in Pearson correlation between the predicted and observed CMO geomagnetic field values when $d|B_N|/dt$ is included from 0.50 to 0.87. The use of $d|B_N|/dt$ as in input variable is spurred by the standard LSTM setup where

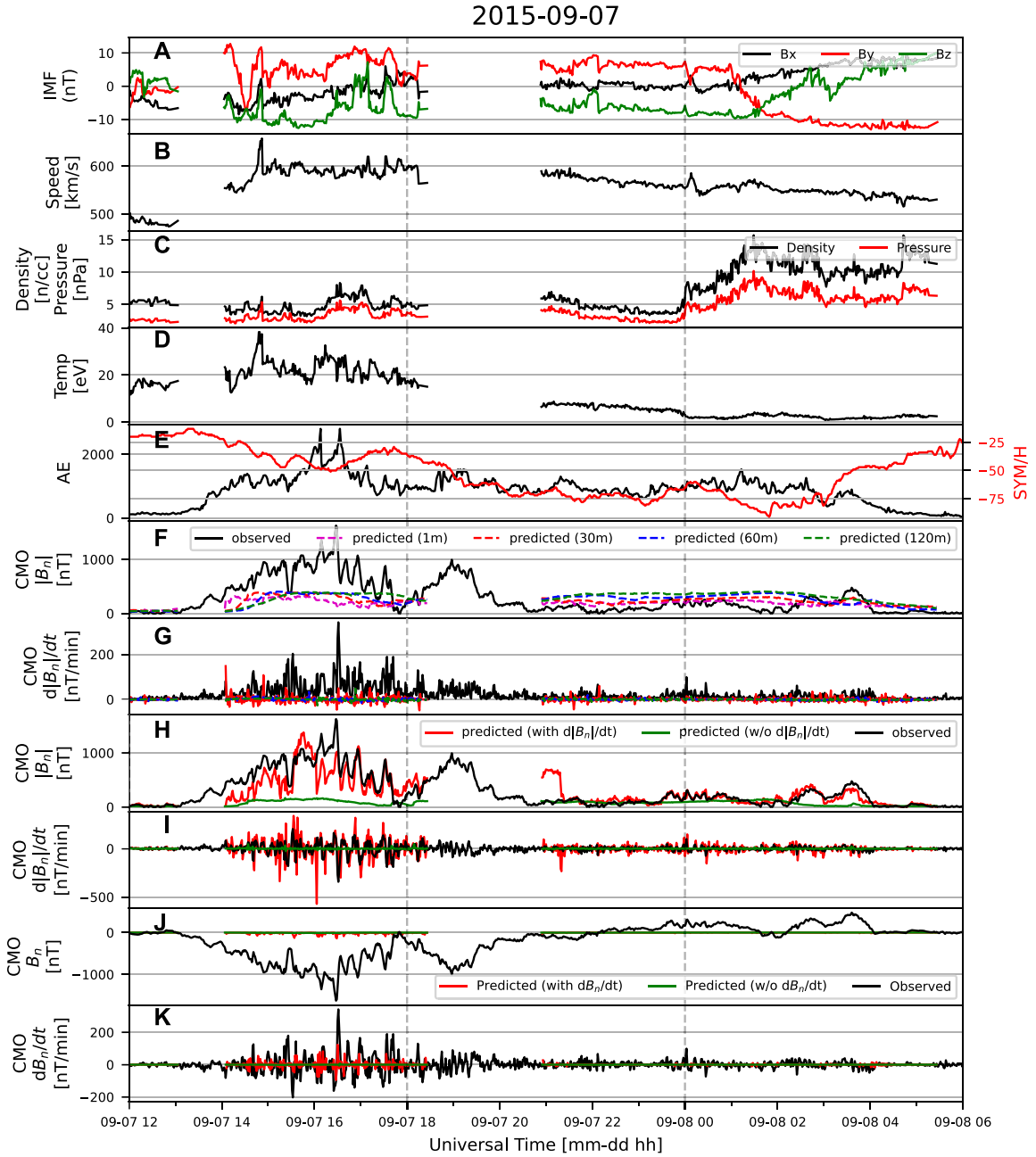

FIGURE 2 | Storm time test predictions of different LSTM configurations for the 07-Sep-2015 storm. Subplots **(A–E)** show the IMF, solar wind, and geomagnetic indices conditions with the model predictions in the subplots beneath. Subplots **(F,G)** show the results of predictions utilizing different time histories of input. From these we see the 1-min time history model consistently predicts values less than that of the other 3 models. Subplots **(H,I)** show the results of predicting for $|B_N|$ utilizing the 30-min time history both with and without $d|B_N|/dt$ as an input parameter. Likewise, subplots **(J,K)** utilize the 30-min time history model to predict B_N with and without dB_N/dt information added as input. From these plots we find that LSTM has an affinity for predicting $|B_N|$ with $d|B_N|/dt$ information included in the input parameters.

previous observations variables are passed as a standard input for the upcoming prediction. In the case of our model, passing $|B_N|$ at t-1 information creates a feed forward network where the model passes a value with high correlation to $|B_N|$ at t-1 as the prediction. The use of $d|B_N|/dt$ removes this outcome while providing the model with general information regarding the strength of fluctuations. We acknowledge concern that by using $d|B_N|/dt$ as input, the LSTM models may act as a first-order Taylor series expansion (i.e., $|B_N|$ (t) = $|B_N|$ (t-1) + $d|B_N|/dt$ (t-1)*dt. However, our approach differentiates from this expansion by not utilizing $|B_N|$ at t-1 as input. Additionally, it uses 30 min of time history of $d|B_N|/dt$ and SW/IMF parameters as input. One

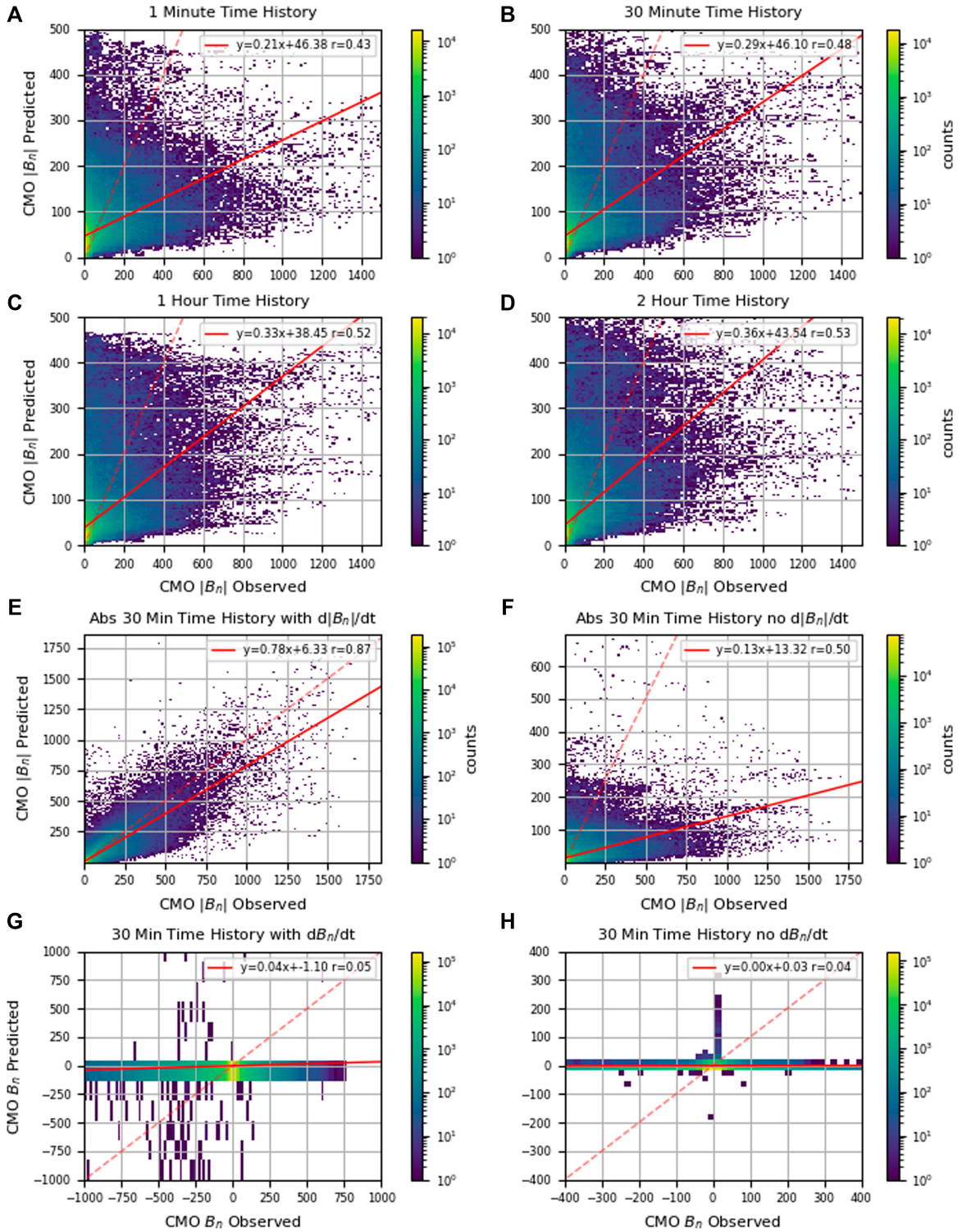

FIGURE 3 | Storm time density plots using 2015 as a test set showing the performance of the models shown in **Figure 2**. The dashed red line indicates a perfect 1:1 correlation between predicted and observed values. The solid red line is a line of best fit correlation between the predicted values and the observed values. Subplots **(A–D)** show density plots of predicted vs observed values for 2015 CMO when utilizing 1–, 30–, 60–, and 120-minutes of time history. Subplots **(E–H)** utilize the selected 30-minute time history model and compare predicting $|B_N|$ with and without $d|B_N|/dt$ information and predicting B_N with and without dB_N/dt information. We can see in subplots **(G)** and **(H)** that LSTM predicting B_N provided significantly poor predictions with a correlation of 0 regardless of utilizing previous dB_N/dt information as an input to the model. However, when predicting $|B_N|$ as seen in subplots **(E)** and **(F)** LSTM performs well, with a significant 74% increase in Pearson correlation when including previous $d|B_N|/dt$ information as an input parameter to the model.

may find concern that the LSTM model will find high correlation between prediction $|B_N|$ (t) and the input feature $d|B_N|/dt$ (t-1), however in testing our model does not show such behavior, suggesting the LSTM model learns more complex behaviors than the first order Taylor series expansion.

Lastly, **Figures 2J,K** show the results of two 30-min time history models trained with (red line) and without (green line) dB_N/dt input predicting B_N. Both of these models performed poorly, predicting a nearly straight line at 0 nT, with density plots in **Figures 3G,H** indicating no correlation between the predicted and observed values for CMO in 2015 while predicting B_N. From the eight models tested we find that the 30-min model with $d|B_N|/dt$ input and $|B_N|$ output retains the best performance regarding predictions during storm time while limiting the complexity of the model. The model configuration is then utilized across all 4 stations with the 2000–2015 training set.

Geomagnetic Field Prediction Across the Alaska Chain

The magnetometer chain of FYU, CMO, PKR, and KAV were chosen to create a perpendicular line prediction of the geomagnetic field with respect to the auroral oval. For each of the stations the 30-min time history was chosen for the performance found in **Section 3.1**. The models employ the same early stopping mechanism described in 3.1 to avoid overfitting of the data. While the models in **Section 3.1** were trained from a smaller limited dataset (i.e., 2009–2015), the models trained for each station utilized data from January 2000 to December 2015 for training with 68.75% of the data as the training set, 25% as the validation set, and 6.25% as the test set. Like the models in **Section 3.1**, the year of 2015 was separated and used as a testing set of the models due to its high prevalence of geomagnetic activity throughout the year. Further, for each station an additional MLR model was made with the same 2000–2015 dataset, utilizing the same input variables and cube root normalization as the LSTM. In some cases the MLR models predicted negative values of $|B_N|$, which would imply an un-physical value of B_N. The negative values were removed from the MLR predictions dataset as they are un-physical quantities for $|B_N|$.

In **Figure 4** the results of the models are plotted for the 07-Sep-2015 geomagnetic storm and subplots **Figures 4A–E** follow the same format as **Figures 2A–E**. In **Figures 4F,G** we see the predictions for the LSTM model (red) and MLR model (green) compared to the observed values (black) for the FYU station with **Figure 4F** showing the $|B_N|$ predictions and **Figure 4G** showing $d|B_N|/dt$ predictions. Likewise, **Figures 4H–M** show the $|B_N|$ and $d|B_N|/dt$ prediction results for PKR, CMO, and KAV, respectively. From these figures we see that the MLR models underestimate $|B_N|$ more than the LSTM models. When considering the $d|B_N|/dt$ plots we find that the LSTM shows more frequent strong fluctuations, with the caveat they are not always predicted at the correct time. In **Figure 5** the LSTM and MLR model predictions are plotted against the observed station values for the year of 2015 in the same format as **Figure 3** to test for correlation. The LSTM models in **Figures 5A,C,E,G** show high Pearson correlation coefficients of 0.80–0.86, while the MLR models in **Figures 5B,D,F,H** show much lower correlation coefficients of 0.68–0.73. Looking at the average of the Pearson correlation values, LSTM shows a 18.2% increase in performance over the MLR models.

Polarity and Coupled Model

One of the problems with the LSTM model is its affinity towards predicting $|B_N|$ over B_N. This means that to obtain the best results we train off the magnitude, thus losing sign information in the process. This lost feature, which we are calling polarity, is a necessary component of the ionospheric current modeling, since ionospheric current directions (i.e., eastward or westward electrojets) can be inferred by the sign of B_N. Initially, LSTM seemed to favor predicting when all values in the observation set were positive. To resolve this problem, we applied multiple scaling methods to the dataset when predicting B_N. These methods comprised of scaling the data between 0 and 1, −1 and 1, and linearly shifting the B_N data above 0 by adding the minimum B_N value to each data point. We developed the LSTM models for predicting the scaled B_N values and reversed the scaling back to the original scale for the finalized prediction. However, these scaling approaches did not show any significant improvement when compared to the predictions of the original cube root normalized dataset. Therefore, a different approach to polarity utilizing a secondary model was developed. To create this model, the polarity was encoded into a 1 for positive values and 0 for negative values, then this observed polarity was trained for using a LSTM kernel looking at the previous minute of MLT, $|B_N|$, and $d|B_N|/dt$ information. With the polarity retained through a secondary model we are able to decode the polarity and multiply it through the geomagnetic field model trained for predicting $|B_N|$ and retain B_N in a coupled model technique. **Figure 1** summarizes the setup of the polarity and coupled models.

Figure 6 shows the results of the coupled model approach applied to all stations for the 09-07-2015 test storm. In the left column we can see the predicted values follow the observed data values, with one false positive prediction around 18 UT for the CMO station. As seen in the right column of **Figure 6**, the Pearson correlation between predicted and observed geomagnetic field increased for all stations from an average 0.84 to 0.88. This indicates improved performance of a coupled model than the LSTM in **Figure 5** while preserving the critical sign information needed for ionospheric current analysis. Further, B_N fluctuations observed in **Figure 6** show persistence of negative enhancements for the 09-07-2015 storm which the coupled model properly captures with minimal unexpected sign flips. For KAV we can see around 14 UT and 17 UT the coupled model properly flips positive for the short positive durations seen in the observed data during a predominately negative enhancement. This makes the coupled model approach a promising modeling method for our study.

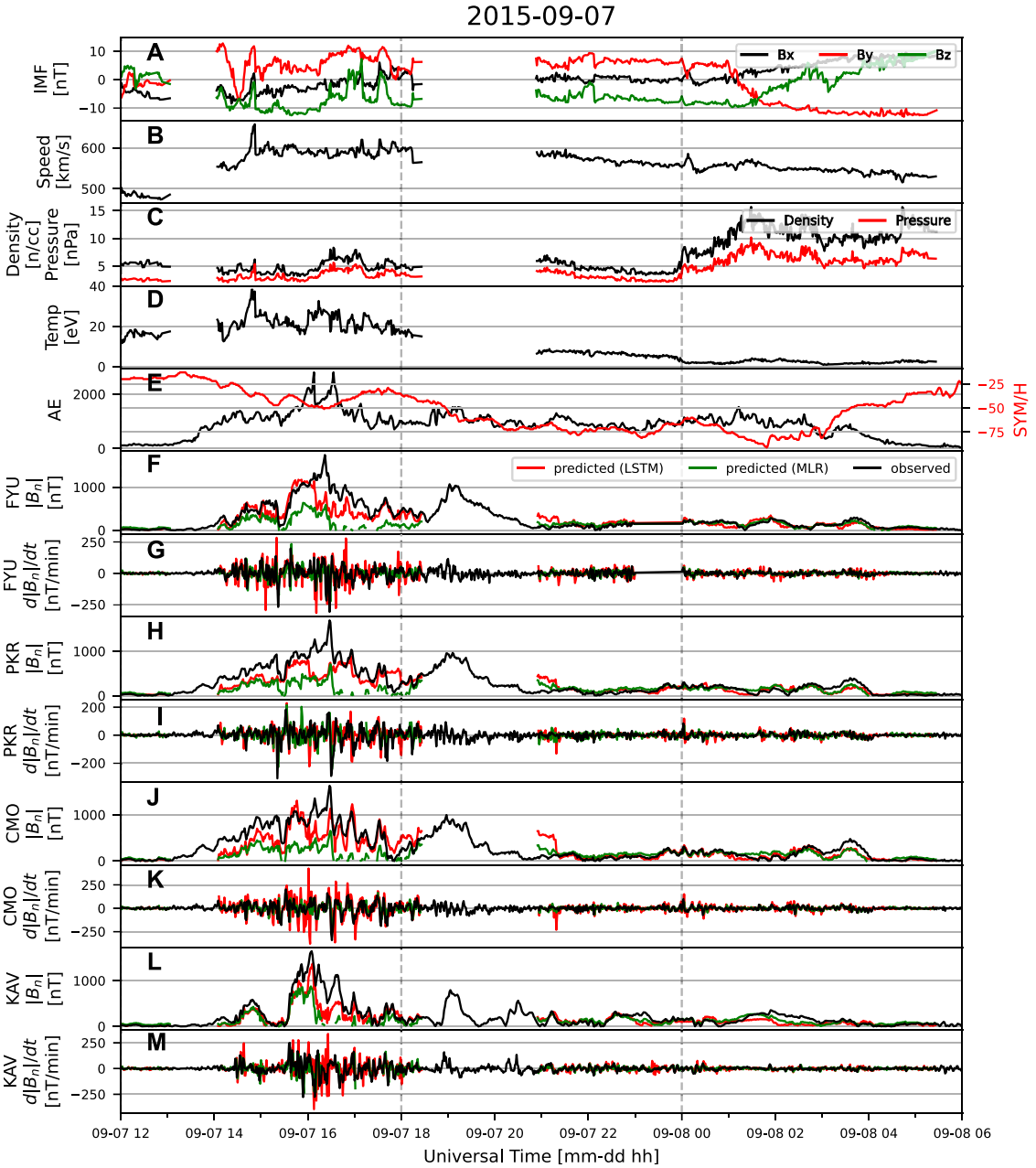

FIGURE 4 | Storm time predictions utilizing LSTM (red) and MLR (green) models showing their ability to predict the 07-Sep-2015 event for each of the 4 individual stations across Alaska. Subplots **(A–E)** show IMF and solar wind properties for the event while subplots **(F–M)** show the observed and predicted $|B_N|$ and $d|B_N|/dt$ for the four selected stations. Here we can see that for all stations the LSTM model performs better at matching $|B_N|$.

DISCUSSION

Skill Scores and Model Performance

Heidke skill scores (HSS) are a widely accepted method of determining machine learned model performance by testing multiple thresholds to understand model sensitivity and variability at discerning the desired output variable. These scores can be see in **Table 1** ranging from 0 (randomprediction) to 1 (perfect prediction) and have been split between scores for $|B_N|$ and $d|B_N|/dt$ sensitivity. The scores are calculated based on whether the predicted and observed values cross the threshold at the same time, with thresholds for $d|B_N|/dt$ selected based off of Pulkkinen et al. (Pulkkinen et al., 2013) and thresholds for $|B_N|$ selected from the 50, 75, 85, and 99 percentile of the $|B_N|$ values over 2000–2014 for the 4 stations. We can see from **Figure 5** and **Table 1** that the LSTM models show a 18.2%

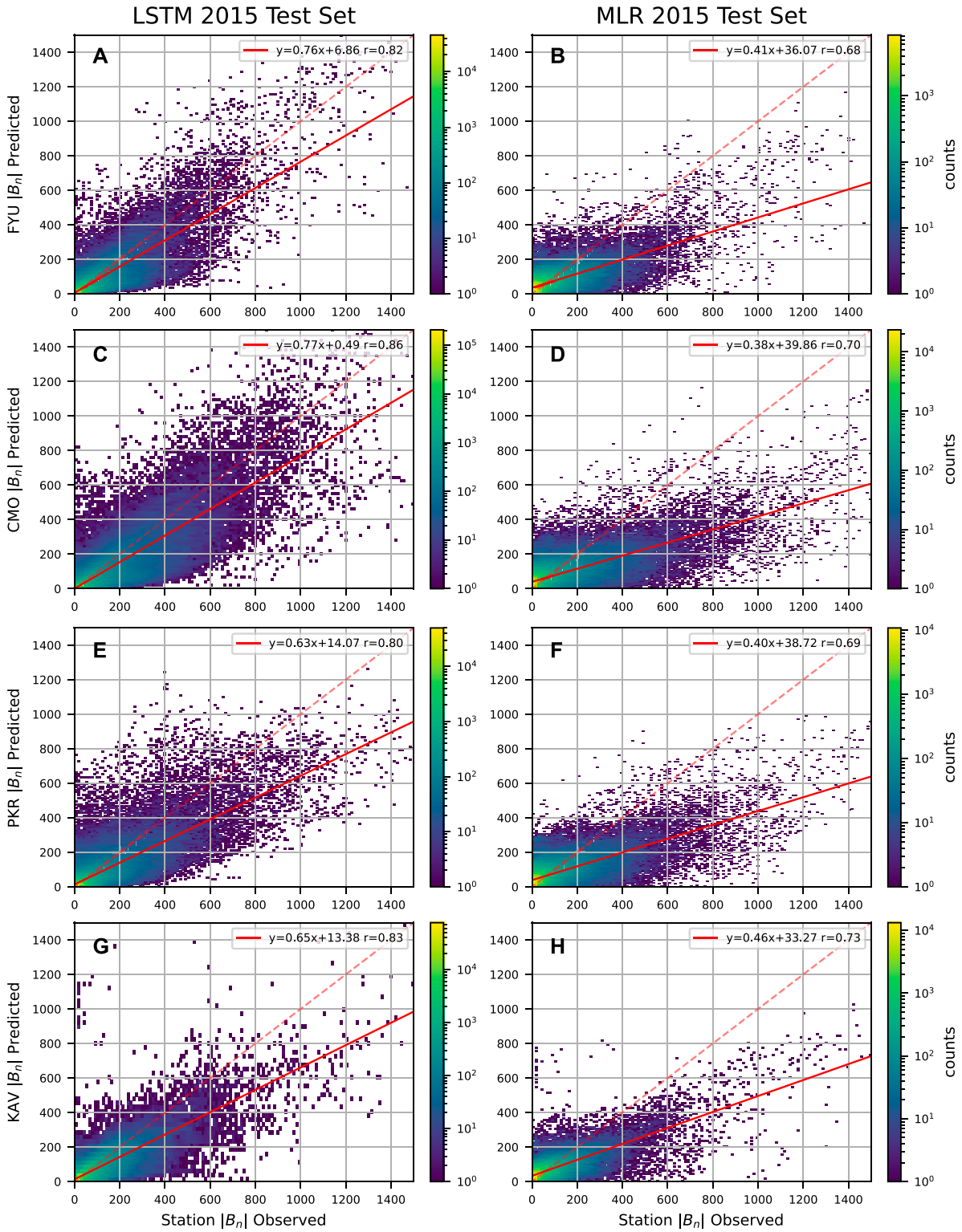

FIGURE 5 | LSTM and MLR density plots using the 2015 test set for all 4 stations. Subplots **(A, C, E, G)** show prediction results for the LSTM models of each station while subplots **(B, D, F, H)** show prediction results for the MLR models of each station. Further, we can see that the lines of best fit (solid red) between the LSTM predictions and observed data are much closer to the perfect prediction lines (dashed red) than those of the MLR predictions and the observed data with a 18.2% increase in the average Pearson correlation values.

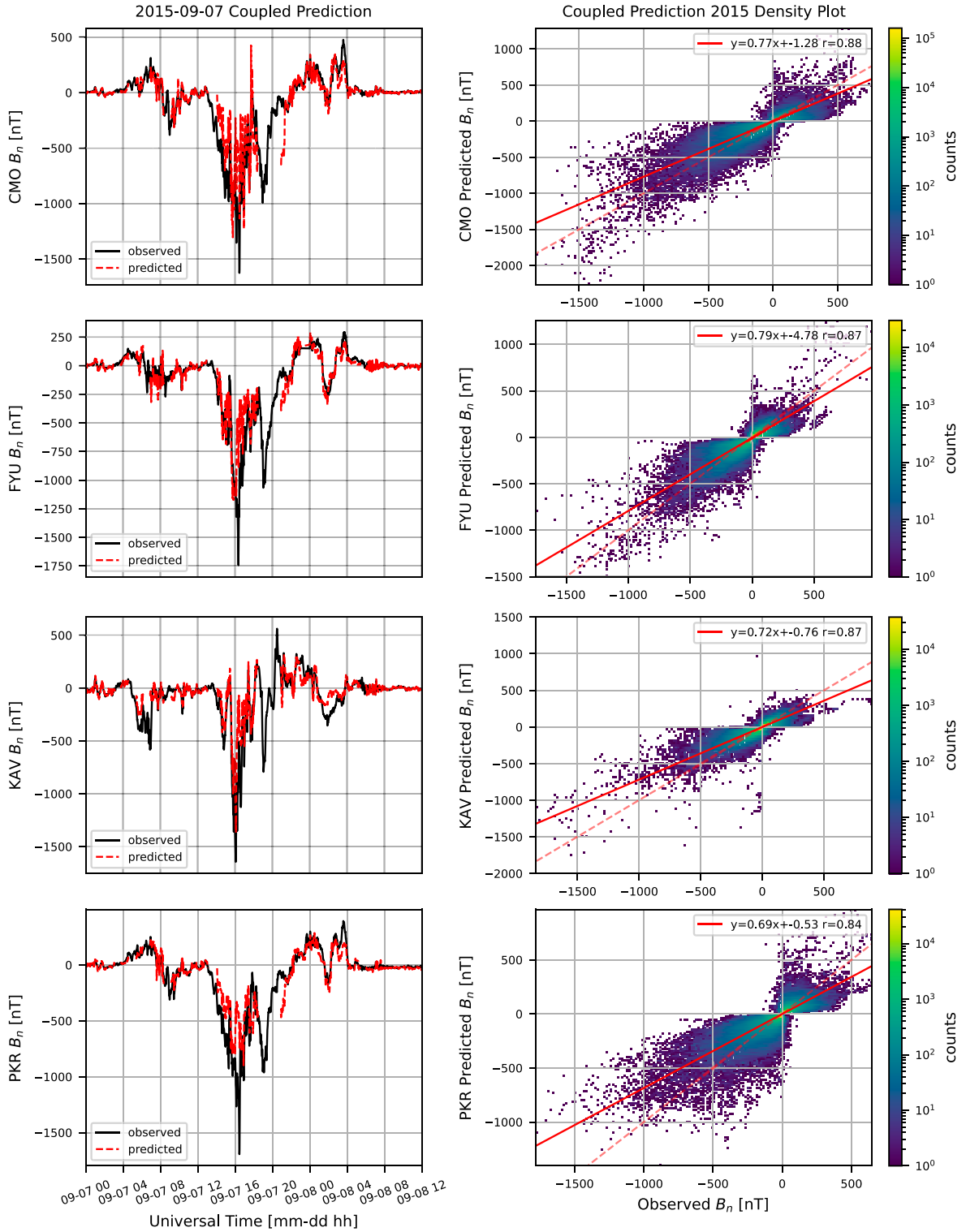

FIGURE 6 | Coupled model storm predictions. In the left column the coupled models ability to predict B_N for the 07-Sep-2015 storm is plotted against the observed data for each station. In the right column the density plot for each station is shown for the 2015 dataset. The predictions of the coupled models provide better Pearson correlation coefficients than the original LSTM $|B_N|$ while following the storm time fluctuations, indicating the Polarity model is providing the necessary information to the $|B_N|$ prediction.

increase in Pearson correlation over the MLR models with the LSTM models achieving HSS above 0.5 for 87.5% of the selected $|B_N|$ thresholds for both storm time and 2015 while MLR scored above 0.5 for 18.75% for the same thresholds. When focused on the 2015 HSS of the 99 percentile $|B_N|$ threshold we can see the LSTM models outperform the MLR models in all cases, showing a 152.2, 66.6, 95.8, and 215.8% increase for FYU, KAV, PKR, and CMO stations, respectively. Likewise, when focusing the 99 percentile $|B_N|$ threshold storm HSS we find a 176.9, 16.1, 510, and 828.6% increase for FYU, KAV, PKR, and CMO, respectively. The HSS results for the $d|B_N|/dt$ thresholds show mixed results between the LSTM and MLR models. For the 2015 $d|B_N|/dt$ HSS, LSTM models generally perform better than the MLR by showing higher HSS for 10 out of 16 thresholds, while for the storm time $d|B_N|/dt$ scores the MLR models perform better by showing higher HSS for 10 out of 16 thresholds. However, our storm time performance testing is limited for only a single geomagnetic storm. For better understanding of storm time model performance, a more comprehensive testing is required with a large number of geomagnetic storms.

The neural networks created are adequate at predicting the overall strength of the field, as indicated by high correlation and good HSS scores for $|B_N|$, but not the variability of the field. This can also be seen in **Figures 4G,I,K**, and m where the $d|B_N|/dt$ of the LSTM models can be seen to exceed that of the observed values at times predating or postdating the observed fluctuations. Additionally, the $d|B_N|/dt$ scores for the LSTM models is less than 0.5, regardless of looking at storm time or the full 2015 test year. The poor $d|B_N|/dt$ scores of LSTM models are understandable since the models were made to predict $|B_N|$ and then derive $d|B_N|/dt$ from predicted $|B_N|$. Further, we see a pattern where the scores for $d|B_N|/dt$ are decreasing as the threshold increases, which is a pattern found in other recent machine learning studies of geomagnetic fields (Camporeale et al., 2020; Smith et al., 2021), implying it is a common challenge for data-driven models.

Despite the low performance in $d|B_N|/dt$, the performance in predicting $|B_N|$ is promising, especially when coupled with the secondary polarity model. The polarity models created scored 0.9 or above at predicting the encoded polarity value, which when multiplied through their respective station the LSTM prediction generally retained or increased in HSS for the $|B_N|$ thresholds of the original $|B_N|$ model. The HSS within **Table 1** show an enhancement within the storm time dB_N/dt scores for 14 out of 16 thresholds when using a coupled model approached, while 15 out of 16 2015 threshold scores stay the same or

TABLE 1 | HSS of selected $|B_N|$ and $d|B_N|/dt$ thresholds for the LSTM and MLR models. Scores are evaluated through direct comparison on a minute by minute basis across the year of 2015 and separately for the 07-Sep-2015 storm. Polarity has been encoded into a value of 0 (–) or 1 (+) and the HSS for this corresponds to accurately assessing a 1.

	$d\|B_N\|/dt$ [nT/min]				$\|B_N\|$ [nT]			
Threshold	18	42	66	90	14.4	41.6	75.3	427.0
FYU LSTM	0.35	0.28	0.23	0.18	0.48	0.61	0.63	0.58
FYU MLR	0.25	0.23	0.20	0.19	0.13	0.31	0.47	0.23
FYU LSTM (storm)	0.29	0.24	0.26	0.12	0.53	0.70	0.68	0.72
FYU MLR (storm)	0.31	0.32	0.18	0.10	0.02	0.29	0.63	0.26
KAV LSTM	0.35	0.29	0.25	0.20	0.58	0.68	0.66	0.50
KAV MLR	0.30	0.26	0.26	0.26	0.17	0.35	0.49	0.30
KAV LSTM (storm)	0.27	0.37	0.23	0.20	0.43	0.54	0.54	0.65
KAV MLR (storm)	0.41	0.41	0.37	0.34	0.05	0.39	0.54	0.56
PKR LSTM	0.36	0.26	0.15	0.10	0.39	0.58	0.63	0.47
PKR MLR	0.24	0.22	0.19	0.15	0.17	0.30	0.48	0.24
PKR LSTM (storm)	0.29	0.25	0.27	0.26	0.37	0.76	0.71	0.61
PKR MLR (storm)	0.30	0.39	0.24	0.10	0.06	0.38	0.60	0.10
CMO LSTM	0.41	0.31	0.25	0.20	0.68	0.75	0.73	0.60
CMO MLR	0.33	0.30	0.25	0.23	0.17	0.35	0.51	0.19
CMO LSTM (storm)	0.36	0.29	0.28	0.34	0.67	0.74	0.75	0.65
CMO MLR (storm)	0.36	0.34	0.30	0.34	0.13	0.39	0.60	0.07
Model	**Polarity**							
	FYU		**KAV**		**PKR**		**CMO**	
Polarity	**0.93**		**0.90**		**0.94**		**0.91**	
	dB_N/dt [nT/min]				**B_N [nT]**			
Threshold	18	42	66	90	14.4	41.6	75.3	427.0
Coupled (FYU)	0.35	0.29	0.23	0.18	0.48	0.61	0.63	0.58
Coupled (FYU) (storm)	0.43	0.35	0.36	0.19	0.61	0.78	0.75	0.72
Coupled (KAV)	0.35	0.29	0.25	0.20	0.58	0.67	0.66	0.50
Coupled (KAV) (storm)	0.37	0.41	0.33	0.28	0.58	0.67	0.61	0.66
Coupled (PKR)	0.32	0.23	0.19	0.15	0.39	0.58	0.63	0.47
Coupled (PKR) (storm)	0.38	0.31	0.36	0.25	0.29	0.81	0.77	0.62
Coupled (CMO)	0.44	0.36	0.30	0.25	0.68	0.75	0.73	0.60
Coupled (CMO) (storm)	0.45	0.46	0.41	0.37	0.77	0.81	0.80	0.67

increase. The minimal increase of the coupled models B_N HSS from the LSTM models $|B_N|$ HSS is understandable because B_N is mainly determined by the $|B_N|$ models and not by the polarity model. However, the coupled model provides important sign information of $|B_N|$, creating overall larger dB_N/dt, and thus boosting the original dB_N/dt HSS of the coupled models from the $d|B_N|/dt$ HSS of the LSTM models. Additionally, there are more data points for both B_N and dB_N/dt of the coupled models than the LSTM models, since the dataset for the LSTM models had the data points where the MLR models predicted negative removed. These additional data may play a role in increasing the B_N and dB_N/dt HSS of coupled models.

There are concerns for the Polarity models to generate false peaks within the $d|B_N|/dt$ predictions due to overpredictions in the $|B_N|$ models before and after a sign transition or a sign change with incorrect timing. However, the occurrence of this is rare because the current models underpredict $|B_N|$ leading to overall lower negative to positive and vice versa jumps within the dataset and thus lead to lower dB_N/dt values. A majority of the unexpectedly large $d|B_N|/dt$ peaks are due to the original $|B_N|$ models predicting a large fluctuation occurring at the wrong time, which persist when multiplied through by Polarity. The Polarity model, while scoring well, fails predominately at times where there is a flip from positive to negative and vice versa. The delayed or preemptive timing of polarity switching may in turn cause unexpected $d|B_N|/dt$ patterns, however the vast majority of these patterns will arise during quiet times where the geomagnetic field is fluctuating minutely around 0 nT.

LSTM Caveats

The LSTM models, while promising, have a few different caveats to them in their implementation. The first and foremost caveat is that the models only predict $|B_N|$ one time step ahead, which is limiting in advanced GIC prediction. For future work we aim to provide a 10–20 min prediction range thus increasing the model's practicability. There are two main methods of achieving this, one of which is utilizing the initial t+1 prediction as the starting point of a secondary network that spans the desired prediction range. This setup requires that the secondary network be initialized with the full training dataset to set the internal states of the model. In practice, this is a time consuming approach, which will only increase as more data is used to create the models. This is opposed to our justification for using the 30-min time history model over the 120-min history model, since complexity of the model will play a factor in time to set the states every time a new prediction is made. Another method is to determine the probability of strong $d|B_N|/dt$ fluctuations in a set period. This type of approach has been utilized in other studies, such as the one by Maimaiti et al. (2019) to determine the probability of geomagnetic substorm onset and the studies by Smith et al. (2021) and Camporeale et al. (2020) to forecast the probability of specific dB/dt (i.e., surface geomagnetic field time derivative) thresholds. However, this approach for GIC prediction has ambiguity in whether the outcome will occur at the beginning, middle, or end of the window, which may be pertinent information to the end user.

Secondly, our models currently only predict $|B_N|$ and in combination with a polarity model, B_N, while GICs are influenced by the surface geomagnetic field which is made up of both B_N and B_E (Ngwira et al., 2008; Bedrosian and Love, 2015; Lotz and Cilliers, 2015). For future work we aim for the models to predict two output variables, $|B_N|$ and $|B_E|$. The surface geomagnetic field information, combined with ground conductivities, can be used to determine geomagnetically induced electric fields within the Earth's surface and GICs in specific electrical systems (Ngwira et al., 2008). The current models are limited in this capacity as local conductivity information is currently unavailable for the Alaska region. Until proper conductivity information is available, the models may still be utilized as an indication system, since GICs oftentime occur with large geomagnetic field perturbations that our models are intended to predict.

Thirdly, our models generally underpredict the geomagnetic field strength and do not properly capture the time variations observed in the data. This underprediction is commonly seen in other Machine Learning models (e.g. Keesee et al. (2020)) and likely due to the choice of training with both quiet and storm time data, where the quiet time data makes a larger portion of the trained set leading to lower overall geomagnetic field predictions. A possible solution around this is to train solely off of storm time data, however this adds an extra layer of complexity to the training process during model training and has been shown to not completely solve the problem on its own (Pinto et al., 2022). However, this approach improves general model performance and is something we plan to implement in the future. Another possible approach would be to create a model that takes the incoming prediction and computes the likely offset for that value to better retain the strength of the observed field data, though such a setup would likely require careful consideration in implementation. Even with the aforementioned approaches, the machine-learning based model may find difficulty in predicting "once-in-a-lifetime" singular events, for example, the Carrington event on 1–2 September 1859 (Green and Boardsen, 2006; Cliver and Dietrich, 2013), because such severe events are very rare. In such a case, a physics-based model is a good alternative for the dB/dt predictions.

Lastly, the internal setup of LSTM requires full time history information of the incoming data, which this study does not have access to due to gaps in data sources. During the training of the models, if data is not split into groups of continuous data, the model will not understand the presence of gaps assuming that the incoming data is continuous. In a real-time situation data outages will occur and the model will continue with these data gaps present in the dataset, unless reset to avoid, which would result in 30 min of time without predictions at a minimum while waiting for a continuous stream of data. With this in mind, we chose to train the model with data gaps included within the dataset, which can be seen in **Figures 2**, **4** where gaps in data are present, allowing the model to pick up immediately when new information is available rather than requiring a down time to fill a continuous segment. This is done after the linear interpolation by passing the model the dataset as-is and allowing the training algorithms to assume the set is continuous. However,

we understand this decision results in performance decreases in datasets with many or expansive gaps in data, such as the PKR dataset which was missing the year of 2010 within the supplied dataset. Despite the known limitations of training from incomplete datasets, we find the models still attain high Pearson correlation coefficients and promising HSS for $|B_N|$ and B_N prediction, and will benefit from on-going efforts for continuous dataset collection.

While the caveats mentioned above are inherent to the utilization of LSTM and our current efforts in producing a model, there is one limitation within our model that puts a restraint on where these models may be used. The models produced so far rely on past known geomagnetic field information (i.e., $\mathrm{d}|B_N|/\mathrm{dt}$ in the previous 30 min) in the local region, which may not be suitable for all areas looking to perform GIC risk assessment. The Alaska region can accommodate this since the magnetometers chosen also provide real-time data which may be used with the models for real-time predictions. Additionally, the Space Weather UnderGround outreach project initiated at the University of New Hampshire (Smith, 2020) and expanded to the University of Alaska Fairbanks, will build a cost efficient and research-capable array of magnetometers across Alaska and New Hampshire with a 1 nT/s resolution, increasing the spatial resolution of data in these regions. Moving forward our models will utilize the datasets provided by these arrays for forecasting GIC risk.

Potential LSTM Model Use and Future Work

With the inclusion of multi-minute output, the LSTM models will be matured enough to create advanced GIC warnings based on likely $\mathrm{d}B_N/\mathrm{dt}$ thresholds without the need to directly predict GICs. The creation of GICs is a complex problem that is dependent on ground conductivity and the properties of the electrical device that it is being influenced. The land conductivities of Alaska have not been thoroughly studied though conductivity maps exist for the mainland of the United States. Due to the complexity of GIC prediction on every electrical system, it becomes practical to provide GIC warning and geomagnetic field predictions to the end user. In this manner, the end user can apply these predictions coupled with the knowledge of their own system to determine risk.

The increase of model performance via multi time history predictions and closer values to the observed data will allow the models to be utilized in ionospheric current predictions. This possible use case was the reason for pursuing B_N predictions instead of $|B_N|$ leading to the creation of the polarity model. Previous studies have created modeling techniques to predict the ionospheric current based on local and/or global geomagnetic field patterns and electrodynamics (Lu et al., 1995; Kihn and Ridley, 2005; Vanhamäki and Juusola, 2020). A local model to determine the auroral oval requires, at a minimum, multi-point field values along a line perpendicular to the oval. Our current study is setup for this with the use of PKR, FYU, KAV, and CMO, which roughly lay on a line perpendicular to the auroral oval. With the inclusion of multi time history predictions our LSTM models coupled with ionospheric current modeling have the potential to forecast north-south motion or expansion of the ionospheric currents. Current patterns would be useful in region-based GIC risk assessment and awareness while also providing information to aurora enthusiasts and citizen scientists since strong ionospheric currents have been connected to auroral activity (Akasofu, 1989; Newell et al., 2001).

SUMMARY AND FUTURE WORK

This study aims to show the progression of LSTM neural networks trained to predict the geomagnetic field at individual stations across Alaska. To achieve this we trained 12 models (4 LSTM, 4 MLR, 4 Polarity) with NASA OMNI IMF and solar wind data coupled with SuperMAG geomagnetic field information from the years 2000–2015 split in a 68.75-25-6.25 training/validation/test set configuration. We produced 8 models to test the configuration of LSTM with our desired inputs and outputs. From these models we chose 30 min of time history utilizing IMF, solar wind, and past $\mathrm{d}|B_N|/\mathrm{dt}$ information as the most effective and applied the configuration to 4 stations across Alaska. We find that the LSTM models generally outperform the MLR models with respect to predicting $|B_N|$, however the results of $\mathrm{d}|B_N|/\mathrm{dt}$ prediction performance are inconsistent between the two modeling methods. Due to the initial model performing best when predicting $|B_N|$ a coupled model approach was utilized to retain B_N output with performance similar to the original $|B_N|$ model it was based on.

The models are limited to single point future predictions and generally underestimate the strength of the geomagnetic field. Future work on these models aims to increase the time history output to 10–20 min of future predictions while also increasing the performance of the models to estimate the geomagnetic field strength and variations in both the North-South and East-West components. With this will come potential integration into ionospheric current models for ionospheric current and auroral activity forecasting. Further, the models will be converted to a regional GIC risk assessment allowing the end user to apply the geomagnetic field predictions to their own electrical systems.

DATA AVAILABILITY STATEMENT

The solar wind and IMF data are available from OMNIWeb at https://omniweb.gsfc.nasa.gov and the ground magnetometer data are available from SuperMAG at http://supermag.jhuapl.edu.

AUTHOR CONTRIBUTIONS

MB was the primary author of the paper, conducted data preparation, model development, and participated in analysis. HC provided guidance on study concept and design, model

development and analysis, interpretation of results, and writing. DO contributed to guidance on model development, data preparation, and general guidance. AK contributed to the study design and overall discussion. VP contributed to model development and methodology, overall discussion, general guidance, and writing. MM contributed to model development and methodology, overall discussion, and general guidance. CN contributed to general guidance and discussions regarding GICs. SP contributed to model methodology, data preparation, and overall discussion.

FUNDING

This work was supported by NSF EPSCoR Award OIA-1920965. HC gracefully acknowledges support from the NASA grants, 80NSSC18K1043, 80NSSC20K1670, and 80MSFC20C0019.

ACKNOWLEDGMENTS

We thank all members of the MAGICIAN team at UNH and UAF that participated in discussions leading to this article. We also thank USGS, INTERMAGNET, and the Geophysical Institute for supporting and maintaining the magnetometers used in this study as well as promoting high standards of magnetic observatory practice. We thank NASA OMNIweb and SuperMAG for providing organized data access that supported this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fspas.2022.846291/full#supplementary-material

REFERENCES

Akasofu, S.-I. (1989). The Dynamic Aurora. *Sci. Am.* 260, 90–97. doi:10.1038/scientificamerican0589-90

Bedrosian, P. A., and Love, J. J. (2015). Mapping Geoelectric Fields during Magnetic Storms: Synthetic Analysis of Empirical united states Impedances. *Geophys. Res. Lett.* 42, 10,160–10,170. doi:10.1002/2015gl066636

Blake, S. P., Gallagher, P. T., McCauley, J., Jones, A. G., Hogg, C., Campanyà, J., et al. (2016). Geomagnetically Induced Currents in the Irish Power Network during Geomagnetic Storms. *Space Weather* 14, 1136–1154. doi:10.1002/2016sw001534

Brownlee, J. (2017). *Long Short-Term Memory Networks with Python: Develop Sequence Prediction Models with Deep Learning*. Jason Brownlee. Available at: https://machinelearningmastery.com/lstms-with-python/.

Camporeale, E., Cash, M. D., Singer, H. J., Balch, C. C., Huang, Z., and Toth, G. (2020). A gray-Box Model for a Probabilistic Estimate of Regional Ground Magnetic Perturbations: Enhancing the Noaa Operational Geospace Model with Machine Learning. *J. Geophys. Res. Space Phys.* 125, e2019JA027684. doi:10.1029/2019JA027684

Carter, B. A., Yizengaw, E., Pradipta, R., Halford, A. J., Norman, R., and Zhang, K. (2015). Interplanetary Shocks and the Resulting Geomagnetically Induced Currents at the Equator. *Geophys. Res. Lett.* 42, 6554–6559. doi:10.1002/2015GL065060

Cliver, E. W., and Dietrich, W. F. (2013). The 1859 Space Weather Event Revisited: Limits of Extreme Activity. *J. Space Weather Space Clim.* 3, A31. doi:10.1051/swsc/2013053

Connor, H. K., Zesta, E., Ober, D. M., and Raeder, J. (2014). The Relation between Transpolar Potential and Reconnection Rates during Sudden Enhancement of Solar Wind Dynamic Pressure: Openggcm-Ctim Results. *J. Geophys. Res. Space Phys.* 119, 3411–3429. doi:10.1002/2013JA019728

Dimmock, A. P., Rosenqvist, L., Hall, J. O., Viljanen, A., Yordanova, E., Honkonen, I., et al. (2019). The GIC and Geomagnetic Response Over Fennoscandia to the 7-8 September 2017 Geomagnetic Storm. *Space Weather* 17, 989–1010. doi:10.1029/2018SW002132

Dimmock, A. P., Rosenqvist, L., Welling, D. T., Viljanen, A., Honkonen, I., Boynton, R. J., et al. (2020). On the Regional Variability of Db/dt and its Significance to Gic. *Space Weather* 18, e2020SW002497. doi:10.1029/2020SW002497

Fiori, R. A. D., Boteler, D. H., and Gillies, D. M. (2014). Assessment of Gic Risk Due to Geomagnetic Sudden Commencements and Identification of the Current Systems Responsible. *Space Weather* 12, 76–91. doi:10.1002/2013sw000967

Gjerloev, J. W. (2012). The SuperMAG Data Processing Technique. *J. Geophys. Res.* 117, A09213. doi:10.1029/2012JA017683

Green, J. L., and Boardsen, S. (2006). Duration and Extent of the Great Auroral Storm of 1859. *Adv. Space Res.* 38, 130–135. doi:10.1016/j.asr.2005.08.054

Gummow, R. A., and Eng, P. (2002). Gic Effects on Pipeline Corrosion and Corrosion Control Systems. *J. Atmos. Solar-Terrestrial Phys.* 64, 1755–1764. doi:10.1016/S1364-6826(02)00125-6

Heyns, M. J., Lotz, S. I., and Gaunt, C. T. (2021). Geomagnetic Pulsations Driving Geomagnetically Induced Currents. *Space Weather* 19, e2020SW002557. doi:10.1029/2020SW002557

Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.* 9, 1735–1780. doi:10.1162/neco.1997.9.8.1735

Keesee, A. M., Pinto, V., Coughlan, M., Lennox, C., Mahmud, M. S., and Connor, H. K. (2020). Comparison of Deep Learning Techniques to Model Connections between Solar Wind and Ground Magnetic Perturbations. *Front. Astron. Space Sci.* 7, 72. doi:10.3389/fspas.2020.550874

Khanal, K., Adhikari, B., Chapagain, N. P., and Bhattarai, B. (2019). HILDCAA-Related GIC and Possible Corrosion Hazard in Underground Pipelines: A Comparison Based on Wavelet Transform. *Space Weather* 17, 238–251. doi:10.1029/2018sw001879

Kihn, E. A., and Ridley, A. (2005). A Statistical Analysis of the Assimilative Mapping of Ionospheric Electrodynamics Auroral Specification. *J. Geophys. Res.* 110 (A7). doi:10.1029/2003JA010371

Liu, L., Yu, Z., Wang, X., and Liu, W. (2019). The Effect of Tidal Geoelectric fields on Gic and Psp in Buried Pipelines. *IEEE Access* 7, 87469–87478. doi:10.1109/ACCESS.2019.2920944

Lotz, S. I., and Cilliers, P. J. (2015). A Solar Wind-Based Model of Geomagnetic Field Fluctuations at a Mid-latitude Station. *Adv. Space Res.* 55, 220–230. doi:10.1016/j.asr.2014.09.014

Lu, G., Lyons, L. R., Reiff, P. H., Denig, W. F., de la Beaujardiére, O., Kroehl, H. W., et al. (1995). Characteristics of Ionospheric Convection and Field-Aligned Current in the Dayside Cusp Region. *J. Geophys. Res.* 100, 11845–11861. doi:10.1029/94JA02665

Maggiolo, R., Hamrin, M., De Keyser, J., Pitkänen, T., Cessateur, G., Gunell, H., et al. (2017). The Delayed Time Response of Geomagnetic Activity to the Solar Wind. *J. Geophys. Res. Space Phys.* 122, 11,109–11,127. doi:10.1002/2016JA023793

Maimaiti, M., Kunduri, B., Ruohoniemi, J. M., Baker, J. B. H., and House, L. L. (2019). A Deep Learning-Based Approach to Forecast the Onset of Magnetic Substorms. *Space Weather* 17, 1534–1552. doi:10.1029/2019SW002251

Newell, P. T., Greenwald, R. A., and Ruohoniemi, J. M. (2001). The Role of the Ionosphere in aurora and Space Weather. *Rev. Geophys.* 39, 137–149. doi:10.1029/1999RG000077

Ngwira, C. M., Pulkkinen, A. A., Bernabeu, E., Eichner, J., Viljanen, A., and Crowley, G. (2015). Characteristics of Extreme Geoelectric fields and Their Possible Causes: Localized Peak Enhancements. *Geophys. Res. Lett.* 42, 6916–6921. doi:10.1002/2015GL065061

Ngwira, C. M., Pulkkinen, A., McKinnell, L.-A., and Cilliers, P. J. (2008). Improved Modeling of Geomagnetically Induced Currents in the South African Power Network. *Space Weather* 6, S11004. doi:10.1029/2008SW000408

Oliveira, D. M., Arel, D., Raeder, J., Zesta, E., Ngwira, C. M., Carter, B. A., et al. (2018). Geomagnetically Induced Currents Caused by Interplanetary Shocks with Different Impact Angles and Speeds. *Space Weather* 16, 636–647. doi:10.1029/2018SW001880

Oliveira, D. M., and Ngwira, C. M. (2017). Geomagnetically Induced Currents: Principles. *Braz. J. Phys.* 47, 552–560. doi:10.1007/s13538-017-0523-y

Pinto, V. A., Keesee, A. M., Coughlan, M., Mukundan, R., Johnson, J. W., Ngwira, C. M., et al. (2022). *Revisiting the Ground Magnetic Firld Perturbations challenge: A Machine Learning Perspective*. Unpublished.

Pirjol, R. J., Viljanen, A. T., and Pulkkineni, A. A. (2007). "Research of Geomagnetically Induced Currents (GIC) in Finland," in 2007 7th International Symposium on Electromagnetic Compatibility and Electromagnetic Ecology, St. Petersburg, Russia, 26-29 June 2007. doi:10.1109/EMCECO.2007.4371707

Pirjola, R. (2005). Effects of Space Weather on High-Latitude Ground Systems. *Adv. Space Res.* 36, 2231–2240. doi:10.1016/j.asr.2003.04.074

Pirjola, R. (2000). Geomagnetically Induced Currents during Magnetic Storms. *IEEE Trans. Plasma Sci.* 28, 1867–1873. doi:10.1109/27.902215

Pirjola, R., Pulkkinen, A., and Viljanen, A. (2003). Studies of Space Weather Effects on the Finnish Natural Gas Pipeline and on the Finnish High-Voltage Power System. *Adv. Space Res.* 31, 795–805. doi:10.1016/s0273-1177(02)00781-0

Pulkkinen, A., Lindahl, S., Viljanen, A., and Pirjola, R. (2005). Geomagnetic Storm of 29-31 October 2003: Geomagnetically Induced Currents and Their Relation to Problems in the Swedish High-Voltage Power Transmission System. *Space Weather* 3 (8). doi:10.1029/2004SW000123

Pulkkinen, A., Rastätter, L., Kuznetsova, M., Singer, H., Balch, C., Weimer, D., et al. (2013). Community-Wide Validation of Geospace Model Ground Magnetic Field Perturbation Predictions to Support Model Transition to Operations. *Space Weather* 11, 369–385. doi:10.1002/swe.20056

Rodger, C. J., Mac Manus, D. H., Dalzell, M., Thomson, A. W. P., Clarke, E., Petersen, T., et al. (2017). Long-Term Geomagnetically Induced Current Observations from New Zealand: Peak Current Estimates for Extreme Geomagnetic Storms. *Space Weather* 15, 1447–1460. doi:10.1002/2017SW001691

Rogers, N. C., Wild, J. A., Eastoe, E. F., Gjerloev, J. W., and Thomson, A. W. P. (2020). A Global Climatological Model of Extreme Geomagnetic Field Fluctuations. *J. Space Weather Space Clim.* 10, 5. doi:10.1051/swsc/2020008

Smith, A. W., Forsyth, C., Rae, I. J., Garton, T. M., Bloch, T., Jackman, C. M., et al. (2021). Forecasting the Probability of Large Rates of Change of the Geomagnetic Field in the uk: Timescales, Horizons, and Thresholds. *Space Weather* 19, e2021SW002788. doi:10.1029/2021SW002788

Smith, C. (2020). Space Weather Underground: A Magnetometer Array with Educational Opportunities. *Scientia Glob.* doi:10.33548/SCIENTIA542

Vanhamäki, H., and Juusola, L. (2020). *Introduction to Spherical Elementary Current Systems*. Cham: Springer International Publishing, 5–33. doi:10.1007/978-3-030-26732-2_2

Verbeek, M. (2017). Using Linear Regression to Establish Empirical Relationships. *IZA World of Labor 2017* 336. doi:10.15185/izawol.336

Welling, D. (2019). *Magnetohydrodynamic Models of B and Their Use in GIC Estimates*. American Geophysical Union AGU, 43–65. chap. 3. doi:10.1002/9781119434412.ch3

Wintoft, P., Wik, M., and Viljanen, A. (2015). Solar Wind Driven Empirical Forecast Models of the Time Derivative of the Ground Magnetic Field. *J. Space Weather Space Clim.* 5, A7. doi:10.1051/swsc/2015008

Zhang, J. J., Wang, C., Sun, T. R., and Liu, Y. D. (2016). Risk Assessment of the Extreme Interplanetary Shock of 23 July 2012 on Low-Latitude Power Networks. *Space Weather* 14, 259–270. doi:10.1002/2015SW001347

Conflict of Interest: CN was employed by the ASTRA LLC.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Astronomy and Space Sciences

ORIGINAL RESEARCH
published: 09 May 2022
doi: 10.3389/fspas.2022.878403

A New Three-Dimensional Empirical Reconstruction Model Using a Stochastic Optimization Method

Xun Zhu[1]*, *Ian J. Cohen*[1], *Barry H. Mauk*[1], *Romina Nikoukar*[1], *Drew L. Turner*[1] and *Roy B. Torbert*[2,3]

[1]The Johns Hopkins University Applied Physics Laboratory, Laurel, MD, United States, [2]Physics Department, University of New Hampshire, Durham, NH, United States, [3]Southwest Research Institute, San Antonio, TX, United States

Motivated by MMS mission observations near magnetic reconnection sites, we have developed a new empirical reconstruction (ER) model of the three-dimensional (3D) magnetic field and the associated plasma currents. Our approach combines both the measurements from a constellation of satellites and a set of physics-based equations as physical constraints to build spatially smooth distributions. This ER model directly minimizes the loss function that characterizes the model-measurement differences and the model departures from linear or nonlinear physical constraints using an efficient stochastic optimization method by which the effects of random measurement errors can be effectively included. Depending on the availability of the measured parameters and the adopted physical constraints on the reconstructed fields, the ER model could be either slightly over-determined or under-determined, yielding nearly identical reconstructed fields when solved by the stochastic optimization method. As a result, the ER model remains valid and operational even if the input measurements are incomplete. Two sets of new indices associated respectively with the model-measurement differences and the model departures are introduced to objectively measure the accuracy and quality of the reconstructed fields. While applying the reconstruction model to observations of an electron diffusion region (EDR) observed by NASA's Magnetospheric Multiscale (MMS) mission, we examine the relative contributions of the errors in the plasma current density arising from random measurement errors and linear approximations made in application of the curlometer technique. It was found that the errors in the plasma current density calculated directly from the measured magnetic fields using a linear approximation were mostly contributed from the nonlinear configuration of the 3D magnetic fields.

Keywords: stochastic optimization, empirical reconstruction model, magnetospheric reconnection, simultaneous perturbation stochastic approximation, loss function

OPEN ACCESS

Edited by:
Olga Verkhoglyadova, NASA Jet Propulsion Laboratory (JPL), United States

Reviewed by:
Bogdan Hnat, University of Warwick, United Kingdom
Yasuhito Narita, Austrian Academy of Sciences (OeAW), Austria

***Correspondence:**
Xun Zhu
xun.zhu@jhuapl.edu

Specialty section:
This article was submitted to Space Physics, a section of the journal Frontiers in Astronomy and Space Sciences

Received: *18 February 2022*
Accepted: *07 April 2022*
Published: *09 May 2022*

Citation:
Zhu X, Cohen IJ, Mauk BH, Nikoukar R, Turner DL and Torbert RB (2022) A New Three-Dimensional Empirical Reconstruction Model Using a Stochastic Optimization Method. Front. Astron. Space Sci. 9:878403. doi: 10.3389/fspas.2022.878403

INTRODUCTION

Visualization of Earth's magnetosphere is an effective way to understand the magnetospheric environment and its associated physical processes. However, historically our exploration and understanding have been limited to either remote sensing (energetic neutral atom imaging, e.g., IMAGE, TWINS) or *in-situ* point-wise measurements made from satellites in space, from either single- (e.g., Geotail, Polar) or multi-satellite (e.g., THEMIS, Cluster, MMS) missions. One technique

to translate discrete point-wise satellite measurements into a 3D visualization is to develop a reconstruction model that captures the fundamental magnetic field (**B**) and plasma field as characterized by the plasma current density (**J**)—measured independently from the magnetic field—in the neighborhood of the measurement domain. Introduction of magnetohydrodynamic (MHD) equations could also lead to the reconstruction of additional field variables such as plasma velocity (**U**) and electric field (**E**). It is understood that MHD is not appropriate just at the localized site of the electron diffusion region (EDR) where the *X*-point becomes a singularity in an ideal MHD model and the diffusion is parameterized by a bulk parameter of resistivity in a resistive MHD model (Priest, 2016). Our goal is to visualize the broader regions surrounding the EDR site. Depending on specific science problems, the magnetic and plasma fields can be reconstructed either from a set of global measurements to yield a climatological configuration covering the entire magnetosphere (e.g., Tsyganenko and Sitnov, 2007) or from a set of *in-situ* measurements along satellite paths to yield a localized configuration in both space and time (e.g., Dunlop et al., 1988; Dunlop et al., 2002). This paper focuses on the localized reconstruction.

Previously, there have been two categories of models for reconstructing localized fields (e.g., **B** and **J**) in Earth's magnetosphere. The first uses the Grad-Shafranov reconstruction (GSR) technique to produce reconstruction field maps of (**B**, **J**) and **U** by solving a set of MHD equations where the measurements are used as boundary conditions to constrain the reconstructed field (e.g., Sonnerup and Guo, 1996; Hasegawa et al., 2004; Hasegawa et al., 2005; Sonnerup and Teh, 2008; Zhu and Lui, 2012; Sonnerup et al., 2016). The GSR technique was developed for a force-free magnetic-field configuration (e.g., Sturrock, 1994) and was mainly used to derive two-dimensional stationary and coherent MHD structure in the magnetosphere (e.g., Sonnerup and Guo, 1996). In this category of approaches, the spatial configuration of the reconstructed fields is determined by solving a full set of self-consistent MHD partial differential equations that extensively describe various physical processes relating different parameters. This reconstruction approach can effectively yield and solve a full set of physics-based model for (**B**, **J**) and **U** using measurements obtained by a single satellite along its trajectory as the boundary conditions.

The second category of reconstruction approaches reconstructs the field maps of (**B**, **J**) by empirically fitting a prescribed spatial configuration of the field maps to the point-wise *in-situ* satellite measurements forming a finite volume with multiple lines and faces in space (e.g., Dunlop et al., 1988; Dunlop et al., 2002; Torbert et al., 2020). We may call this category of techniques an "empirical reconstruction" (ER). This ER approach is especially effective and useful for reconstructing (**B**, **J**) fields from multi-satellite measurements. Unlike the GSR techniques where the spatial configuration of the fields (**B**, **J**) and **U** is solved from the measurements based on a full set of MHD equations, the ER models prescribe the spatial configurations of (**B**, **J**) guided by *in-situ* measurements and use only limited number of physical equations as constraints, such as

$$\mu_0 \mathbf{J} = \nabla \times \mathbf{B} \text{ and} \tag{1a}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{1b}$$

to determine the model parameters. In **Eq. 1a**, μ_0 is the permeability of free space. Note that the above two equations do not form a closed set of equations for a system. There are six individual dependent variables for four component equations. As a result, ER models heavily rely on the measurements to construct smooth fields.

Assuming a linear approximation for the spatial variation of the modeled **B**, Dunlop et al. (1988) introduced a curlometer technique to reconstruct the **J** field solely from the measured **B** based on one MHD equation (**Eq. 1a**). The authors also proposed an objective index called the "quality indicator" to measure the accuracy or quality of the reconstructed **J** field. In Torbert et al. (2020), an ER model for both **B** and **J** fields produced by assuming a nonlinear function for **B** was developed based on point-wise measurements of (**B**, **J**) from MMS and physical constraints from **Eqs. 1a, b**. For a reconstruction model with a nonlinear variation in **B**, we expect the reconstructed **B** and **J** fields to be more accurate and of higher quality than those derived from the curlometer technique, which is founded upon a linear approximation for the **B** field. Such an improvement is especially important near EDRs where the magnetic field lines are expected to be highly curved and the plasma field plays an important role in the localized reconnection process. Note that the ER model by Torbert et al. (2020) was developed as an evenly determined problem, i.e., the numbers of unknown parameters and constraints are equal, from the perspective of the more general data analysis technique for which an extra constraint is needed to add to the model that will affect the quality of the reconstructed fields. In addition, the quality and the factors affecting the reconstruction quality are difficult to quantify.

In this paper, we develop a new 3D ER model by using a stochastic optimization method to construct the smooth fields. This new ER model is a generalization of the previous ER models for which additional measurements and MHD equations can be flexibly introduced in the same model framework. In addition, the model effectively considers and quantifies the effects of random errors arising from uncertainties in the *in-situ* measurements. Furthermore, this stochastic optimization approach introduces additional flexibility into the model by allowing it to work regardless of whether the parameters considered are over- or under-defined. The central idea of the previous ER models is the utilization of the MHD **Eq. 1a** that derives **J** field from a prescribed analytic **B** field to fit the point-wise measurements and to perform the reconstructions. Note that **Eq. 1a** is derived by neglecting the displacement current in Ampere's Law and is one of several important equations in an MHD system. The validity of **Eq. 1a** is based on the MHD fundamental assumption that the fields vary on the same time and length scales as the plasma parameters (Boyd and Sanderson, 2003). Two other important MHD equations similar to **Eq. 1a** are Ohm's Law

$$\mathbf{E} = (\eta/\mu_0)(\nabla\times \mathbf{B}) - \mathbf{U}\times\mathbf{B} \quad (2)$$

which derives the electric field **E** from the plasma velocity **U** for a given **B** field, and Faraday's Law of Induction, which relates the plasma resistivity (η) to the rest of the fields (Boyd and Sanderson, 2003)

$$\frac{\partial \mathbf{B}}{\partial t} = -\mathbf{B}(\nabla\cdot\mathbf{U}) + (\mathbf{B}\cdot\nabla)\mathbf{U} - (\mathbf{U}\cdot\nabla)\mathbf{B} + \frac{\eta}{\mu_0}\nabla^2\mathbf{B} + \nabla\eta\times\frac{(\nabla\times\mathbf{B})}{\mu_0}. \quad (3)$$

Here, **Eq. 2** plays a role similar to **Eq. 1a** in that an analytic **E** field can be derived from a prescribed analytic **U** field for given (**B**, η). Note that the plasma resistivity η can be considered a parametric measure of the particle acceleration and energy conversion near EDRs. Alternatively, it can also be considered as a phenomenological parameter to be used as a proxy to locate the EDRs of reconnection (e.g., Scudder, 2016; Yamada et al., 2016). In particular, the ultimate inclusion of this aspect of particle acceleration—and thus connection to recent MMS energetic particle observations near EDRs (e.g., Cohen et al., 2021; Turner et al., 2021)—motivated development of this new reconstruction approach.

For an ER model that only adopts one or two MHD linear equations, the problem can be solved by a traditional least-squares method that solves a set of linear algebraic equations (e.g., Dunlop et al., 1988; Dunlop et al., 2002; Denton et al., 2020; Torbert et al., 2020). When additional and, more importantly, nonlinear MHD equations such as **Eqs. 2**, **3** are included, a more practical approach is to solve the model parameters by directly minimizing a "loss function" that characterizes the model-measurement differences and the model departures from the above MHD equations (**Eqs. 1**–**3**). The new ER model introduced here solves for the reconstruction parameters by directly minimizing this loss function, which will be discussed in detail in **Section 2**. Note that the term "model departure" here means violation of a physical constraint—e.g., a violation of **Eq. 1b** in the reconstructed model. Such a violation arises either from the measurement errors on which the ER model is built or from the nature of the ER with a prescribed configuration—e.g., a linear or quadratic functional form in **B** field.

The new 3D ER model presented here has been built on the basis of directly minimizing the loss function L (or χ) using a stochastic optimization method. For a linear system such as one using only **Eqs. 1a**, b, the model parameters could also be solved by the traditional least-squares method if the reconstruction is formulated in an even-determined or an over-determined problem. Comparing to the traditional least-squares method that solves a set of linear algebraic equations, this alternative method has several merits. First, the system could be nonlinear or the loss function L is not necessarily in a quadratic form with respect to the model parameters. The nonlinearity becomes unavoidable when the plasma resistivity is included in an ER model that uses MHD **Eqs. 1**–**3**. The loss function L, to be discussed in detail in **Section 2** for the present ideal MHD ER model, has a quadratic form for which the model parameters could also be derived by solving a set of linear algebraic equations when an additional constraint is used to formulate the problem into an even-determined one (e.g., Denton et al., 2020; Torbert et al., 2020). However, our detailed discussions on how to specify and select different components of L clearly also show the flexibility of the new model that allows other constraints corresponding to the point-wise measurements of (**U**, **E**) fields and **Eqs. 2**, **3** to be added to the reconstruction without much change in the algorithmic structure. Second, the effect of the measurement errors is explicitly included in the reconstruction model (see **Section 3.1**). While by nature all parameters of stochastic algorithms are random variables, there are two sources of uncertainties in practice for a physical problem: 1) the measurements carry random errors and 2) physical relations used in the loss function constraints are not perfect. Both uncertainty sources are included in the stochastic optimization method, which gives a solution with its accuracy limited by the error term ε_σ in **Eq. 8b**. Of course, algorithmically, one may choose a very small error term or set $\sigma \to 0$ in **Eq. 8b**—i.e., assuming perfect measurements and physical constraints - to recover a quasi-mathematically deterministic solution (e.g., Zhu and Spall, 2002). Finally, we adopted a simultaneous perturbation stochastic approximation (SPSA) algorithm to solve the stochastic optimization problem that makes directly minimizing the loss function efficient or practically feasible when the number of the model parameters gets large. The ability of SPSA algorithms to efficiently evaluate the loss function gradient at each iteration makes stochastic optimization a powerful tool for various applications models and simulations (e.g., Spall, 2003; Bhatnagar et al., 2013).

In **Section 2**, we describe how to build an ER model that includes two critical steps: 1) design of a loss function and 2) use of an efficient optimization technique to solve for the model parameters. **Section 3** defines several indices that measure the accuracy and quality of the reconstruction model and presents the model results for a test case near a previously-studied EDR event (Torbert et al., 2018; Torbert et al., 2020) observed in the magnetotail by the Magnetospheric Multiscale (MMS) mission (Burch et al., 2016). **Section 4** provides a few concluding remarks.

MODEL DESCRIPTION

The first step to build an ER model is to design a "loss function" based on the available measurements and a set of adopted MHD equations such as those shown in **Eqs. 1**–**3**. In general, an analytic and smooth specification of the field variables (**B**, **U**, η) will automatically lead to analytic and smooth functions for (**J**, **E**) fields by use of **Eqs. 1a**, **2**. This procedure allows analytic evaluations of all modeled fields at any space-time grids to be compared with the available measurements. The loss function is defined as a collection of various constraints corresponding to the model-measurement differences and the model departures from the adopted MHD equations, such as **Eqs. 1**–**3**. In practice, other complementary physical equations may serve as additional constrains. For example, just as to **Eq. 1b** that imposes a strong constraint on the reconstructed **B** field, the plasma velocity **U** may satisfy an approximate continuity equation $\nabla\cdot$

$\mathbf{U} = 0$ (Priest, 2016), which can serve as an additional constraint on the **U** field in addition to **Eqs. 2**, **3** and the point-wise **U** measurements.

The second step to build an ER model is to solve for the model parameters by minimizing the defined loss function. When the MHD equations are linear, such as those shown in **Eqs. 1a**, b, the model parameters can be derived by a traditional least-squares method that solves a set of linear algebraic equations. Alternatively, the model parameters can also be solved by directly minimizing the loss function. This approach is especially useful when the adopted MHD equations contain nonlinear components, which generally cannot be converted to a set of linear algebraic equations. In this paper, we use a stochastic optimization method called the "simultaneous perturbation stochastic approximation" (SPSA) method to directly minimize the loss function regardless of whether or not the system contains nonlinear terms (e.g., Spall, 1998a; Zhu and Spall, 2002; Spall, 2003). In addition, random errors are treated directly in the loss function and the SPSA solution procedure so that the effects of measurement uncertainties can be examined. Once the model parameters are obtained, the last step to build an ER model is to diagnose the accuracy and the quality of the reconstructed fields. Such a post-diagnostic procedure is necessary because the ER models are built on both measurements that contain random measurement errors and adopted MHD equations that do not form a closed system.

Design of the Loss Function for the Reconstruction Model for an Ideal Magnetohydrodynamic System

To demonstrate how the aforementioned three steps are implemented, we first apply this new 3D ER model to an MHD system that only contains point-wise measurements of (**B**, **J**) fields together with MHD **Eqs 1a**, b as has been extensively investigated by the traditional least-squares method (e.g., Denton et al., 2020; Torbert et al., 2020). This reconstruction model can be considered an ER model for an ideal MHD system because the effect of resistivity (η) is not included. Extension to a more comprehensive nonlinear ER model that uses **Eqs. 1**–**3** with point-wise measurements of (**B**, **J**) and (**U**, **E**) fields and incorporates the effects of plasma resistivity contained in **Eqs. 2**, **3** near the EDRs will be presented in our future investigations.

Here, we follow Torbert et al. (2020) and prescribe the form of the reconstructed field by expressing the time-independent magnetic field **B** as a quadratic function of the spatial coordinate **r** by a second-order Taylor expansion of a vector field

$$\mathbf{B}(\mathbf{r}) \approx \mathbf{B}(\mathbf{r}_0) + [D_{\mathbf{r}}\mathbf{B}(\mathbf{r}_0)](\mathbf{r}-\mathbf{r}_0) + \frac{1}{2}(\mathbf{r}-\mathbf{r}_0)^T\left[D^2_{\mathbf{r}}\mathbf{B}(\mathbf{r}_0)\right](\mathbf{r}-\mathbf{r}_0), \quad (4)$$

where $\mathbf{r}_0 = (1/4)\sum_{\alpha=1:\,4}\mathbf{r}_\alpha$ is the barycenter of the tetrahedron defined at its four vertices by the locations of the four MMS spacecraft ($\mathbf{r}_\alpha(\alpha = 1, 2, 3, 4)$), with $D_{\mathbf{r}}\mathbf{B}(\mathbf{r}_0)$ and $D^2_{\mathbf{r}}\mathbf{B}(\mathbf{r}_0)$ being the first- and second-order derivatives of **B** at $\mathbf{r}_0$, respectively. The new ER model presented here is independent of the coordinate system though we have chosen to employ Geocentric Solar Ecliptic (GSE) coordinates. The terminology, notations and various manipulations of the tetrahedron geometry formed by a four-point satellite configuration have been discussed previously (e.g., Chanteur, 1998; Harvey, 1998; Robert et al., 1998; Dunlop et al., 2002). In addition to the barycenter, we may also define four face-centers ($\mathbf{r}_{F\alpha} = (1/3)\sum_{\beta\neq\alpha}\mathbf{r}_\beta$) and six edge-centers ($\mathbf{r}_{\alpha\beta} = (\mathbf{r}_\alpha + \mathbf{r}_\beta)/2$) of the tetrahedron that can be easily calculated from the coordinates of the vertices. In practice, the coefficients of the derivatives in **Eq. 4** will be determined by the reconstruction model based on the measurements. Hence, we may define the reconstruction model by rewriting **Eq. 4** into the following explicit form for the ith component of the magnetic field

$$B_i(\mathbf{r}) = B_{0i} + \sum_{j=1}^{3} C_{0i,j}\Delta x_j + \frac{1}{2}\sum_{j,k=1}^{3} D_{0i,jk}\Delta x_j \Delta x_k,\ i = 1, 2, 3, \quad (5)$$

where $\mathbf{r} = (x_1, x_2, x_3)$ and $\Delta x_j = x_j - x_{0,j}$. The resulting smooth 3D magnetic field will be determined by thirty model parameters $\{B_{0i}, C_{0i,j}, D_{0i,jk}\}$ constrained by the MMS measurements. Given these model parameters, the spatial derivatives of the **B** field and the associated divergence ($\nabla\cdot\mathbf{B}$) and vorticity ($\nabla\times\mathbf{B}$) fields can be evaluated analytically and thus their valuations are available at any spatial point, though the measurements ($\hat{B}, \hat{J}$) are only available at the four vertices. Note that, physically, $\delta(\mathbf{r}) \equiv \nabla\cdot\mathbf{B}(\mathbf{r}) = \sum_{i=1}^{3}\partial B_i(\mathbf{r})/\partial x_i \equiv 0$ for any value of **r**. Specifically, $\delta(\mathbf{r}_0) = 0$ leads to $\sum_{i=1}^{3} C_{0i,i} = 0$ and $\sum_{i=1}^{3}\sum_{j=1}^{3} D_{0i,ij}\Delta x_j = 0$ for the quadratic expression of **B** given in **Eq. 5**. Likewise, the plasma current density **J** can also be evaluated analytically from the modeled **B** field by **Eq. 1a**. When using these analytic expressions, the field values and constraints evaluated at barycenter, four vertices and four face centers, such as $\mathbf{J}(\mathbf{r}_{F\alpha}) = \mathbf{J}_{F\alpha}$ and $\delta(\mathbf{r}_0) = \delta(\mathbf{r}_\alpha) = \delta(\mathbf{r}_{F\alpha}) = 0$, are of particular importance.

Given MMS measurements at the vertices ($\mathbf{r}_\alpha$) of the magnetic field ($\hat{\mathbf{B}}$) from the MMS Fluxgate Magnetometer (FGM) instruments (Russell et al., 2016) and particle current density ($\hat{\mathbf{J}}$) from the Fast Plasma Investigation (FPI) sensors (Pollock et al., 2016), the model parameters $\{B_{0i}, C_{0i,j}, D_{0i,jk}\}$ in **Eq. 5** can often be derived by minimizing a loss function as defined below. Here, the loss function characterizes 1) the model-measurement differences between the modeled (**B**, **J**) and measured ($\hat{\mathbf{B}}, \hat{\mathbf{J}}$) parameters and 2) the model departures corresponding to the violation of the MHD **Eqs. 1a**, b. For a linear system, the minimization procedure can also be reduced to solving a set of linear algebraic equations (e.g., Menke, 1989). Depending on whether the number of the adopted constraints is smaller than, equal to, or greater than the number of model parameters, the solution derived from the least-squares method could be under-, even-, or over-determined, respectively. Previous reconstruction models have focused on the even-determined solutions of a quadratic loss function (e.g., Dunlop et al., 1988; Torbert et al., 2020), for which the measurement errors were not explicitly considered. The new ER model presented here adopts a new method that derives the model parameters by directly minimizing a generalized loss function using a stochastic optimization method that contains random

measurement errors and consists of a flexible number of constraints. As a result, the solution is always programmatically feasible regardless of whether the physical constraints defined by the MHD equations are linear or nonlinear and whether the system is under-, even-, or over-determined.

The generalized loss function (L) has the following form

$$L = L_O + w_A\varepsilon_A L_A + w_B\varepsilon_B L_B + w_C\varepsilon_C L_C, \tag{6}$$

where the individual components of the loss function (L_O, L_A, L_B, L_C) are given by

$$L_O = \frac{1}{12}\sum_{\alpha=1}^{4}\sum_{i=1}^{3}\left[B_i(\mathbf{r}_\alpha) - \hat{B}_{\alpha,i}\right]^2, \tag{7a}$$

$$L_A = \frac{1}{12}\sum_{\alpha=1}^{4}\sum_{i=1}^{3}\left[J_i(\mathbf{r}_\alpha) - \hat{J}_{\alpha,i}\right]^2, \tag{7b}$$

$$L_B = \frac{1}{9}\left[\delta^2(\mathbf{r}_0) + \sum_{\alpha=1}^{4}\delta^2(\mathbf{r}_\alpha) + \sum_{\alpha=1}^{4}\delta^2(\mathbf{r}_{F\alpha})\right]\text{or}$$
$$L_B^* = \frac{1}{5}\left[\delta^2(\mathbf{r}_0) + \sum_{\alpha=1}^{4}\delta^2(\mathbf{r}_\alpha)\right],\text{and} \tag{7c}$$

$$L_C = \frac{1}{4}\sum_{\alpha=1}^{4}\left[\mu_0\mathbf{J}(\mathbf{r}_{F\alpha})\cdot\left(\Delta\mathbf{r}_{\beta\gamma}\times\Delta\mathbf{r}_{\beta\delta}\right) - \left(\bar{\mathbf{B}}_{\beta\gamma}\cdot\Delta\mathbf{r}_{\beta\gamma} + \bar{\mathbf{B}}_{\gamma\delta}\cdot\Delta\mathbf{r}_{\gamma\delta} + \bar{\mathbf{B}}_{\delta\beta}\cdot\Delta\mathbf{r}_{\delta\beta}\right)\right]^2 \tag{7d}$$

with $\Delta\mathbf{r}_{\beta\gamma} = (\mathbf{r}_\gamma - \mathbf{r}_\beta)$ being the edge vector connecting the vertices $\mathbf{r}_\beta$ and $\mathbf{r}_\gamma$ and $\bar{\mathbf{B}}_{\beta\gamma} = (\hat{\mathbf{B}}_\beta + \hat{\mathbf{B}}_\gamma)/2$ being the mean magnetic field on the edge $\Delta\mathbf{r}_{\beta\gamma}$ calculated by the measured $\hat{\mathbf{B}}$ field by using a linear approximation to obtain the field along an edge, between two spacecraft measurements. In **Eq. 7**, we use i to denote the dimensional index ranging 1-3 and use Greek letters to denote tetrahedron points or faces ranging 1–4. The components L_O and L_A each consist of twelve terms or twelve constraints and represent the differences of the modeled and measured fields at the vertices $\mathbf{r}_\alpha$. Thus, L_O and L_A correspond to the model-measurement differences in the loss function. The component L_B consists of nine physical constraints, which requires minimization of $\delta^2(\mathbf{r}) = (\nabla\cdot\mathbf{B})^2$ at nine particular spatial points (i.e., one barycenter $\mathbf{r}_0$, four vertices $\mathbf{r}_\alpha$ and four face centers $\mathbf{r}_{F\alpha}$). Because the measurements do not directly enter the expression, L_B corresponds to the model departures or violations from the above MHD equations. L_B can be replaced by L_B^*, which neglects the face-center constraints. The component L_C consists of four approximate physical constraints derived from the generic MHD equation obtained by applying Stokes' Theorem to Ampere's Law ($\mu_0\iint\tilde{\mathbf{J}}\cdot d\mathbf{S} = \oint\hat{\mathbf{B}}\cdot dl$) on the four tetrahedron faces, which derives the current density components normal to the tetrahedron faces ($\tilde{\mathbf{J}}$) by using the linear curlometer technique from the measured $\hat{\mathbf{B}}$ (Dunlop et al., 1988). A minimization between $\tilde{\mathbf{J}}$ and $\mathbf{J}$ projecting onto the normal directions of four tetrahedron faces yields L_C. Thus, L_C also possesses the nature of the model-measurement differences. Note that, as previously denoted, each face-center $\mathbf{r}_{F\alpha}$ in **Eq. 7d** is defined by other three vertices $(\mathbf{r}_\beta, \mathbf{r}_\gamma, \mathbf{r}_\delta)$. Specification of the weighting factors (w_A, w_B, w_C) in **Eq. 6** determines the selection of the loss function components to be included in the reconstruction model. The scaling parameters $(\varepsilon_A, \varepsilon_B, \varepsilon_C)$ in **Eq. 6** depend on the characteristic length scale of the tetrahedron and the dimensional factors of the loss functions. We will discuss the settings of these parameters in more detail below.

We first note the similarities and differences between L_A and L_C in **Eqs. 6**, **7**. Both loss function components adopt the differences in current densities as constraints. L_A is the difference between the modeled $\mathbf{J}$ and the measured particle current density $\hat{\mathbf{J}}$ at four vertices whereas L_C is the difference between the modeled $\mathbf{J}$ components and the current density $\tilde{\mathbf{J}}$ components derived from the curlometer technique (i.e., using $\hat{\mathbf{B}}$) on the four tetrahedron face-centers. When both $\hat{\mathbf{B}}$ and $\hat{\mathbf{J}}$ are available and include direct measurement errors of the same order, L_A is more accurate to be included in the generalized loss function L than L_C because the $\tilde{\mathbf{J}}$ value used in L_C contains additional errors due to the linear approximation assumed in the curlometer technique. On the other hand, if only $\hat{\mathbf{B}}$, but not $\hat{J}$, is available (in which case L_A will not be available) or if the errors in $\hat{\mathbf{J}}$ are far greater than those in $\hat{\mathbf{B}}$, then L_C is preferred to L_A for inclusion in L. In Denton et al. (2020), $\tilde{\mathbf{J}}$ derived from the curlometer technique is used to modify the particle current density $\hat{\mathbf{J}}$ to produce a composite current density, which together with the measured $\hat{\mathbf{B}}$ is used to build the reconstruction model. Our approach of introducing different constraints L_A and L_C for different current densities $\hat{\mathbf{J}}$ (measured directly by FPI) and $\tilde{\mathbf{J}}$ (derived from the curlometer technique) evaluated at different spatial locations provides a clear physical significance and algorithmic flexibility.

Application of a Stochastic Optimization Algorithm to Solve for Model Parameters and Selection of Loss Function Components

In this new 3D ER model, the model parameters in **Eq. 5** are solved by directly minimizing the loss function L defined in **Eq. 6** using a stochastic optimization algorithm called the SPSA method (Spall, 1998a; Spall, 1998b; Spall, 2003) through an iterative procedure that also naturally incorporates the errors for the measured fields $(\hat{\mathbf{B}}, \hat{\mathbf{J}})$. A comprehensive introduction to the algorithm with detailed procedures of implementation to the current problem is presented in **Supplementary Appendix A**. Note that the generalized loss function L defined by **Eqs. 6**, **7** is in a quadratic form with respect to the model parameters $\{B_{0i}, C_{0i,j}, D_{0i,jk}\}$ because the MHD **Eqs. 1a**, b are linear. Minimization of a quadratic loss function is equivalent to solving a set of linear algebraic equations for the model parameters (e.g., Menke, 1989; Axelsson, 1996). When the model parameters are obtained by directly minimizing the loss function L the corresponding MHD system could be either linear or nonlinear. The nonlinearity occurs in our new 3D ER model when **Eqs. 2**, **3** are also included as additional constraints. Nonlinear systems are not unusual in various empirical models. For example, in Roelof et al. (1993), the loss function for reconstructing global magnetospheric images based on the

extreme ultraviolet (EUV) and energetic neutral atom (ENA) measurements is highly nonlinear, for which the loss function can only be directly minimized. Furthermore, when the loss function contains measurements, it also contains random measurement errors. The SPSA method effectively solves problems containing random errors by including the errors in the solutions. In addition, we will show later through examples that the SPSA method can solve slightly under-determined problems that could not be solved directly by the traditional least-squares approach.

In practice, the SPSA method solves for the model parameters that minimize the following dimensionless loss function (y) with a random perturbation that characterizes the measurement errors (**Supplementary Eqs. A4a, b** in **Supplementary Appendix A**)

$$L_\theta = \sqrt{L}\Big/B_{00} \text{ and} \tag{8a}$$

$$y = L_\theta + \varepsilon_\sigma, \tag{8b}$$

where L is given by **Eq. 6**, B_{00} is the measured mean magnetic field (defined by **Supplementary Eq. A1** in **Supplementary Appendix A**) used to normalize the general loss funciton L, $\varepsilon_\sigma = N(0, \sigma^2)$ represents a random variable having a normal distribution with zero mean and σ^2 variance that characterizes the random measurement errors. The first-order SPSA algorithm is adopted to solve for the model parameters in this paper. The specifications of various model parameters including the weighting coefficients and scaling parameters in **Eq. 6** and the algorithmic procedures of the recursive formulations are presented in **Supplementary Appendix A**. In **Section 3**, we will detail the application of this SPSA-based ER model to a specific EDR case using MMS measurements and discuss the relationship between the SPSA model variance σ^2 in **Eq. 8b** and the variances of the random errors of the measured $\hat{\mathbf{B}}$ and $\hat{\mathbf{J}}$ fields, σ_B^2 and σ_J^2, respectively. Note that L_θ in **Eq. 8a** is a deterministic variable whereas y in **Eq. 8b** is a randome variable. A stochastic optimation method such as SPSA algorithm optimizes loss functons associated with randome variables.

When all the loss function components in **Eq. 6** (L_O, L_A, L_B, L_C) are included, then, the total number of constraints is thirty-seven (37). This number is greater than the number of model parameters (30) and the problem is significantly over-determined. Because L_C adopts a linear approximation in the curlometer technique it is expected to introduce additional errors in the modeled fields near the reconnection regions where the field curvature is large. As a result, our default setting for the reconstruction model is to set $w_C = 0$, i.e., to not include L_C in the generalized loss function L. This reduces the total number of constraints for the default setting to thirty-three (33) and thus renders the problem, i.e., solving thirty model parameters, only slightly over-determined. The model departures in the loss function component L_B shown in **Eq. 7c** are the application of the MHD equation $\nabla \cdot \mathbf{B} = 0$ to nine particular points on the tetrahedron (the four vertices, the four face-centers, and the barycenter). When L_B is replaced by L_B^* that only applies $\nabla \cdot \mathbf{B} = 0$ to the barycenter plus four vertices, the total number of the constraints is reduced to twenty-nine (29) and the problem becomes slightly under-determined. Our numerical experiments show that the model parameters resulting from the SPSA method yield only slight and negligible (~1–3%) differences when the problem is changed between slightly over-determined and slightly under-determined. On the other hand, the numerical solution to a set of under-determined linear algebraic equations no longer exists or cannot be calculated directly if the problem were solved by the traditional least-squares method (e.g., Menke, 1989).

To explain why using L_B and L_B^* does not lead to significantly different solutions, we first note that for an even-determined or an over-determined problem with a quadratic loss function, a unique solution can be derived either by directly solving an optimization problem or by solving a set of linear algebraic equations (e.g., Axelsson, 1996; Chong and Zak, 2001). It is also noted that for an over-determined problem, the inclusion of additional measurements or constraints may not change noticeably the existing solution if the newly added constraints are redundant (e.g., Menke, 1989). For an under-determined problem where the number of constraints is less than that of the model parameters, however, the set of linear algebraic equations becomes undetermined and one is no longer able to uniquely solve for the model parameters. Returning to the expressions of the loss functions in **Eqs. 6–8**, we note that the roles of model parameters and constraints (e.g., $\mathbf{B}$ vs. $\hat{B}$, or $\mathbf{J}$ vs. $\hat{\mathbf{J}}$) do not show preference to one or the other. A minimized or a least-squares solution is always formally available for given numbers of model parameters and constraints regardless of their relative magnitudes. Adding four constraints of $\nabla \cdot \mathbf{B} = 0$ to the four tetrahedron faces is expected to be largely redundant to the already existing constraints of $\nabla \cdot \mathbf{B} = 0$ at the barycenter and four vertices, thus leading to only slight modifications to the model parameters. Again, it is noted that unlike $(\hat{\mathbf{B}}, \hat{\mathbf{J}})$ that are only available on the four vertices, the analytic $\mathbf{B}$-field as expressed by **Eq. 5** and all its derived fields such as $\mathbf{J}$ and $\nabla \cdot \mathbf{B}$ are available on any spatial point. Furthermore, in terms of the uniqueness of the solution, either the random noise term or the under-determined constraints in the loss function y in **Eq. 8** could lead to the non-uniqueness of the solution. Note that the stochastic optimization algorithm minimizes the random variable y defined by **Eq. 8b** rather than the deterministic physical loss function L_θ defined by **Eq. 8a**. We will discuss this issue in more detail in the next section. From the perspective of constraint redundancy, it is also noted that given the analytic expression in **Eq. 5** for $\mathbf{B}$, the relation $\nabla \cdot \mathbf{J} = \nabla \cdot (\nabla \times \mathbf{B}/\mu_0) \equiv 0$ will be automatically satisfied regardless of what the model parameters are. As a result, one cannot introduce a constraint component for $\mathbf{J}$ similar to L_B based on the redundant relation of $\nabla \cdot \mathbf{J} = 0$.

RESULTS

To test our new model and to also demonstrate the third step of diagnosing the accuracy and the quality of the reconstructed fields while building an ER model, we use MMS measurements from the magnetotail EDR event of 11 July 2017. During this

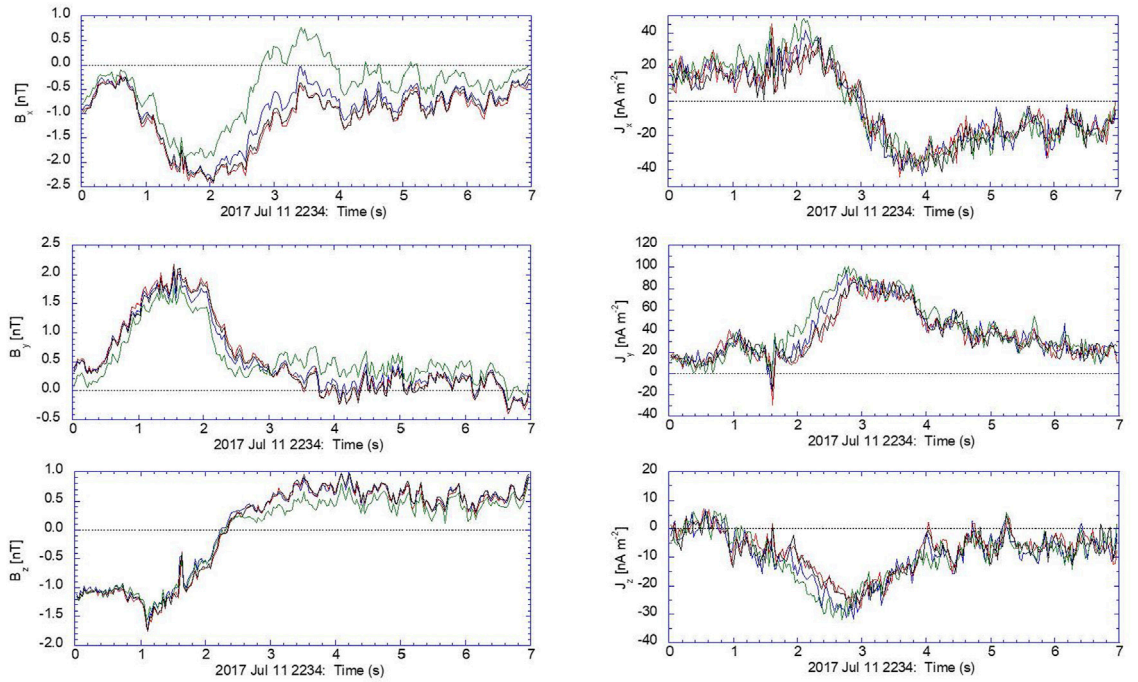

FIGURE 1 | Measurements from the four MMS spacecraft (MMS1, MMS2, MMS3, MMS4) on 11 July 2017 showing (left) the measured magnetic field $\hat{B} = (B_x, B_y, B_z)$ from FGM (Russell et al., 2016) and (right) particle current density $\hat{J} = (J_x, J_y, J_z)$ from FPI (Pollock et al., 2016) in GSE coordinates.

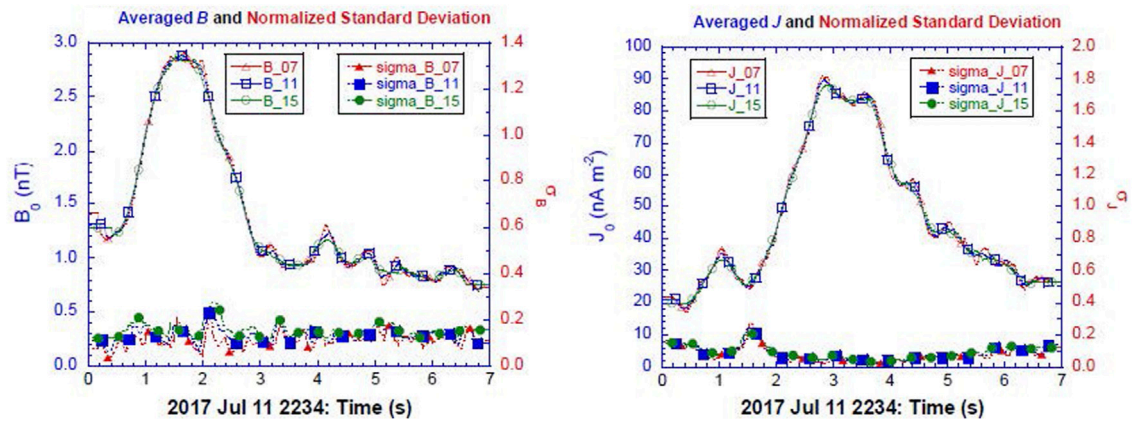

FIGURE 2 | Mean and normalized standard deviation fields derived from the MMS measurements. The means and standard deviations are calculated on a moving window with a width of 7 (red), 11 (blue), and 15 (green) time steps, respectively. Panels **(A)** and **(B)** correspond to the **B** field and **J** field, respectively.

event, the MMS constellation traversed a reconnection region in the earthward and northward directions while remaining near the neutral plane (Torbert et al., 2018). **Figure 1** shows 7 s of magnetic field ($\hat{\mathbf{B}}$) and particle current density ($\hat{\mathbf{J}}$) measurements starting at 22:34 UT. Since $\hat{\mathbf{B}}$ and $\hat{\mathbf{J}}$ are measured and processed at different sampling rates and the loss function L shown in **Eq. 6** is assumed to be evaluated simultaneously, we have interpolated the measured fields onto the same time resolution with a time interval of $\Delta t = 0.0293$ s, which corresponds to a sampling frequency of 34 Hz.

Error Consideration and Quality Indicators

Note that the measurement errors here include uncertainties in both the instrumentation and subsequent processing of the data. However, the random errors in **Eq. 8** are associated with the unbiased instrument noise. Here, we estimate the errors by directly calculating the parameter variability included in the data series. In **Figure 2**, we show both the means (B_0, J_0) and the normalized standard deviations (σ_B, σ_J) of the magnitudes for the measured $\hat{\mathbf{B}}$ and $\hat{\mathbf{J}}$ fields. The averages are taken over the four spacecraft and over moving windows with widths of 7, 11, and

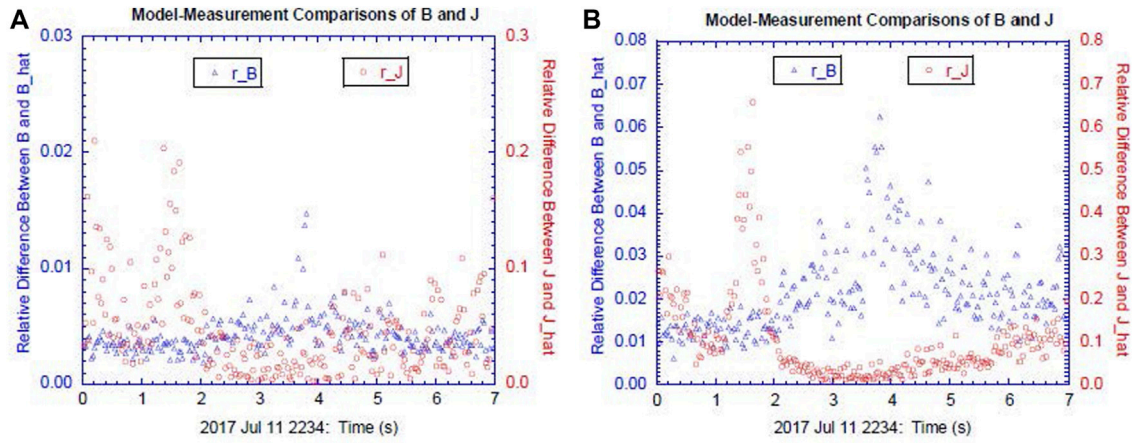

FIGURE 3 | Relative differences (γ_B, γ_J) of the reconstructed **(B,J)** fields based on MMS measurements shown in **Figure 1**. Panels **(A)** and **(B)** correspond to two cases of a standard setting of σ in **Eq. 8** with $\sigma = 0.1\bar{\sigma}_B$ and an alternative setting of $\sigma = \bar{\sigma}_B$, respectively.

15 time steps, respectively. The calculated B_0 is approximately equal to the characteristic value of the temporal mean of the magnetic field B_{00} defined in **Supplementary Eq. A1**. It is noted that the mean fields are not noticeably sensitive to the width of the moving window. This implies that the sampling rate of the measurements is high enough to resolve the temporal variability of the fields. It is also noted from **Figure 2** that there is no systematic variation of σ_B with respect to B_0, whereas σ_J is inversely proportional to J_0. The weighting factor w_A in **Eq. 6** is proportional to the σ_B^2/σ_J^2 parameter that can be calculated from the values shown in the figure. The weighting factors (w_B, w_C) are prescribed to $(1, 0)$ for the default setting of the reconstruction model. Note that setting $w_B = 1$ here also means that we give no preference between the model-measurement differences L_0 and the model departures L_B. To set the final model parameter σ used in **Eq. 8**, we note that σ_B directly derived from the measured $\hat{\mathbf{B}}$ contains both the unbiased random errors required for the construction of ε_σ in **Eq. 8** and possibly also the biased errors associated with the parameter retrieval and data processing issues. In addition, a smaller σ in the random loss function γ will yield a more numerically accurate solution, though its usefulness may be limited by the measurement errors; any numerical accuracy achieved that is higher than the measurement errors after setting $\sigma \to 0$ does not contain additional information as the results are ultimately limited by the uncertainty in the measurements. As a result, our default setting for σ in the algorithm as shown in **Eq. 8b** takes a conservative value of $\sigma = 0.1\bar{\sigma}_B$, where $\bar{\sigma}_B$ is the time-averaged standard deviation of $\hat{\mathbf{B}}$ as shown in **Figure 2A**.

Given model parameters $\{B_{0i}, C_{0i,j}, D_{0i,jk}\}$, a smooth 3D solution for $(\mathbf{B}, \mathbf{J})$ can be plotted to be visulized. But before addressing these visualizations, we begin our discussions here with evaluations of the quality factors associated with these results. In **Figure 3**, we show the relative differences of the fields $(\mathbf{B}, \mathbf{J})$ reconstructed at every time step based on the MMS-measured fields $(\hat{\mathbf{B}}, \hat{\mathbf{J}})$ shown in **Figure 1**. The indices (γ_B, γ_J) can be considered as the normalized loss function components (L_O, L_A) corresponding to the model-measurement differences, which can be used as a set of accuracy indicators of the reconstruction model and are defined as:

$$\gamma_B = \sqrt{\frac{\sum_{\alpha=1}^{4}\sum_{i=1}^{3}\left[B_i(\mathbf{r}_\alpha) - \hat{B}_{\alpha,i}\right]^2}{\sum_{\alpha=1}^{4}\sum_{i=1}^{3}\hat{B}_{\alpha,i}^2}} \text{ and} \tag{9a}$$

$$\gamma_J = \sqrt{\frac{\sum_{\alpha=1}^{4}\sum_{i=1}^{3}\left[J_i(\mathbf{r}_\alpha) - \hat{J}_{\alpha,i}\right]^2}{\sum_{\alpha=1}^{4}\sum_{i=1}^{3}\hat{J}_{\alpha,i}^2}}. \tag{9b}$$

The results from a pair of reconstructions with $\sigma = 0.1\bar{\sigma}_B$ and $\sigma = \bar{\sigma}_B$, respectively, are presented in **Figure 3**. The default setting, which has a smaller measurement noise of $\sigma = 0.1\bar{\sigma}_B$, yields a more accurate reconstruction field as characterized by smaller indices (γ_B, γ_J). On the other hand, if the measurement noise in the loss function γ amounts to $\bar{\sigma}_B$, such that $\sigma \sim \bar{\sigma}_B \sim 0.1$ as shown in **Figure 2**, then, a numerical solution of $\mathbf{B}$ with $\gamma_B < \bar{\sigma}_B$ can be considered to be an acceptable or valid solution. Our default setting of $\sigma = 0.1\bar{\sigma}_B$ leads to a numerical solution of $\mathbf{B}$ with $\gamma_B \ll \bar{\sigma}_B$, which can be considered an accurate solution. It should also be noted that because of the existence of measurement errors in $\hat{\mathbf{B}}$ (i.e., $\sigma > 0$), a deterministic and idealized solution with $\gamma_B \equiv 0$ is considered to be as accurate as one with $\gamma_B < \sigma$. **Figure 3** shows that far greater errors exist in the modeled current density γ_J than those in the magnetic field γ_B. This is largely expected since the modeled current density $\mathbf{J}$ is a quantity derived from the prescribed $\mathbf{B}$ field and contains fewer free parameters and therefore is expected to lead to greater errors in $\mathbf{J}$ than in $\mathbf{B}$. This is another reason for us to set σ so that it is much smaller than $\bar{\sigma}_B$ in **Eq. 8**, which yields a solution also with an acceptable error in the reconstructed $\mathbf{J}$ field. Comparison between the two panels in **Figure 3** shows that the magnitude of the errors in the reconstruction model is

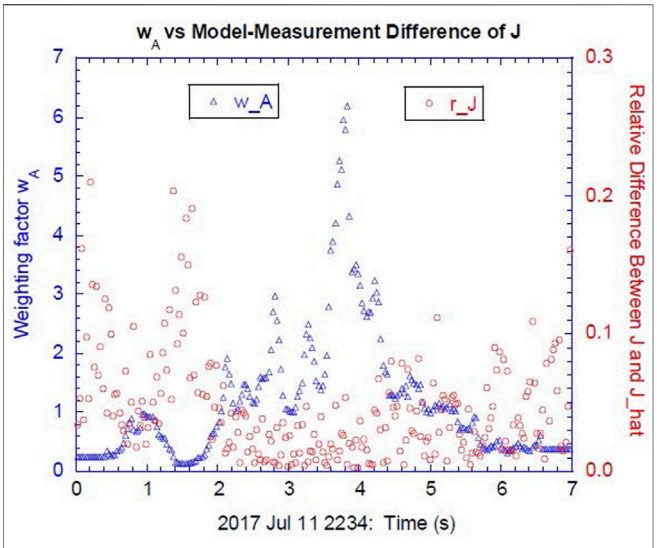

FIGURE 4 | Weighting factor w_A versus the relative difference γ_J for the reconstructed current density **J** shown in **Figure 3**.

sensitive to the measurement errors. This feature can be further confirmed by examining γ_J variation with the time-dependent measurement errors. We note from **Figure 2** that σ_J changes more significantly with time than σ_B, which leads to a significant variation in the weighting factor w_A. **Figure 4** shows both w_A ($= \sigma_B^2/\sigma_J^2$) and γ_J for a standard setting of $\sigma = 0.1\bar{\sigma}_B$. The figure shows a negative correlation between w_A and γ_J. Since errors in $\hat{\mathbf{B}}$ are nearly constant, **Figure 4** shows a strong positive correlation between the errors in the measured $\hat{\mathbf{J}}$ and the modeled **J**. Overall, **Figures 2–4** show that the accuracy of the reconstructed fields from the stochastic optimization algorithm is limited by the measurement uncertainties, with more accurate measurements unsurprisingly resulting in a more accurate solution for the reconstruction based upon those measurements.

Unlike reconstruction models based on the GSR technique, where the reconstructed fields are mainly derived by various physical relations, the new ER model presented here is mainly data-driven, directly fitting the modeled fields to the measured fields. Since the design of the general loss function highlighted in **Eqs. 6**, **7** also contains a component of the model departures characterizing a few physical constraints, the validity of those constraints can be used as a measure of the quality of the ER model in addition to the two indices (γ_B, γ_J) for measuring the accuracy of the solution. Here, one important constraint is the vanishing of the divergence of the magnetic field ($\nabla \cdot \mathbf{B} = 0$), which is also used as a constraint of the loss function component L_B in **Eq. 7**. Dunlop et al. (1988) introduced an index of the ratio of the divergence to the vorticity of the magnetic field as a quality indicator to measure the robustness of the reconstructed current density **J** field. In **Figure 5**, we show the following two quality indicators Q_{model} and Q_{curl} representing the ratio of the divergence to the vorticity of the magnetic field based on the reconstructed **B** field and the measured $\hat{\mathbf{B}}$ field by curlometer technique, respectively. These are defined as

$$Q_{\text{model}} = \sqrt{\frac{\sum_{\alpha=1}^{4} (\nabla \cdot \mathbf{B})_\alpha^2}{\sum_{\alpha=1}^{4}\sum_{i=1}^{3} (\nabla \times \mathbf{B}|_{\alpha,i})^2}} \text{ and} \tag{10a}$$

$$Q_{\text{curl}} = \left[\frac{|\nabla \cdot \hat{\mathbf{B}}|}{|\nabla \times \hat{\mathbf{B}}|}\right]_{\text{curlometer}}. \tag{10b}$$

In the above, Q_{model} is calculated by evaluating $\nabla \cdot \mathbf{B}$ and $\nabla \times \mathbf{B}$ analytically based on the modeled **B** field at the four vertices, whereas Q_{curl} is calculated by evaluating the volume-averaged $|\nabla \cdot \hat{\mathbf{B}}|$ and $|\nabla \times \hat{\mathbf{B}}|$ based on the measured $\hat{\mathbf{B}}$ field following the schemes shown in Dunlop et al. (2002) and Middleton and Masson (2016). Note that Q_{curl} has also been used as an objective index that measures the quality of the **J** fields reconstructed from the curlometer technique (Dunlop et al., 2002). On the other hand, the index Q_{model} defined in **Eq. 10a**

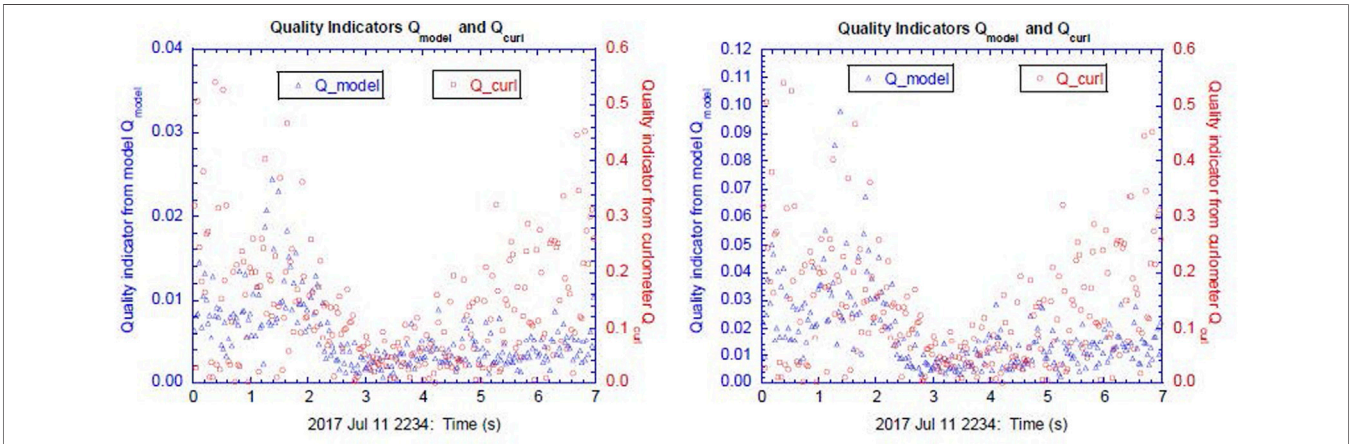

FIGURE 5 | Quality indicators Q_{model} and Q_{curl} representing the ratio of the divergence to the vorticity of the magnetic field based on the reconstructed **B** field and the measured $\hat{B}$ field by curlometer technique, respectively. The left and right panels correspond to two cases of a standard setting of σ in **Eq. 8** with $\sigma = 0.1\bar{\sigma}_B$ and an alternative setting of $\sigma = \bar{\sigma}_B$, respectively.

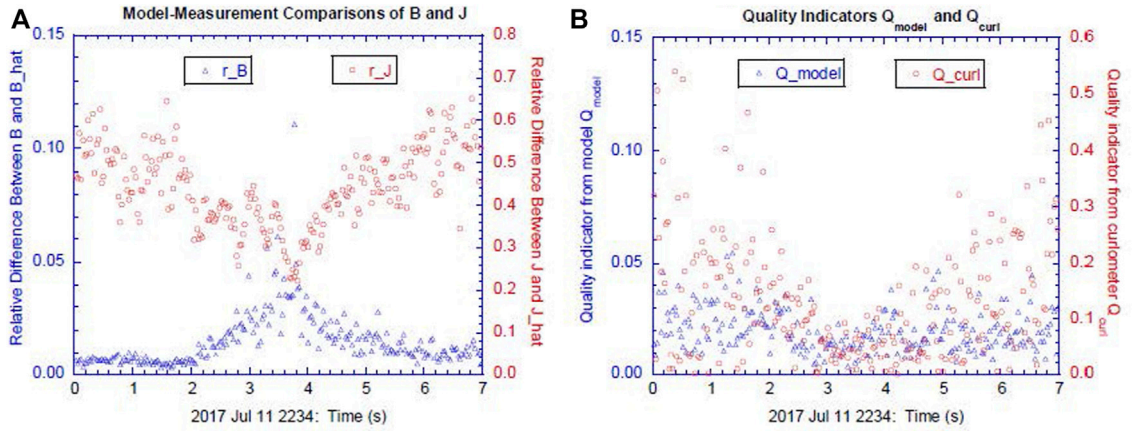

FIGURE 6 | (A) Relative differences (γ_B, γ_J) and **(B)** quality indicators $(Q_{\text{model}}, Q_{\text{curl}})$ for a sensitivity run with weighting factors w_B and w_C being both set to 1: $(w_B, w_C) = (1, 1)$.

measures the quality or robustness of both **B** and **J** fields derived from the ER model. **Figure 5** shows that typically $Q_{\text{model}} \leq 0.01$ for a standard setting of the model parameter $\sigma = 0.1\bar{\sigma}_B$, whereas the typical values of Q_{curl} are much greater, $Q_{\text{curl}} \geq 0.1$. Since Q_{curl} is calculated directly by a linear approximation from the measured $\hat{B}$ field, it contains errors from both the linear approximation and measurement errors (Dunlop et al., 1988; Dunlop et al., 2002). Comparison between the two panels in **Figure 5** also shows that the increase in the measurement errors by changing $\sigma = 0.1\bar{\sigma}_B$ to $\sigma = \bar{\sigma}_B$ only increases the quality indicator Q_{model} by a factor of ~3 (note the changed scale on the ordinate between the two panels for just the "model" result, not the "curl" result). The relation of $Q_{\text{model}} \ll Q_{\text{curl}}$ is still valid for $\sigma = \bar{\sigma}_B$. As a result, we can conclude based on **Figure 5** that the uncertainties in the reconstructed fields based on the MMS-measured $\hat{\mathbf{B}}$ near the EDR using the curlometer technique are mostly contributed by the linear approximation used in the technique. It should be pointed out that when an ER model is formulated and solved as an even-determined problem based on the traditional least-squares method, one may impose a condition of vanishing Q_{model} everywhere $(Q_{\text{model}} \equiv 0)$. In this case, an alternative constraint corresponding to model departures, say, a vanishing variance of the modeled **B** field in a particular diretion M, i.e., $\partial^2\mathbf{B}/\partial M^2 = 0$, needs to be introduced into the reconstruciton model (e.g., Torbert et al., 2020).

We now turn to the loss function component L_C. The default setting of the weighting factors is $(w_B, w_C) = (1, 0)$. This means that the constraint of the precise physical relation of $\nabla \cdot \mathbf{B} = 0$ is fully utilized whereas the constraint of matching the modeled current components to the ones derived from the curlometer technique on the tetrahedron faces is neglected. Again, setting $w_B = 1$ here also means that we give no preference between two sets of constraints of the model-measurement differences and the model departures. Our analysis of the quality indicators $(Q_{\text{model}}, Q_{\text{curl}})$ derived from the runs without L_C shown in **Figure 5** can be considered as a rationale for the default setting of $w_C = 0$. It is noted that the curlometer technique developed in Dunlop et al. (1988) and Middleton and Masson (2016) applies a linear approximation to the entire volume of the tetrahedron, whereas in L_C the linear approximation applies only to the four individual tetrahedron faces. Hence, it is worthwhile examining quantitatively the effect of the loss function component L_C on the performance of the reconstruction model. In **Figure 6**, we show the indices (r_B, r_J) and the quality indicators $(Q_{\text{model}}, Q_{\text{curl}})$ for a sensitivity run of the reconstruction model with all parameters in default settings, except w_C that is set to 1. Comparing **Figure 6A** with **Figure 3A**, we find that the inclusion of L_C significantly reduces the accuracy of the reconstructed fields. This is expected because the inclusion of L_C not only introduces a linear approximation in the calculation of the current density **J** from the magnetic field **B**, but also enhances the degree of over-determination of the model. Both of these are expected to increase the errors of a least-squares solution. Comparing **Figure 6B** with **Figure 5A** on the modeled quality indicators Q_{model} derived from different runs underscores the same conclusion—i.e., that the inclusion of L_C makes the model performance worse. However, **Figure 6B** also shows that Q_{model} is still significantly smaller than Q_{curl} (note the different scales), indicating that a linear approximation in a loss function component only partially affects the model performance. This sensitivity investigation of setting $w_C = 1$ also demostrates the flexibility of the new ER model that directly minimizes the general loss function with its components being able to be included or excluded without changing the model framework.

At this stage, it is also interesting to examine a largely under-determined setting of excluding both L_B and L_C in the generalized loss function L by setting $(w_B, w_C) = (0, 0)$ in **Eq. 6**. There are only twenty-four (24) constraints in L_O and L_A, all given by the MMS measurements, whereas the reconstruction model contains thirty (30) model parameters that need to be determined. Hence, the problem is largely under-determined and the solution cannot be uniquely solved. For a stochastic optimization, such as the one based on the SPSA method, there

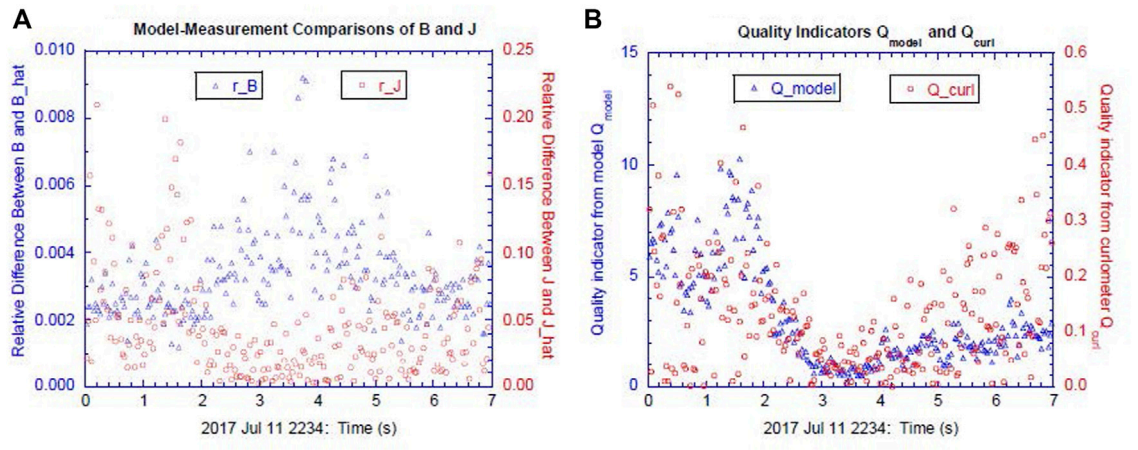

FIGURE 7 | (A) Relative differences (γ_B, γ_J) and **(B)** quality indicators $(Q_{\text{model}}, Q_{\text{curl}})$ for a sensitivity run with weighting factors w_B and w_C being both set to 0: $(w_B, w_C) = (0, 0)$.

is no fundamental difference in the non-uniqueness of the solution either due to the lack of constraints or due to random errors in the loss function. In other words, the unknown parameters for the reconstruction model can always be formally solved by minimizing the generalized loss function y in **Eq. 8**. **Figure 7** shows the indices (γ_B, γ_J) and the quality indicators $(Q_{\text{model}}, Q_{\text{curl}})$ for a sensitivity run of the reconstruction model that sets $(w_B, w_C) = (0, 0)$. We note from **Figure 7A** that the relative differences between the modeled and measured fields (γ_B, γ_J) are much less than those shown in **Figures 3A**, **6A**. However, the quality indicator Q_{model} shown in **Figure 7B** is much greater than those derived by any approach shown above including Q_{curl} derived by the curlometer technique. **Figure 7** shows that even though one can construct an empirical model that leads to a very good fit between the modeled and the measured fields at the prescribed spatial points, the fields may not necessarily satisfy some physical relations, such as $\nabla \cdot \mathbf{B} = 0$. This is due to the following two facts: 1) the fields contain errors, either in the measured field or in the modeled field derived from the measurements and 2) the numerical evaluation of the physical relation based on the discrete measurements involves a small difference between two large quantities. The divergence of a vector field contains two components of variation corresponding to variations in the magnitude and direction of the vector. For a deformation vector field that is mainly confluent-diffluent—i.e., divergence is mainly caused by the change in direction—the calculation of the divergence of the vector field generally involves a small difference of two large quantities (e.g., Holton, 2004). In this case, small errors in the **B** field will be greatly amplified in calculating $\nabla \cdot \mathbf{B}$ unless an additional constraint or assumption of $\nabla \cdot \mathbf{B} = 0$, or $|\nabla \cdot \mathbf{B}|$ being small, is explicitly included in the model or algorithm development. A similar assumption of "charge neutrality" in plasma physics is also used as an explicitly imposed constraint in developing various MHD models (e.g., Gurnett and Bhattacharjee, 2005).

The other more important implication of this test run for a largely under-determined setting with only 24 constraints for a 30-parameter reconstruction model is that the current ER model can be directly applied to reconstructing fields with a set of incomplete measurements. The stochastic optimization algorithms can solve for model parameters under the same algorithmic framework regardless whether the problem is over- or under-determined. For example, for a default setting of the current ER model with 33 constraints, the algorithm can be directly applied to an incomplete set of MMS measurements if the $(\hat{\mathbf{B}}, \hat{\mathbf{J}})$ measurements from one spacecraft are not available. Under such a circumstance, the same algorithm with 27 (= 33–6) constraints will produce a reconstruction field $(\mathbf{B}, \mathbf{J})$ that fits the measured $(\hat{\mathbf{B}}, \hat{\mathbf{J}})$ at three vertices having available measurements plus $\nabla \cdot \mathbf{B} = 0$ being satisfied at all four vertices, all within the measurement errors. The results shown in **Figure 7** also suggest that the deterioration of the reconstructed fields due to lack of the needed constraints is gradual. On the other hand, the algorithm based on the traditional least-squares method that solves a set of linear algebraic equations (e.g., Torbert et al., 2020) becomes inapplicable once the problem changes from an even- to under-determined one.

The Reconstructed Fields

We now present the reconstructed fields based on the MMS measurements shown in **Figure 1**. A reconstruction model can be developed in either an *L-M-N* coordinate system derived from the minimum variance analysis or in a fixed system, such as GSE that is used in the present reconstruction model. One purpose of adopting the *L-M-N* coordinate system to develop a reconstruction model is to take advantage of the ability to neglect changes in the minimum variance direction to convert a slightly under-determined problem into an even-determined one (Denton et al., 2020; Torbert et al., 2020). When the reconstructed field varies rapidly with time, the constructed *L-M-N* coordinate may also change accordingly. Under such a

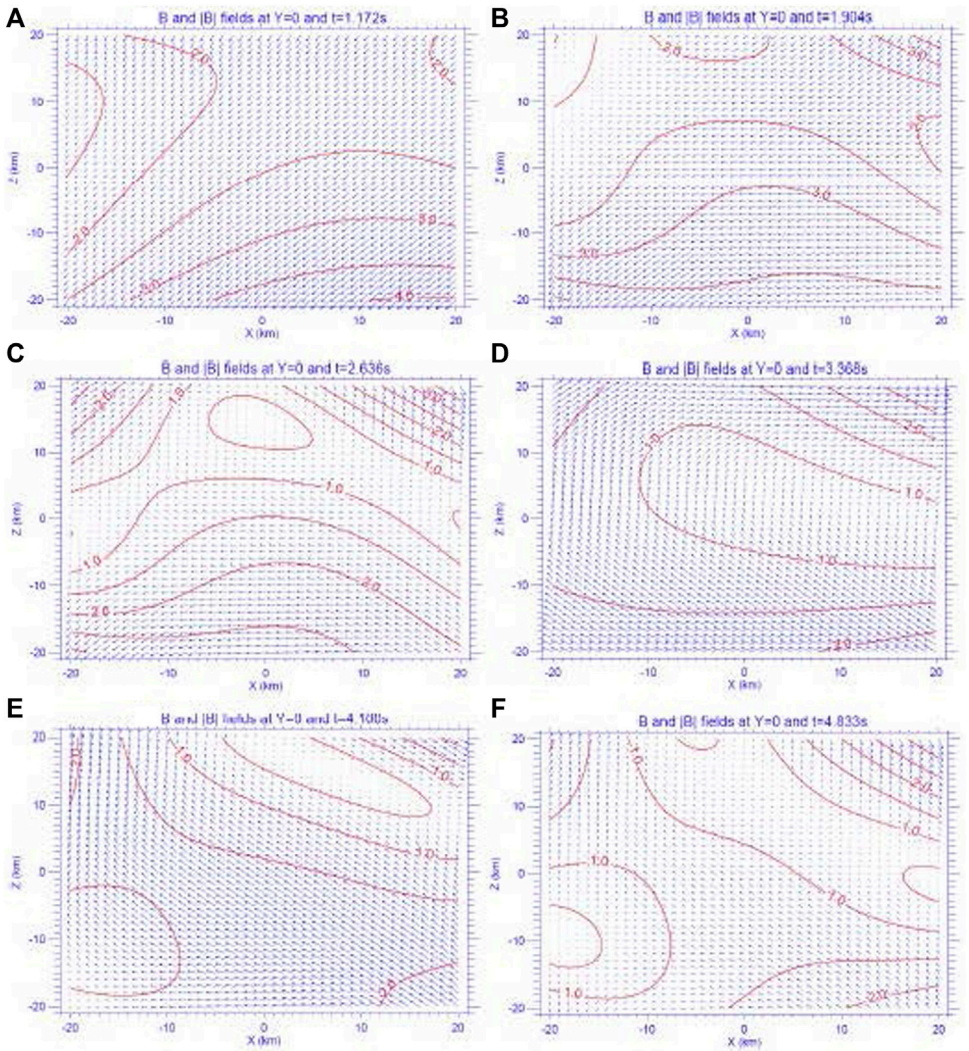

FIGURE 8 | Modeled **B** fields projected into and its magnitude |**B**| (in nT) evaluated on the *X*-*Z* plane of *Y* = 0 at six time instances of **(A)** *t* = 1.172 s, **(B)** *t* = 1.904 s, **(C)** *t* = 2.636 s, **(D)** *t* = 3.368 s, **(E)** *t* = 4.100 s, and **(F)** *t* = 4.833 s after 22:34 UT on 11 July 2017.

circumstance, the reconstructed fields at different time instances cannot be directly compared with each other. Our reconstruction model based on the SPSA stochastic optimization method can automatically accommodate an over-determined or under-determined setting of the model as discussed above. As a result, the fields reconstructed at different temporal instances but on the common, fixed GSE coordinate system can be directly compared.

We present the reconstructed fields in a local GSE coordinate *X*-*Y*-*Z* such that

$$(X, Y, Z) = \left(X', Y', Z'\right) - (X_0, Y_0, Z_0), \tag{11}$$

where $X' - Y' - Z'$ define the generic GSE coordinate system and $(X_0, Y_0, Z_0) = (-1.373 \times 10^5, 2.70 \times 10^4, 2.32 \times 10^4)$ km is determined by the satellite constellation, which corresponds to the GSE coordinate of the mean barycenter averaged over the measurement time shown in **Figure 1**. In **Figure 8**, we show the reconstructed **B** fields projected into and its magnitude $|\mathbf{B}|$ $(= \sqrt{B_1^2 + B_2^2 + B_3^2}$, in nT) evaluated on the *X*-*Z* plane of $Y = 0$ at six time instances of (a) $t = 1.172$ s, (b) $t = 1.904$ s, (c) $t = 2.636$ s, (d) $t = 3.368$ s, (e) $t = 4.100$ s, and (f) $t = 4.833$ s after 22:34 UT. The figure shows that both the magnetic configuration and the intensity of the magnetic field change noticeably with time. The reconnection region is characterized by a weak $|\mathbf{B}|$ and a reversal of the orientation (or a near anti-parallization) of the **B** vectors across the region. It is noted that a weak $|\mathbf{B}|$ also means a weak confinement to the motions of energetic electrons. This will lead to localized very fine-scale energy spectra and angular distributions that could be correlated with the remote magnetic topologies through the gyro-sounding process as revealed by the data from the Fly's Eye Energetic Particle Spectrometer (FEEPS) onboard the MMS spacecrafts (Cohen

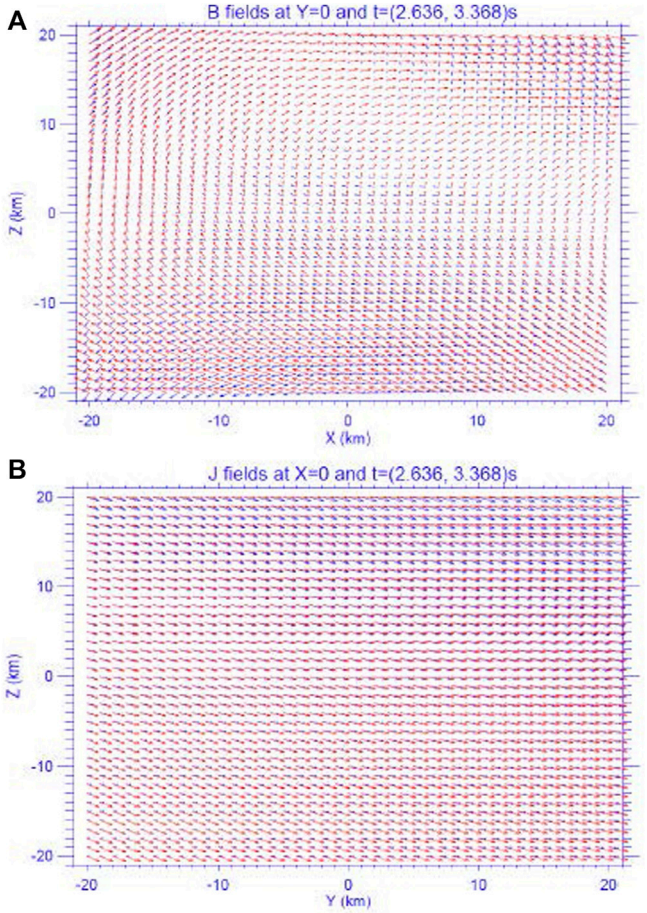

FIGURE 9 | (A) Modeled **B** fields projected into the *X-Z* plane of *Y* = 0 at two times of *t* = 2.636 s (blue) and *t* = 3.368 s (red) after 22:34 UT on 11 July 2017. **(B)** Modeled **J** fields projected into the *Y-Z* plane of *X* = 0 at the same two instances as in panel **(A)**. A significant intensification in **J** fields at *t* = 3.368 s as indicated by a plot with dominant red arrows.

et al., 2021; Turner et al., 2021). The development and evolvement of these two features can be easily identified in this figure. To provide a better view on the development of the reconnection region, we show in **Figure 9A** the superimposed **B** fields at two neighboring time instances of $t = 2.636$ s and $t = 3.368$ s on the same plot. The figure shows the development of an anti-parallel **B** field having a nearly opposite direction and an equal magnitude with a significantly weak **B** field sandwiched between the two regions at $t = 3.368$ s and especially in the region of $X > 0$. Though the reconstruction in the present model is under the *X-Y-Z* coordinate whereas the reconstruction in Torbert et al. (2020) was presented in the *L-M-N* coordinate, the configuration of the reconstructed **B**-field shown in **Figure 9A** is qualitatively similar to that shown in Torbert et al. (2020). **Figure 9B** shows the corresponding cross tail current **J** on the *Y-Z* plane of $X = 0$ that shows a significant intensification in its magnitude due to the development of the reconnection event.

CONCLUSION

A new ER model for the 3D magnetic field and plasma current field has been developed by use of a stochastic optimization method called SPSA. This reconstruction model adopts an empirical approach by fitting the prescribed analytic functions for the magnetic and plasma fields to the point-wise measurements from a constellation of satellites with a set of physical constraints determined by the MHD equations. The fitness is defined by a general loss function that consists of the model-measurement differences and the model departures from linear or nonlinear physical constraints. The new ER model directly minimizes the loss function using a stochastic optimization method called SPSA algorithm for which the effect of the random measurement errors is also included. We presented the concrete steps of how to implement this ER model to a special case of having the MMS-measured fields $(\hat{\mathbf{B}}, \hat{\mathbf{J}})$ combined with a set of physical constraints corresponding to an ideal MHD system of **Eqs. 1a**, b, which has been extensively investigated by traditional least-squares method (e.g., Denton et al., 2020; Torbert et al., 2020). Most SPSA applications contain the loss functions that only involve the difference between the modeled and measured quantities (e.g., Chin, 1999; Spall, 2003). On the other hand, the constraints contained in the generalized loss function (6) include not only the model-measurement differences but also the model departures derived from the physical constraints **Eqs. 1a, b**, which in turn characterizes the physical robustness of the fields reconstructed by an empirical model.

We have introduced the indices (r_B, r_J) in **Eq. 9** that calculate the relative differences between the modeled $(\mathbf{B}, \mathbf{J})$ fields and the measured $(\hat{\mathbf{B}}, \hat{\mathbf{J}})$ fields. This set of indices (γ_B, γ_J) provides an objective measure of the accuracy to the modeled fields. In addition, the concept of the quality indicator Q_{curl} introduced in Dunlop et al. (1988) has been extended to a new model quality indicator Q_{model} shown in **Eq. 10**. This index provides an objective measure to the robustness of the modeled field in terms of its physical property of $\nabla \cdot \mathbf{B} = 0$. These two sets of new indices are respectively associated with the two sets of constraints of model-measurement differences and the model departures used in designing the general loss function for the new ER model. The new ER model was applied to the measurements of an EDR observed by the MMS mission (Torbert et al., 2018). By conducting various sensitivity investigations of the reconstruction model, we were able to examine the sources of the errors in the reconstructed fields previously noted by the curlometer technique. It is now found that the errors in the plasma current density calculated directly from the measured magnetic fields based on curlometer technique were mostly contributed from the linear approximation to a nonlinear configuration of the 3D magnetic fields. A more comprehensive nonlinear ER model that uses **Eqs. 1–3** with point-wise measurements of (**B**, **J**) and (**U**, **E**) fields and effectively includes the effects of plasma resistivity contained in **Eqs. 2**, **3** near the EDRs will be presented in our future investigations.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

XZ: Proposed the original idea, developed the model, and wrote the paper. IC: Provided the data, refined the idea, and revised the paper. BM: Refined the idea and revised the paper. RN: Refined the idea and revised the paper. DT: Refined the idea through various discussions. RT: Refined the idea.

FUNDING

This research was supported by the Magnetospheric Multiscale (MMS) mission of NASA's Science Directorate Heliophysics Division *via* subcontract to the Southwest Research Institute (NNG04EB99C) and NASA Grant NNX10AB84G.

ACKNOWLEDGMENTS

Constructive comments on SPSA applications from James C. Spall and Richard Denton are greatly appreciated. We also thank Daniel J. Gershman and others on MMS team for the FPI current density and magnetic field measurements used in this paper. Constructive comments from two reviewers are also appreciated.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fspas.2022.878403/full#supplementary-material

REFERENCES

Axelsson, O. (1996). *Iterative Solution Methods*. Cambridge: Cambridge University Press, 654.

Bhatnagar, S., Prasad, H. L., and Prashanth, L. A. (2013). *Stochastic Recursive Algorithms for Optimization - Simultaneous Perturbation Method*. New York: Springer, 302.

Boyd, T. J., and Sanderson, J. J. (2003). *The Physics of Plasmas*. Cambridge: Cambridge University Press, 532.

Burch, J. L., Moore, T. E., Torbert, R. B., and Giles, B. L. (2016). Magnetospheric Multiscale Overview and Science Objectives. *Space Sci. Rev.* 199, 5–21. doi:10.1007/s11214-015-0164-9

Chanteur, G. (1998). "Spatial Interpolation for Four Spacecraft: Theory," in *Analysis Methods for Multi-Spacecraft Data*. Editors G. Paschmann and P. W. Daly (Bern, Switzerl, and Paris, France: The International Space Science InstituteEuropean Space Agency), 395–418. Chap. 12.

Chin, D. C. (1999). Simultaneous Perturbation Method for Processing Magnetospheric Images. *Opt. Eng.* 38, 606–611. doi:10.1117/1.602104

Chong, E. K. P., and Zak, S. H. (2001). *An Introduction to Optimization*. Second Edition. New York: John Wiley & Sons, 476.

Cohen, I. J., Turner, D. L., Mauk, B. H., Bingham, S. T., Blake, J. B., Fennell, J. F., et al. (2021). Characteristics of Energetic Electrons Near Active Magnetotail Reconnection Sites: Statistical Evidence for Local Energization. *Geophys. Res. Lett.* 48, e2020GL090087. doi:10.1029/2020GL090087

Denton, R. E., Torbert, R. B., Hasegawa, H., Dors, I., Genestreti, K. J., Argall, M. R., et al. (2020). Polynomial Reconstruction of the Reconnection Magnetic Field Observed by Multiple Spacecraft. *J. Geophys. Res. Space Phys.* 125, e2019JA027481. doi:10.1029/2019JA027481

Dunlop, M. W., Balogh, A., Glassmeier, K.-H., and Robert, P. (2002). Four-point Cluster Application of Magnetic Field Analysis Tools: The Curlometer. *J. Geophys. Res.* 107 (A11), 1384. doi:10.1029/2001JA005088

Dunlop, M. W., Southwood, D. J., Glassmeier, K.-H., and Neubauer, F. M. (1988). Analysis of Multipoint Magnetometer Data. *Adv. Space Res.* 8, 273–277. doi:10.1016/0273-1177(88)90141-x

Gurnett, D. A., and Bhattacharjee, A. (2005). *Introduction to Plasma Physics – with Space and Laboratory Applications*. Cambridge: Cambridge University Press, 452.

Harvey, C. C. (1998). "Spatial Gradients and the Volumetric Tensor," in *Analysis Methods for Multi-Spacecraft Data*. Editors G. Paschmann and P. W. Daly (Bern, Switzerl, and Paris, France: The International Space Science InstituteEuropean Space Agency), 307–322. Chap. 12.

Hasegawa, H., Sonnerup, B. U. Ö., Dunlop, M. W., Balogh, A., Haaland, S. E., Klecker, B., et al. (2004). Reconstruction of Two-Dimensional Magnetopause Structures from Cluster Observations: Verification of Method. *Ann. Geophys.* 22, 1251–1266. doi:10.5194/angeo-22-1251-2004

Hasegawa, H., Sonnerup, B. U. Ö., Klecker, B., Paschmann, G., Dunlop, M. W., and Rème, H. (2005). Optimal Reconstruction of Magnetopause Structures from Cluster Data. *Ann. Geophys.* 23, 973–982. doi:10.5194/angeo-23-973-2005

Holton, J. R. (2004). *An Introduction to Dynamic Meteorology*. Fourth Edition. New York: Elsevier Academic Press, 535.

Menke, W. (1989). *Geophysical Data Analysis: Discrete Inverse Theory*. Revised Edition. New York: Academic Press, 289.

Middleton, H., and Masson, A. (2016). *The Curlometer Technique: A Beginner's Guide*. CSA Technical Note. ESDC-CSA-TN-0001. Madrid, Spain: European Space Astronomy Centre, 19.

Pollock, C., Moore, T., Jacques, A., Burch, J., Gliese, U., Saito, Y., et al. (2016). Fast Plasma Investigation for Magnetospheric Multiscale. *Space Sci. Rev.* 199, 331–406. doi:10.1007/s11214-016-0245-4

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in Fortran. The Arts of Scientific Computing*. Second Edition. Cambridge: Cambridge University Press, 963.

Priest, E. (2016). "MHD Structures in Three-Dimensional Reconnection," in *Magnetic Reconnection – Concepts and Applications*. Editors W. Gonzalez and E. Parker (Switzerland: Springer), 101–142. doi:10.1007/978-3-319-26432-5_3

Robert, P., Roux, A., Harvey, C. C., Dunlop, M. W., Daly, P. W., and Glassmeier, K. H. (1998). "Tetrahedron Geometric Factors," in *Analysis Methods for Multi-Spacecraft Data*. Editors G. Paschmann and P. W. Daly (Bern, Switzerl, and Paris, France: The International Space Science InstituteEuropean Space Agency), 323–348. Chap. 13.

Roelof, E. C., Mauk, B. H., and Meier, R. R. (1993). Simulations of EUV and ENA Magnetospheric Images Based on the Rice Convection Model. *Proc. Spie, Instrumentation Magnetospheric Imagery* 2008, 202–213.

Russell, C. T., Anderson, B. J., Baumjohann, W., Bromund, K. R., Dearborn, D., Fischer, D., et al. (2016). The Magnetospheric Multiscale Magnetometers. *Space Sci. Rev.* 199, 189–256. doi:10.1007/s11214-014-0057-3

Scudder, J. D. (2016). "Collisionless Reconnection and Electron Demagnetization," in *Magnetic Reconnection – Concepts and Applications*. Editors W. Gonzalez and E. Parker (Switzerland: Springer), 33–100. doi:10.1007/978-3-319-26432-5_2

Sonnerup, B. U. Ö., and Guo, M. (1996). Magnetopause Transects. *Geophys. Res. Lett.* 23, 3679–3682. doi:10.1029/96gl03573

Sonnerup, B. U. Ö., Hasegawa, H., Denton, R. E., and Nakamura, T. K. M. (2016). Reconstruction of the Electron Diffusion Region. *J. Geophys. Res. Space Physics* 121, 4279–4290. doi:10.1002/2016JA022430

Sonnerup, B. U. Ö., and Teh, W.-L. (2008). Reconstruction of Two-Dimensional Coherent MHD Structures in a Space Plasma: The Theory. *J. Geophys. Res.* 113. doi:10.1029/2007JA012718

Spall, J. C. (2000). Adaptive Stochastic Approximation by the Simultaneous Perturbation Method. *IEEE Trans. Automat. Contr.* 45, 1839–1853. doi:10.1109/tac.2000.880982

Spall, J. C. (1998a). An Overview of the Simultaneous Perturbation Method for Efficient Optimization. *Johns Hopkins APL Tech. Dig.* 19, 482–492.

Spall, J. C. (1998b). Implementation of the Simultaneous Perturbation Algorithm for Stochastic Optimization. *IEEE Trans. Aerosp. Electron. Syst.* 34 (3), 817–823. doi:10.1109/7.705889

Spall, J. C. (2003). *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. New York: John Wiley & Sons, 595.

Spall, J. C. (1992). Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation. *IEEE Trans. Automat. Contr.* 37, 332–341. doi:10.1109/9.119632

Sturrock, P. A. (1994). *Plasma Physics: An Introduction to the Theory of Astrophysical, Geophysical and Laboratory Plasmas*. Cambridge: Cambridge University Press, 335.

Torbert, R. B., Burch, J. L., Phan, T. D., Hesse, M., Argall, M. R., Shuster, J., et al. (2018). Electron-scale Dynamics of the Diffusion Region during Symmetric Magnetic Reconnection in Space. *Science* 362 (6421), 1391–1395. doi:10.1126/science.aat2998

Torbert, R. B., Dors, I., Argall, M. R., Genestreti, K. J., Burch, J. L., Farrugia, C. J., et al. (2020). A New Method of 3-D Magnetic Field Reconstruction. *Geophys. Res. Lett.* 47, e2019GL085542. doi:10.1029/2019gl085542

Tsyganenko, N. A., and Sitnov, M. I. (2007). Magnetospheric Configurations from a High-Resolution Data-Based Magnetic Field Model. *J. Geophys. Res.* 112. doi:10.1029/2007JA012260

Turner, D. L., Cohen, I. J., Bingham, S. T., Stephens, G. K., Sitnov, M. I., Mauk, B. H., et al. (2021). Characteristics of Energetic Electrons Near Active Magnetotail Reconnection Sites: Tracers of a Complex Magnetic Topology and Evidence of Localized Acceleration. *Geophys. Res. Lett.* 48, e2020GL090089. doi:10.1029/2020GL090089

Yamada, M., Yoo, J., and Zenitani, S. (2016). "Energy Conversion and Inventory of a Prototypical Magnetic Reconnection Layer," in *Magnetic Reconnection – Concepts and Applications*. Editors W. Gonzalez and E. Parker (Switzerland: Springer), 143–179. doi:10.1007/978-3-319-26432-5_4

Zhu, X., and Lui, A. T. Y. (2012). Reconstruction of Neighboring Plasma Environment along a Satellite Path by a Barotropic Plasma Model. *J. Atmos. Solar-Terrestrial Phys.* 77, 46–56. doi:10.1016/j.jastp.2011.11.005

Zhu, X., and Spall, J. C. (2002). A Modified Second-Order SPSA Optimization Algorithm for Finite Samples. *Int. J. Adapt. Control. Signal. Process.* 16, 397–409. doi:10.1002/acs.715

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Astronomy and Space Sciences

ORIGINAL RESEARCH
published: 10 May 2022
doi: 10.3389/fspas.2022.867947

Domain-Agnostic Outlier Ranking Algorithms—A Configurable Pipeline for Facilitating Outlier Detection in Scientific Datasets

Hannah R. Kerner[1], Umaa Rebbapragada[2], Kiri L. Wagstaff[2], Steven Lu[2], Bryce Dubayah[1], Eric Huff[2], Jake Lee[2], Vinay Raman[3] and Sakshum Kulshrestha[1]*

[1]University of Maryland, College Park, Maryland, MD, United States, [2]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, United States, [3]Montgomery Blair High School, Silver Spring, MD, United States

Automatic detection of outliers is universally needed when working with scientific datasets, e.g., for cleaning datasets or flagging novel samples to guide instrument acquisition or scientific analysis. We present Domain-agnostic Outlier Ranking Algorithms (DORA), a configurable pipeline that facilitates application and evaluation of outlier detection methods in a variety of domains. DORA allows users to configure experiments by specifying the location of their dataset(s), the input data type, feature extraction methods, and which algorithms should be applied. DORA supports image, raster, time series, or feature vector input data types and outlier detection methods that include Isolation Forest, DEMUD, PCA, RX detector, Local RX, negative sampling, and probabilistic autoencoder. Each algorithm assigns an outlier score to each data sample. DORA provides results interpretation modules to help users process the results, including sorting samples by outlier score, evaluating the fraction of known outliers in *n* selections, clustering groups of similar outliers together, and web visualization. We demonstrated how DORA facilitates application, evaluation, and interpretation of outlier detection methods by performing experiments for three real-world datasets from Earth science, planetary science, and astrophysics, as well as one benchmark dataset (MNIST/Fashion-MNIST). We found that no single algorithm performed best across all datasets, underscoring the need for a tool that enables comparison of multiple algorithms.

Keywords: astrophysics, planetary science, Earth Science, outlier detection, novelty detection, out-of-distribution detection

OPEN ACCESS

Edited by:
Bala Poduval, University of New Hampshire, United States

Reviewed by:
Chathurika S.Wickramasinghe, Virginia Commonwealth University, United States
Jérôme Saracco, Institut Polytechnique de Bordeaux, France

****Correspondence:***
Hannah R. Kerner
hkerner@umd.edu

Specialty section:
This article was submitted to Astrostatistics, a section of the journal Frontiers in Astronomy and Space Sciences

Received: *01 February 2022*
Accepted: *04 April 2022*
Published: *10 May 2022*

Citation:
Kerner HR, Rebbapragada U, Wagstaff KL, Lu S, Dubayah B, Huff E, Lee J, Raman V and Kulshrestha S (2022) Domain-Agnostic Outlier Ranking Algorithms—A Configurable Pipeline for Facilitating Outlier Detection in Scientific Datasets. Front. Astron. Space Sci. 9:867947. doi: 10.3389/fspas.2022.867947

1 INTRODUCTION

The ability to automatically detect out-of-distribution samples in large data sets is of interest for a wide variety of scientific domains. Depending on the application setting, this capability is also commonly referred to as anomaly detection, outlier detection, or novelty detection. More broadly, this is referred to as out-of-distribution (OOD) detection. In general, the goal of OOD detection systems is to identify samples that deviate from the majority of samples in a dataset in an unsupervised manner (Pimentel et al., 2014). In machine learning, these methods are commonly used for identifying mislabeled or otherwise invalid samples in a dataset (Liang et al., 2018; Böhm and Seljak, 2020). When working with science datasets, OOD detection can be used for

cleaning datasets, e.g., flagging ground-truth labels with GPS or human entry error or identifying wrongly categorized objects in a catalog (Wagstaff et al., 2020a; Lochner and Bassett, 2021). It could also be used for discovery, e.g., to flag novel samples in order to guide instrument acquisition or scientific analysis (Wagstaff et al., 2013; Kerner et al., 2020a; Kerner et al., 2020b; Wagstaff et al., 2020b). Another application is the detection of rare objects that are known to exist but the known examples are too few to create a large enough labeled dataset for supervised classification algorithms (Chein-I Chang and Shao-Shan Chiang, 2002; Zhou et al., 2016).

Despite wide differences in applications, data types, and dimensionality, the same underlying machine learning algorithms can be employed across all of these domains. A challenge for applying them however is that domain scientists do not always have the programming or machine learning background to apply the algorithms themselves using existing tools. Given the widespread applicability and transferability of OOD methods, the scientific community would benefit from a tool that made it easy for them to apply popular outlier detection algorithms to their science datasets. We created DORA (Domain-agnostic Outlier Ranking Algorithms) to provide a tool for applying outlier detection algorithms to a variety of scientific data sets with minimal coding required. Users need only to specify details for their data/application including the data type, location, and algorithms to run in an experiment configuration file. DORA supports image, raster, time series, or feature vector input data types and outlier detection methods that include Isolation Forest, Discovery via Eigenbasis Modeling of Uninteresting Data (DEMUD) (Wagstaff et al., 2013), principal component analysis (PCA), Reed-Xiaoli (RX) detector (Reed and Yu, 1990), Local RX, negative sampling (Sipple, 2020), and probabilistic autoencoder (PAE). Each algorithm assigns an outlier score to each sample in a given dataset. DORA provides results organization and visualization modules to help users process the results, including sorting samples by outlier score, evaluating outlier recall for a set of known/labeled outliers, clustering groups of similar outliers together, and web visualization. We demonstrated how DORA facilitates application, evaluation, and interpretation of outlier detection methods by performing experiments for three real-world datasets from Earth science, planetary science, and astrophysics, as well as one benchmark dataset (MNIST/Fashion-MNIST).

The key contributions of this paper are:

- A new pipeline, DORA, for performing outlier detection experiments using several AI algorithms that reduces the effort and expertise required for performing experiments and comparing results from multiple algorithms
- Using experiments for a diverse set of real world datasets and application areas, we show that no single algorithm performs best for all datasets and use cases, underscoring the need for a tool that compares multiple algorithms
- We provide publicly available code for running and contributing to the DORA pipeline and datasets that can be used for reproducing experiments or benchmarking outlier detection methods

2 RELATED WORK

Methods for outlier detection have been surveyed extensively and can be differentiated primarily based on how they score outliers (Markou and Singh, 2003a; Markou and Singh, 2003b; Chandola et al., 2009; Pimentel et al., 2014). Reconstruction-based methods construct a model of a dataset by learning a mapping between the input data and a lower-dimensional representation that minimizes the loss between the input and its reconstruction from the low-dimensional representation (Kerner et al., 2020a). The reconstruction error is used as the outlier score because samples that are unlike the data used to fit the model will be more poorly reconstructed compared to inliers. Reconstruction-based methods include PCA (Jablonski et al., 2015), autoencoders (Richter and Roy, 2017), and generative adversarial networks (Akcay et al., 2018). Distance-based methods score outliers based on their distance from a "background" which can be defined in a variety of ways. For example, the Reed-Xiaoli (RX) detector computes the Mahalanobis distance between each sample and the background dataset defined by its mean and covariance matrix (Reed and Yu, 1990). Sparsity-based methods such as isolation forest (Liu et al., 2008) and local outlier factor (Breunig et al., 2000) score outliers based on how isolated or sparse samples are in a given feature space. Probability distribution and density based methods estimate the underlying distribution or probability density of a dataset and score samples using likelihood. Examples include the probabilistic autoencoder, which scores samples based on the log likelihood under the latent space distribution (Böhm and Seljak, 2020), Gaussian mixture Models, and kernel density estimators (Chandola et al., 2009). Other methods formulate outlier detection as supervised classification, usually with only one class constituted by known normal samples. Such methods include one-class support vector machines (Schölkopf et al., 1999) and negative sampling (Sipple, 2020).

In astrophysics, outlier detection methods have been used to identify astrophysical objects with unique characteristics (Hayat et al., 2021) as well as data or modeling artifacts in astronomical surveys (Wagstaff et al., 2020a; Lochner and Bassett, 2021). Example outlier detection applications in Earth science include detecting anomalous objects or materials (Zhou et al., 2016), data artifacts or noise (Liu et al., 2017), change (Touati et al., 2020), and ocean extremes (Prochaska et al., 2021). Planetary science applications have mostly focused on prioritizing samples with novel geologic or geochemical features for follow-up targeting or analysis (Wagstaff et al., 2013; Kerner et al., 2020a). These examples show the benefit of applying outlier detection methods in a variety of real-world science use cases. However, the effort required to apply and evaluate the many available algorithms is non-trivial and can be daunting for non-ML experts, thus impeding the uptake of outlier detection methods in science applications. There is a need for tools that make it easier for domain scientists to apply outlier detection methods as well as compare results across datasets. While there have been some efforts to develop tools for facilitating the application of outlier

detection methods (Zhao et al., 2019), they cover limited data formats and algorithms. DORA aims to fill the need for tools that facilitate application, evaluation, and interpretation of outlier detection methods.

3 METHODS

Figure 1 illustrates the architecture of DORA including data loading, feature extraction, outlier ranking, and results organization and visualization modules. In order to improve the readability and execution speed of the code, we adopted object-oriented and functional programming practices. We designed DORA to be readily extensible to support additional data types or formats, outlier detection algorithms, and results organization or visualization methods by writing new modules that follow the DORA API. Experimental settings are controlled by configuration files in which users can specify the input data, feature extraction methods, normalization method, outlier ranking methods, and results organization methods. DORA is implemented in Python 3.

3.1 Data Loaders

We chose to implement data loaders for four data types that are commonly used by the machine learning and domain science communities: time series, feature vectors, images (grayscale or RGB), and *N*-band rasters. *N*-band rasters are images or grids in which every pixel is associated with a location (e.g., latitude/longitude in degrees); most satellite data are distributed as rasters. A data loader for each data type locates the data by the path(s) defined in the configuration file and loads samples into a dictionary of numpy arrays indexed by the sample id. This data_dict is then passed to each of the ranking algorithms.

3.2 Outlier Ranking Algorithms

We implemented seven unsupervised algorithms for scoring and ranking samples by outlierness. We chose these algorithms to include a diverse set of approaches to scoring outliers since different algorithms may perform better for different datasets and use cases. We describe each approach to scoring outliers and the associated methods below.

3.2.1 Reconstruction Error

Principal component analysis (PCA) has been used for outlier detection by scoring samples using the reconstruction error (here, the L2 norm) between inputs and their inverse transformation from the principal subspace (Kerner et al., 2020a). DEMUD (Wagstaff et al., 2013) differs from other outlier ranking methods: instead of independently scoring all observations, DEMUD incrementally identifies the most unusual remaining item, then incorporates it into the model of "known" (non-outlier) observations before selecting the next most unusual item. DEMUD's goal is to identify diverse outliers and avoid redundant selections. Once an outlier is found, repeated occurrences of that outlier are deprioritized. Methods that score samples independently maximize coverage of outliers, while DEMUD maximizes fast discovery of distinct outlier types.

3.2.2 Distance

The Reed-Xiaoli (RX) detector is commonly used for anomaly detection in multispectral and hyperspectral remote sensing. RX scores samples using the Mahalanobis distance between a sample and a background mean and covariance (Reed and Yu, 1990). The local variant of RX (Local RX or LRX) can be used for image or raster data and scores each pixel in an image with respect to a window "ring" of pixels surrounding it (Molero et al., 2013). LRX requires two parameters to define the size of the outer window

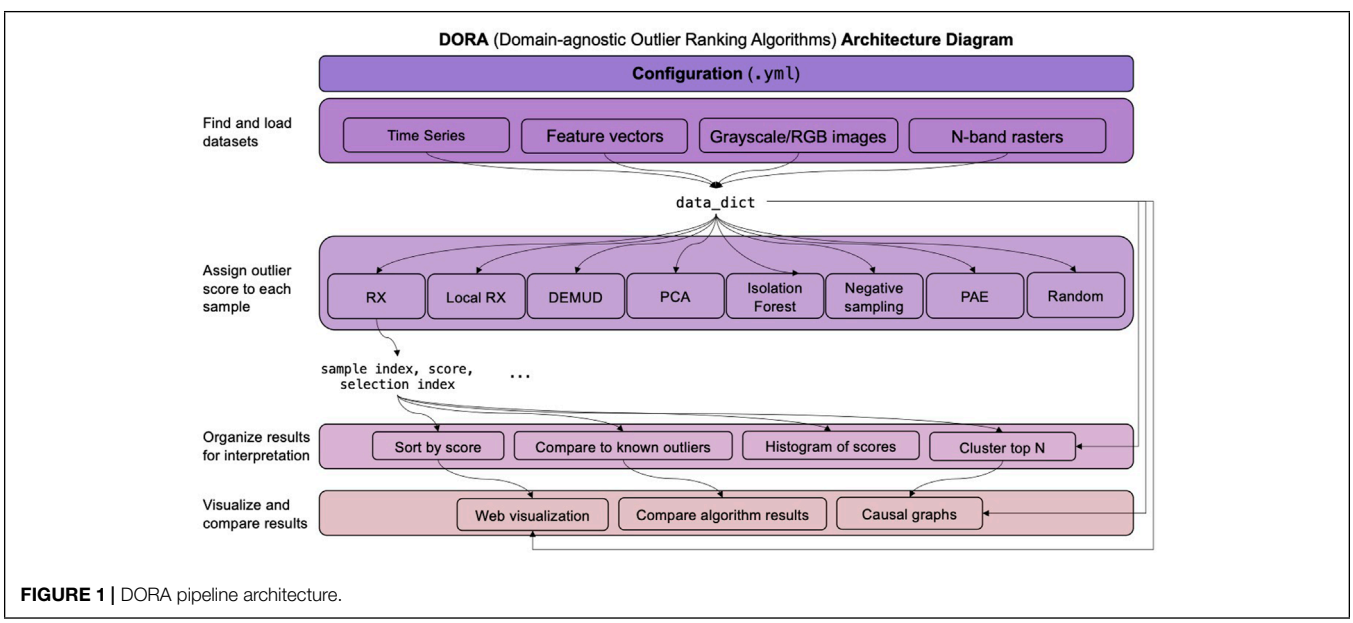

FIGURE 1 | DORA pipeline architecture.

surrounding the pixel and the inner window around the target pixel to exclude from the background distribution.

3.2.3 Sparsity

Isolation forest (iForest) is a common sparsity-based method that constructs many random binary trees from a dataset (Liu et al., 2008). The outlier score for a sample is quantified as the average distance from the root to the item's leaf. Shorter distances are indicative of outliers because the number of random splits required to isolate the sample is small.

3.2.4 Probability

The negative sampling algorithm is implemented by converting the unsupervised outlier ranking problem into a semi-supervised problem (Sipple, 2020). Negative (anomalous) examples are created by sampling from an expanded space defined by the minimum and maximum values of each dimension of the positive (normal) examples. The negative and positive examples are then used to train a random forest classifier. We use the posterior probabilities of the random forest classifier as outlier scores, which means that the observations with higher posterior probabilities are more likely to be outliers. The probabilistic autoencoder is a generative model consisting of an autoencoder trained to reconstruct input data which is interpreted probabilistically after training using a normalizing flow on the autoencoder latent space (Böhm and Seljak, 2020). Samples are scored as outliers using the log likelihood in the latent distribution, the autoencoder reconstruction error, or a combination of both.

3.3 Results Interpretation

Each of the outlier ranking algorithms returns an array containing the sample index, outlier score, and selection index (index after sorting by outlier score). DORA provides organization and visualization modules intended to help users interpret and make decisions based on these outputs. The simplest module saves a CSV of the samples sorted by their outlier score (i.e., selection order). Clustering the top N outlier selections can enable users to investigate the different types of outliers that might be present in the dataset; this could be especially useful for separating outliers caused by noise or data artifacts vs scientifically interesting samples. We implemented the K-means and self-organizing maps (SOMs) algorithms for clustering the top-N outliers. For use cases in which an evaluation dataset containing known outliers is available, we provide a module to assess how well algorithm selections correlate with known outliers. This is done by plotting the number of known outliers vs number of selections made. We provide a module for plotting histograms of outlier scores to visualize the distribution of scores in the dataset (which may be, e.g., multimodal or long-tailed). We developed a desktop application to easily visualize DORA results with the Electron application framework and React frontend library. This enables fast and easy comparison of the results from different methods. We developed a desktop application to easily visualize DORA results with the Electron application framework and React frontend library. The application loads the DORA configuration file to locate the dataset and result CSVs. Then, it displays the ranked samples and their scores in a table sorted by their selection order. This allows for fast and easy comparison of the results of different methods. **Figure 2** shows a screenshot of the "Aggregate Table" view, which displays all results from different algorithms side-by-side.

4 DATASETS

We constructed three datasets to evaluate the utility of DORA and algorithm performance for a variety of scientific domains (astrophysics, planetary science, and Earth science). We also included a benchmark dataset that uses MNIST and Fashion-MNIST. **Table 1** summarizes the number of unlabeled samples used for training and evaluation for each dataset. We describe each dataset in detail below.

4.1 Astrophysics: Objects in Dark Energy Survey

Astronomical data sets are large and growing. Large modern optical imaging surveys are producing catalogs of order 10^8 stars and galaxies, with dozens or hundreds of distinct measured features for each entry. Discovery science becomes difficult at this data volume: the scale is too large for expert human inspection, and separating real astrophysical anomalies from non-astrophysical sources like detector artifacts or satellite trails is a challenging problem for current methods.

The Dark Energy Survey (DES) is an ongoing imaging survey of $5,000$ deg^2 of the southern sky from the Cerro-Tololo Inter-American Observatory in Chile (Zuntz et al., 2018). The resulting galaxy catalogs produced have provided some of the strongest constraints to date on the physical properties of dark energy and accelerated expansion of the Universe. The first version of this catalog, released June 2018, incorporated only cuts on signal-to-noise and resolution, masks against known detector anomalies and data quality indicators, and the automated data quality flags produced during processing to filter outliers. In December 2019, the full catalog was released after 18 months of extensive manual vetting. We used the samples that were removed in the second version of the catalog as a set of known outliers for evaluating anomaly detection methods on the first version.

We compared all methods on a dataset of 100K galaxy objects observed by the Dark Energy Survey (DES) sampled from the initial June 2018 release. We labeled the 25,339 objects from this 100 K set that did not appear in the later December 2019 release, thus were likely eliminated during the manual vetting process, as outliers. While the remaining 74,661 objects may also contain outliers, we assume them to be inliers in this experiment. We used publicly-available photometry from the *g*–, *r*–, *i*– and *z*– band DES exposures. We transformed the photometry into luptitudes[1]. The input features were the *r*-band

[1]A "Luptitude" (Lupton et al., 1999) is an arcsinh-scaled flux, with properties quantitatively equal to traditional astronomical magnitudes for bright sources, but which gracefully handles non-detections and negative fluxes.

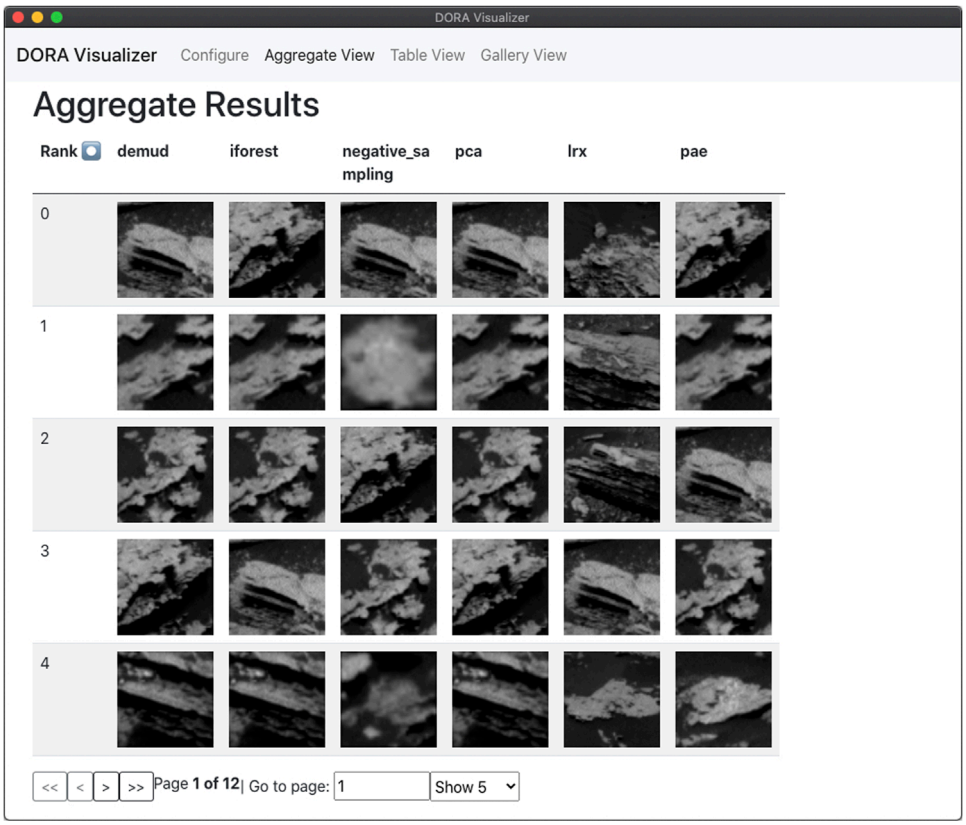

FIGURE 2 | A screenshot of the DORA visualizer displaying results from the planetary science dataset.

luptitude, colors computed as band differences between $g - r$, $i - r$, and $z - r$, and associated observational errors, for a total of eight features.

4.2 Planetary: Targets in Mars Rover Images

Mars exploration is fundamentally an exercise in discovery with the goal of increasing our understanding of Mars's history, evolution, composition, and currently active processes. Outliers identified in Mars observations can inspire new discoveries and inform the choice of which areas merit follow-up or deeper investigation (Kerner et al., 2020a; Wagstaff et al., 2020b). We collected 72 images from the Navigation camera (Navcam) on the Mars Science Laboratory (MSL) rover and employed Rockster (Burl et al., 2016) (currently used by onboard rover software) to identify candidate rock targets with an area of at least 100 pixels, yielding 1,050 targets. We cropped out a 64 × 64 pixel image centered on each target.

We simulated the operational setting in which the rover has observed targets up through mission day (sol) s and the goal is to rank all subsequent targets (after sol s) to inform which recent targets merit further study. Our rover image data covers sols 1,343 to 1703. We partitioned the images chronologically to assess outlier detection in the 10 most recent sols, using "prior" set $D_{1343-1693}$ ($n = 992$) and "assessment" set $D_{1694-1703}$ ($n = 58$) for evaluation. We collaborated with an MSL science team member to independently review the targets in $D_{1694-1703}$ and identify those considered novel by the mission ($n_{outlier} = 9$). Our goal for this application is to assess how well the selections made by each algorithm correlate with human novelty judgments to determine which methods would be most suitable for informing onboard decisions about follow-up observations.

TABLE 1 | Number of samples in the training and test sets for each dataset.

Dataset	Training Unlabeled	Test Outliers	Inliers
Astrophysics	100,000	25,339	74,661
Planetary	992	9	49
Earth	6,757	37	76
F-MNIST	60,000	1,000	1,000

4.3 Earth: Satellite Time Series for Ground Observations

Many Earth science applications using satellite Earth observation (EO) data require ground-truth observations for identifying and modeling ground-identified objects in the satellite observations. These ground observations also serve as labels that are paired with satellite data inputs for machine learning models. For example,

a model trained to classify crop types in satellite observations requires ground-annotated labels of crop type. A widespread challenge for ground-annotated labels is that there are often points with erroneous location or label information (e.g., due to GPS location error or human entry error) that need to be cleaned before the labels can be used for machine learning or other downstream uses. Automatically detecting these outliers could save substantial time required for cleaning datasets and improve the performance of downstream analyses that rely on high-quality datasets.

We used a dataset of ground annotations of maize crops collected by the UN Food and Agriculture Organization (FAO). This dataset includes 6,757 samples with location (latitude/longitude) metadata primarily in Africa and Southeast Asia. Most locations coincide with crop fields but there are many outliers that coincide with other land cover types such as water, buildings, or forests. We constructed an evaluation set of all samples in Kenya ($n = 113$) and manually annotated whether each sample was located in a crop field (inlier) or not (outlier) using high-resolution satellite images in Collect Earth Online ($n_{inlier} = 76$, $n_{outlier} = 37$). We used the Sentinel-1 synthetic aperture radar (SAR) monthly median time series for each sample location from the year the sample was collected. We used SAR data because it is sensitive to ground texture and penetrates clouds, which is important for the often-cloudy region covered by the dataset. Our goal for this application was to assess how well the selections made by each algorithm correlate with outliers determined by visual inspection of the satellite images.

4.4 Benchmark: MNIST and Fashion-MNIST

We used MNIST and Fashion-MNIST (F-MNIST) to demonstrate DORA with a traditional benchmark dataset. We used 60,000 images from F-MNIST as the training set and a test set of 1,000 images each from MNIST (outliers) and F-MNIST (inliers).

5 RESULTS

The experimental setup for each dataset was to fit or train a model for each ranking algorithm using a larger unlabeled dataset and then apply the models to compute the outlier scores for a smaller test dataset for which labels of known outliers were available (**Table 1**). For each test set, we created a plot of the number of known outliers detected out of the top N selections. We also reported the Mean Discovery Rate (MDR) in the legend for each algorithm to give a quantitative comparison across the datasets. We defined MDR as:

$$MDR = \frac{\sum_{i=1}^{N_s} n_i}{\sum_{i=1}^{N_s} s_i} \tag{1}$$

where $i \in [1, N_s]$ is the selection index, N_s is the total number of selections, s_i is the number of selections made up to index i, and n_i is the number of known outliers (true positives) among s_i selections. We also reported the precision at $N = n_{outlier}$ for each test set where $n_{outlier}$ is the number of known outliers, i.e., the precision obtained when the number of selections is the same as the total number of outliers. Precision at N is the number of known outliers divided by the number of selections N (Campos et al., 2016). **Table 2** compares the precision at $N = n_{outlier}$ for each dataset and ranking algorithm. We calculated a random selection baseline which we refer to as "Theoretical Random" using the expected value of n_i for i random selections:

$$\mathbf{E}\left[n_i, i \in [1, N_s]\right] = \frac{\sum_{j=0}^{i} \binom{n_{outlier}}{j} \binom{D - n_{outlier}}{i - j} j}{\binom{D}{i}} \tag{2}$$

$$= \frac{n_{outlier} i}{D} \tag{3}$$

For the astrophysics dataset (**Figure 3A**), DEMUD was omitted due to computational time and LRX was omitted as it applies only to image data. Of the remaining methods, PCA achieved the highest precision, followed by RX. Negative sampling performs well initially before its performance drops off. The PAE finds the most outliers overall.

For the planetary dataset, we found that the Isolation Forest achieved the highest precision (best outlier detection) when allowed to select only 9 images. **Figure 3B** shows the complete (cumulative) outlier detection performance for each algorithm when ranking all 58 target images in $D_{1694–1703}$. We could not employ RX since the data dimensionality ($64 \times 64 = 4,096$) exceeded the data set size.

For the Earth dataset, negative sampling had the best performance in both metrics. DEMUD, PCA, and PAE tied for the lowest precision at $N = n_{outlier}$ while DEMUD and PCA tied for the lowest MDR (**Figure 3C**). We did not evaluate LRX for this time series dataset because LRX can only be applied to gridded image or raster data types.

For the MNIST and F-MNIST dataset, PCA and DEMUD tied for the highest precision at $N = n_{outlier}$ while DEMUD, PCA, and PAE tied for the highest MDR (**Figure 3D**). Negative sampling had the lowest performance in both metrics.

TABLE 2 | Precision at $N = n_{outlier}$ for four datasets; the best result for each data set is in bold.

Algorithm	Astro	Planetary	Earth	F-MNIST
PCA	**0.42**	0.44	0.41	**0.84**
DEMUD	—	0.44	0.41	**0.84**
RX	0.40	—	0.43	0.82
LRX	—	0.33	—	0.56
IForest	0.34	**0.56**	0.46	0.74
PAE	0.35	0.44	0.41	0.83
Neg. Sampling	0.32	0.33	**0.49**	0.43
Random	0.25	0.14	0.32	0.50

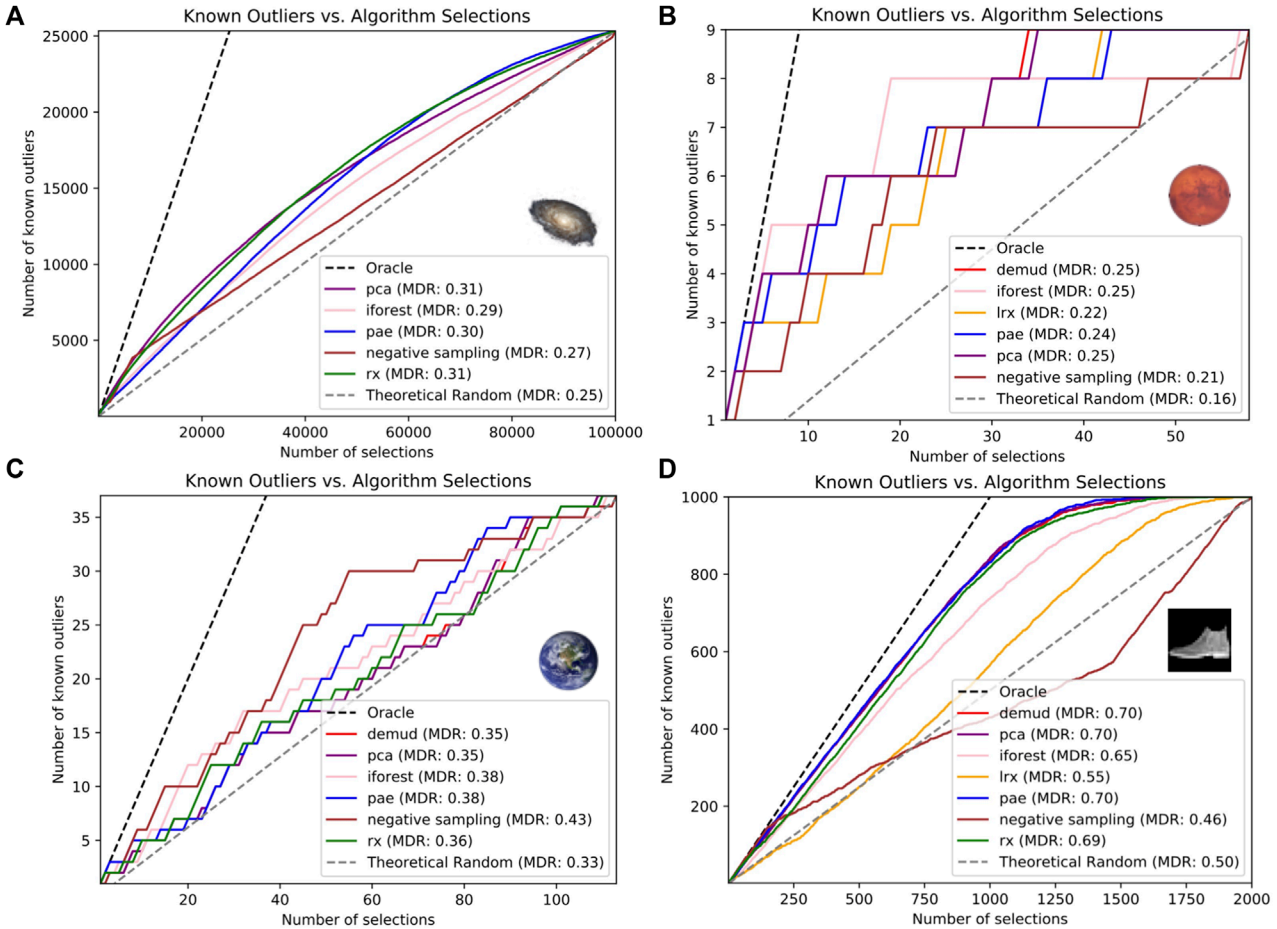

FIGURE 3 | Number of known outliers ranked in top *N* selections for the **(A)** astrophysics, **(B)** planetary, **(C)** Earth, and **(D)** FMNIST datasets.

6 DISCUSSION

6.1 Algorithm Performance

No one algorithm had the best performance across all four datasets. PCA had the best performance for the astrophysics and F-MNIST datasets, while negative sampling and isolation forest was best for the Earth science and planetary datasets respectively. This illustrates the importance of including a diverse set of algorithms and tools for easily inter-comparing them in DORA, since the best algorithm will vary for different datasets. The purpose of this study was to demonstrate how DORA could be used to facilitate outlier detection experiments and compare results across datasets from different domains. Thus we did not perform hyperparameter tuning which could improve results for each dataset; we leave this for future work.

6.2 Evaluation in Outlier Detection

Prior work has emphasized the difficulty of creating standardized metrics for outlier detection that represents how models will perform in real world settings while also enabling intercomparison between datasets (Campos et al., 2016). We chose two complementary metrics with this in mind: precision at $N = n_{outliers}$, which measures the fraction of selections that are known outliers when the number of selections is equivalent to the number of outliers, and Mean Discovery Rate, which measures the fraction of selections that are known outliers on average. Designing experiments to evaluate outlier detection methods for real-world use cases is also difficult because it is difficult, or sometimes impossible, to obtain labeled samples of outliers, inliers, or both for evaluation. In addition, labels are often subjective or uncertain, especially in the case of scientific datasets. For example, a dataset of known outliers was available for the astrophysics dataset from human annotation in prior work, but the remainder of samples in the dataset used for evaluation were not known to be inliers or outliers. This can result in evaluation metrics that are deceptively low because unlabeled samples that might actually be outliers (as was found to be common in prior work (Wagstaff et al., 2020a)) are counted as false positives.

6.3 Open Code and Data

Our goal is for DORA to enable increased application and benefit of outlier detection methods in real-world scientific use cases. We have designed DORA to make it as easy as possible for scientists to apply algorithms and to compare and interpret their results. Users need only to specify the specifics of their data (e.g., path, data type) in a configuration file to start running experiments and seeing results for their own datasets and use cases. DORA is publicly available and can be installed using pip via Github, making it easy to integrate into existing scientific workflows. The datasets used in this study are also publicly available via Zenodo. This enables DORA to be improved and expanded by the machine learning and domain science communities. If a researcher wants to use DORA for a dataset with a type that is not yet supported, they can contribute a new data loader by creating a subclass that extends the DataLoader abstract base class. Similarly, new results interpretation modules can be added by creating a subclass of the ResultsOrganization abstract base class. A new outlier ranking algorithm can be added by writing a new python module that defines a subclass of the OutlierDetection abstract base class and implements the required functions for scoring and ranking samples, following the existing algorithm modules named *_outlier_detection.py. In addition, DORA will be infused into the scientific workflows for the three use cases we demonstrated results for in this study. The DORA code can be accessed at https://github.com/nasaharvest/dora and datasets at https://doi.org/10.5281/zenodo.5941338.

7 CONCLUSION

The ability to automatically find outliers in large datasets is critical for a variety of scientific and real-world use cases. We presented Domain-agnostic Outlier Ranking Algorithms (DORA), a configurable pipeline that facilitates application and evaluation of outlier detection methods in a variety of domains. DORA minimizes the coding and ML expertise required for domain scientists since users need only to specify their experiment details in a configuration file to get results from all available algorithms. This is particularly important because the experiments for three cross-domain science datasets in this study showed that no one algorithm performs best for all datasets. DORA will be publicly accessible as a python package to make it easy to integrate into existing scientific workflows. The will be open-sourced to enable continued improvement and expansion of DORA to serve the needs of the science community. The datasets used in this study will also be public and can serve as real-world benchmarks for future outlier detection methods.

In future work, we will continue to improve DORA based on the experience of deploying it in the workflows of the domain scientists associated with the datasets in this study and add additional interpretation modules including causal inference graphs.

DATA AVAILABILITY STATEMENT

The datasets generated and analyzed for this study can be found in the "Multi-Domain Outlier Detection Dataset" repository on Zenodo (https://doi.org/10.5281/zenodo.5941338).

AUTHOR CONTRIBUTIONS

HK, UR, and KW developed the conception of the project. HK oversaw the overall implementation and led the analysis of the Earth science dataset. UR led the analysis of the astrophysics dataset. KW led the analysis of the planetary science dataset. SL led the development of the DORA software architecture. BD led the analysis of the FMNIST/MNIST dataset. EH supported the analysis of the astrophysics dataset. JL developed the desktop application for results visualization. VR supported the analysis of the Earth science dataset. SK supported the analysis of the planetary science dataset. All authors contributed to the overall system implementation and manuscript writing/editing.

FUNDING

This work was funded by a grant from the NASA Science Mission Directorate (under the NASA Cross-Divisional AI Use Case Demonstration grants; Grant No. 80NSSC21K0720). Part of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

ACKNOWLEDGMENTS

We thank the UN Food and Agriculture Organization (FAO) and Ritvik Sahajpal (University of Maryland/NASA Harvest) for providing the maize ground observation dataset. We thank Raymond Francis for analyzing the Navcam rover images to identify targets considered to be "novel" for evaluating outlier detection in the planetary rover scenario.

REFERENCES

Akcay, S., Atapour-Abarghouei, A., and Breckon, T. P. (2018). "Ganomaly: Semi-supervised Anomaly Detection via Adversarial Training," in Asian Conference on Computer Vision (Springer), 622–637.

Böhm, V., and Seljak, U. (2020). *Probabilistic Auto-Encoder*. *arXiv preprint arXiv:2006.05479*.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). "Lof," in Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 29. 93–104. SIGMOD Rec. doi:10.1145/335191.335388

Burl, M. C., Thompson, D. R., deGranville, C., and Bornstein, B. J. (2016). Rockster: Onboard Rock Segmentation through Edge Regrouping. *J. Aerospace Inf. Syst.* 13, 329–342. doi:10.2514/1.i010381

Campos, G. O., Zimek, A., Sander, J., Campello, R. J. G. B., Micenková, B., Schubert, E., et al. (2016). On the Evaluation of Unsupervised Outlier Detection:

Measures, Datasets, and an Empirical Study. *Data Min Knowl Disc* 30, 891–927. doi:10.1007/s10618-015-0444-8

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly Detection. *ACM Comput. Surv.* 41, 1–58. doi:10.1145/1541880.1541882

Chein-I Chang, C. I., and Shao-Shan Chiang, S. S. (2002). Anomaly Detection and Classification for Hyperspectral Imagery. *IEEE Trans. Geosci. Remote Sensing* 40, 1314–1325. doi:10.1109/tgrs.2002.800280

Hayat, M. A., Stein, G., Harrington, P., Lukić, Z., and Mustafa, M. (2021). Self-supervised Representation Learning for Astronomical Images. *ApJL* 911, L33. doi:10.3847/2041-8213/abf2c7

Jablonski, J. A., Bihl, T. J., and Bauer, K. W. (2015). Principal Component Reconstruction Error for Hyperspectral Anomaly Detection. *IEEE Geosci. Remote Sensing Lett.* 12, 1725–1729. doi:10.1109/lgrs.2015.2421813

Kerner, H., Hardgrove, C., Czarnecki, S., Gabriel, T., Mitrofanov, I., Litvak, M., et al. (2020). Analysis of Active Neutron Measurements from the mars Science Laboratory Dynamic Albedo of Neutrons Instrument: Intrinsic Variability, Outliers, and Implications for Future Investigations. *J. Geophys. Res. Planets* 125, e2019JE006264. doi:10.1029/2019je006264

Kerner, H. R., Wagstaff, K. L., Bue, B. D., Wellington, D. F., Jacob, S., Horton, P., et al. (2020). Comparison of novelty Detection Methods for Multispectral Images in Rover-Based Planetary Exploration Missions. *Data Min Knowl Disc* 34, 1642–1675. doi:10.1007/s10618-020-00697-6

Liang, S., Li, Y., and Srikant, R. (2018). "Enhancing the Reliability of Out-Of-Distribution Image Detection in Neural Networks," in 6th International Conference on Learning Representations (ICLR 2018).

Liu, F. T., Ting, K. M., and Zhou, Z. H. (2008). "Isolation forest," in 2008 8th IEEE International Conference on Data Mining (IEEE), 413–422. doi:10.1109/icdm.2008.17

Liu, Q., Klucik, R., Chen, C., Grant, G., Gallaher, D., Lv, Q., et al. (2017). Unsupervised Detection of Contextual Anomaly in Remotely Sensed Data. *Remote Sensing Environ.* 202, 75–87. doi:10.1016/j.rse.2017.01.034

Lochner, M., and Bassett, B. A. (2021). Astronomaly: Personalised Active Anomaly Detection in Astronomical Data. *Astron. Comput.* 36, 100481. doi:10.1016/j.ascom.2021.100481

Lupton, R. H., Gunn, J. E., and Szalay, A. S. (1999). A Modified Magnitude System that Produces Well-Behaved Magnitudes, Colors, and Errors Even for Low Signal-To-Noise Ratio Measurements. *Astronomical J.* 118, 1406–1410. doi:10.1086/301004

Markou, M., and Singh, S. (2003). Novelty Detection: a Review-Part 1: Statistical Approaches. *Signal. Process.* 83, 2481–2497. doi:10.1016/j.sigpro.2003.07.018

Markou, M., and Singh, S. (2003). Novelty Detection: a Review-Part 2:. *Signal. Process.* 83, 2499–2521. doi:10.1016/j.sigpro.2003.07.019

Molero, J. M., Garzon, E. M., Garcia, I., and Plaza, A. (2013). Analysis and Optimizations of Global and Local Versions of the Rx Algorithm for Anomaly Detection in Hyperspectral Data. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing* 6, 801–814. doi:10.1109/jstars.2013.2238609

Pimentel, M. A. F., Clifton, D. A., Clifton, L., and Tarassenko, L. (2014). A Review of novelty Detection. *Signal. Process.* 99, 215–249. doi:10.1016/j.sigpro.2013.12.026

Prochaska, J. X., Cornillon, P. C., and Reiman, D. M. (2021). Deep Learning of Sea Surface Temperature Patterns to Identify Ocean Extremes. *Remote Sensing* 13, 744. doi:10.3390/rs13040744

Reed, I. S., and Yu, X. (1990). Adaptive Multiple-Band Cfar Detection of an Optical Pattern with Unknown Spectral Distribution. *IEEE Trans. Acoust. Speech, Signal. Process.* 38, 1760–1770. doi:10.1109/29.60107

Richter, C., and Roy, N. (2017). Safe Visual Navigation via Deep Learning and novelty Detection. *Robotics: Sci. Syst.* doi:10.15607/rss.2017.xiii.064

Schölkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J., Platt, J. C., et al. (1999). Support Vector Method for novelty Detection. *Neural Inf. Process. Syst. (Citeseer)* 12, 582–588.

Sipple, J. (2020). "Interpretable, Multidimensional, Multimodal Anomaly Detection with Negative Sampling for Detection of Device Failure," in Proceedings of the 37th International Conference on Machine Learning, 119. 9016–9025.

Touati, R., Mignotte, M., and Dahmane, M. (2020). Anomaly Feature Learning for Unsupervised Change Detection in Heterogeneous Images: A Deep Sparse Residual Model. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing* 13, 588–600. doi:10.1109/jstars.2020.2964409

Wagstaff, K. L., Francis, R., Kerner, H., Lu, S., Nerrise, F., et al. (2020). "Novelty-driven Onboard Targeting for mars Rovers," in Proceedings of the International Symposium on Artificial Intelligence (Robotics and Automation in Space).

Wagstaff, K. L., Huff, E., and Rebbapragada, U. (2020). "Machine-assisted Discovery through Identification and Explanation of Anomalies in Astronomical Surveys," in Proceedings of the Astronomical Data Analysis Software and Systems Conference.

Wagstaff, K. L., Lanza, N. L., Thompson, D. R., Dietterich, T. G., and Gilmore, M. S. (2013). "Guiding Scientific Discovery with Explanations Using DEMUD," in Proceedings of the Twenty-Seventh Conference on Artificial Intelligence, 905–911.

Zhao, Y., Nasrullah, Z., and Li, Z. (2019). Pyod: A python Toolbox for Scalable Outlier Detection. *J. Machine Learn. Res.* 20, 1–7.

Zhou, J., Kwan, C., Ayhan, B., and Eismann, M. T. (2016). A Novel Cluster Kernel Rx Algorithm for Anomaly and Change Detection Using Hyperspectral Images. *IEEE Trans. Geosci. Remote Sensing* 54, 6497–6504. doi:10.1109/tgrs.2016.2585495

Zuntz, J., Sheldon, E., Samuroff, S., Troxel, M. A., Jarvis, M., MacCrann, N., et al. (2018). Dark Energy Survey Year 1 Results: Weak Lensing Shape Catalogues. *Monthly Notices R. Astronomical Soc.* 481, 1149–1182.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Astronomy and Space Sciences

ORIGINAL RESEARCH
published: 20 May 2022
doi: 10.3389/fspas.2022.875985

Classification of Cassini's Orbit Regions as Magnetosphere, Magnetosheath, and Solar Wind *via* Machine Learning

Kiley L. Yeakel[1]*, *Jon D. Vandegriff*[1], *Tadhg M. Garton*[2,3], *Caitriona M. Jackman*[3], *George Clark*[1], *Sarah K. Vines*[1], *Andrew W. Smith*[4] *and Peter Kollmann*[1]

[1]Johns Hopkins University Applied Physics Laboratory, Laurel, MD, United States, [2]Department of Physics and Astronomy, University of Southampton, Southampton, United Kingdom, [3]School of Cosmic Physics, Dublin Institute for Advanced Studies, Dublin, Ireland, [4]Mullard Space Science Laboratory, University College London, London, United Kingdom

Several machine learning algorithms and feature subsets from a variety of particle and magnetic field instruments on-board the Cassini spacecraft were explored for their utility in classifying orbit segments as magnetosphere, magnetosheath or solar wind. Using a list of manually detected magnetopause and bow shock crossings from mission scientists, random forest (RF), support vector machine (SVM), logistic regression (LR) and recurrent neural network long short-term memory (RNN LSTM) classification algorithms were trained and tested. A detailed error analysis revealed a RNN LSTM model provided the best overall performance with a 93.1% accuracy on the unseen test set and MCC score of 0.88 when utilizing 60 min of magnetometer data ($|B|$, B_θ, B_ϕ and B_R) to predict the region at the final time step. RF models using a combination of magnetometer and particle data, spanning H^+, He^+, He^{++} and electrons at a single time step, provided a nearly equivalent performance with a test set accuracy of 91.4% and MCC score of 0.84. Derived boundary crossings from each model's region predictions revealed that the RNN model was able to successfully detect 82.1% of labeled magnetopause crossings and 91.2% of labeled bow shock crossings, while the RF model using magnetometer and particle data detected 82.4 and 74.3%, respectively.

Keywords: recurrent neural network (RNN) long short-term memory (LSTM), random forest, machine learning, magnetosphere, boundary crossings, Saturn, Cassini-Huygens

OPEN ACCESS

Edited by:
Olga Verkhoglyadova, NASA Jet Propulsion Laboratory (JPL), United States

Reviewed by:
Tomas Karlsson, Royal Institute of Technology, Sweden
Primoz Kajdic, National Autonomous University of Mexico, Mexico

***Correspondence:**
Kiley L. Yeakel
kiley.yeakel@jhuapl.edu

Specialty section:
This article was submitted to Space Physics, a section of the journal Frontiers in Astronomy and Space Sciences

Received: *14 February 2022*
Accepted: *19 April 2022*
Published: *20 May 2022*

Citation:
Yeakel KL, Vandegriff JD, Garton TM, Jackman CM, Clark G, Vines SK, Smith AW and Kollmann P (2022) Classification of Cassini's Orbit Regions as Magnetosphere, Magnetosheath, and Solar Wind via Machine Learning. Front. Astron. Space Sci. 9:875985. doi: 10.3389/fspas.2022.875985

1 INTRODUCTION

Preliminary to any detailed studies of space physics phenomena is the detection and statistical quantification of large quantities of example "events" in data sets from orbiting spacecraft. At present, the detection and cataloging of such events is done primarily by visual inspection of the data sets by domain experts. Yet, as the current and near-future space missions continue to fly evermore data-intensive sensors, the space physics community is rapidly approaching a point in which the data volume vastly exceeds the analysis capacity of the domain experts (Azari et al., 2020). Additionally, manual detection and cataloging of the events embeds the bias of the individual observer into the curated catalog, consequently precluding the inter-comparison of results from two independent observers. Semi-automation of the event detection by using, for instance, a set of explainable threshold criteria to define an event, has helped to combat some of the inter-observer bias and reduce

the time needed to build event catalogs relative to a purely manual method. Yet, it can be the case that such rigid threshold criteria fail to replicate the subtle event detection/inspection process of the domain experts, or to appropriately account for the complexities introduced by the varying observer (spacecraft) position. Machine learning (ML) presents a viable alternative to the current best practice of manual inspection or semi-automated methodologies given the proven ability in other fields to comb through vast data reserves to find events of interest. In the space domain, with its exponentially increasing data archives, ML is becoming a necessity.

A common feature to identify in spacecraft data sets is the encounter of a spacecraft with magnetospheric boundaries such as the bow shock or magnetopause. The regions adjacent to these boundaries have very particular characteristics: the magnetosphere is dominated by planetary field and plasma; the magnetosheath is a region of turbulent, compressed, heated, shocked solar wind plasma, and the solar wind upstream of the bow shock can reflect a pattern of regular corotating interaction regions as well as revealing the presence of solar wind transients such as coronal mass ejections. There are many physical phenomena which occur in these different regions—e.g., from magnetic reconnection to wave-particle interactions—and robust region identification (magnetosphere vs magnetosheath vs solar wind) is often a necessary step prior to doing focused event detection surveys. At Earth, various studies have developed algorithms to detect magnetopause and bow shock boundaries based on changes in the time-based variance of the magnetic field, orientation of the magnetic field and the composition and properties of the local plasma from *in situ* spacecraft data (Ivchenko et al., 2000; Jelínek et al., 2012; Case and Wild, 2013; Olshevsky et al., 2021). Similar studies have been applied to splitting heliospheric measurements into categories based on *in situ* solar wind observation and using techniques such as Gaussian process classification (Xu and Borovsky, 2015; Camporeale et al., 2017). On the sun-ward side of the bow shock there is also a foreshock region, which displays properties similar to the magnetosheath. This is especially prominent at Earth where the orientation of interplanetary magnetic field (IMF) drives a quasi-parallel bow shock over a large extent of the boundary, and the resulting foreshock propagates shocked ions and magnetic field perturbations far upstream, obfuscating the solar wind population. The distinct characteristics of the quasi-parallel foreshock region at Earth has prompted ML-based region classification algorithm approaches to identify the foreshock as a fourth region in addition to the magnetopshere, magnetosheath and solar wind (Olshevsky et al., 2021). In contrast to Earth, the Parker spiral angle at Saturn is found to be larger at approximately 86.8 ± 0.3° (Jackman et al., 2008). Thus, Saturn's bow shock is primarily quasi-perpendicular to the IMF, and the foreshock will be pushed to the dawn side of the planet. While some studies have found evidence of quasi-parallel foreshocks at Saturn present in the Cassini data (Bertucci et al., 2007), in general, quasi-perpendicular bow shock crossings dominate (Sulaiman et al., 2016).

At Saturn, there are still large unknowns concerning physically processes within Saturn's magnetosphere as well as its interaction with the solar wind. For example, the role of dayside reconnection in controlling the magnetospheres of giant planets is still not fully understood (Guo et al., 2018). Increasing the event list that can enable detailed studies of physical phenomena, i.e., magnetic reconnection, can have impactful results in our understanding of physical drivers of magnetospheric dynamics such as particle injections and auroral pulsations (Guo et al., 2018). Therefore, in

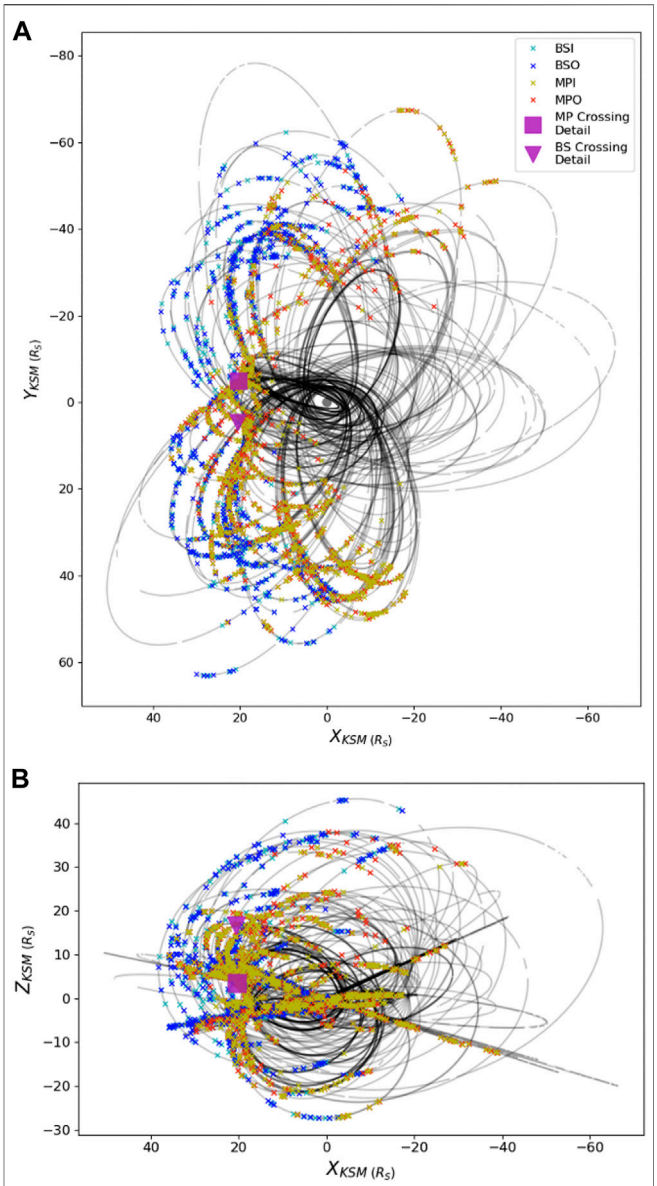

FIGURE 1 | Depiction of Cassini's orbits spanning 1 November 2004 (the end of the first capture orbit) through 15 September 2017 along with associated labeled boundary crossings in the X-Y KSM plane **(A)** and X-Z KSM plane **(B)**. The legend refers to the color denoting the four types of crossings in the data set: bow shock inbound (BSI, cyan), bow shock outbound(BSO, blue), magnetopause inbound (MPI, yellow) and magnetopuase outbound (MPO, red). The magenta box and triangle highlight case study MP and BS crossings, which were fully encapsulated in the test set and occurred on 3 May 2008 and 8 March 2008, respectively. These case studies are highlighted in **Figure 2**.

this study, we will focus on observations from the Cassini-Huygens mission, which orbited the Saturn system from 2004–2017, sampling Saturn's dynamic magnetosphere from a diversity of vantage points as highlighted in **Figure 1**. The variable orbit design of the Cassini mission meant that while most of the mission was spent taking measurements within the planetary magnetosphere, the magnetosheath and upstream solar wind was also frequently sampled. Each of the three regimes all have uniquely identifying characteristics of field and plasma, with transitions between the three regimes occasionally seen to occur at different times depending on the identifying data set being used (i.e., magnetometer data versus low-energy plasma or energetic particle data). Early studies included the publication of lists of boundary crossings (Pilkington et al., 2015), while other work included the development of empirical models to describe the shape and location of the magnetopause (Kanani et al., 2010) and bow shock (Went et al., 2011).

Since the conclusion of the Cassini mission in 2017, the full data set has been visually inspected and a list of bow shock and magnetopause crossings has been made available (Jackman et al., 2019). This list uses magnetometer data as the primary descriptor, with augmentation from plasma data (electron spectrometer) until the failure of the CAPS sensor in 2012. The list focuses on clear crossings of the boundaries and does not consider very short excursions (with duration < 2–3 min). It is a common issue that the timing of boundary crossings may appear slightly different as seen from different instrument platforms, due to the cadence of the measuring instruments, or to physical reasons such as finite gyroradius effects. The Jackman et al., 2019 list upon which we base this work placed the crossings at the location most closely aligned with the largest change in magnetic field and this property of the time labels must be remembered for subsequent analysis and interpretation. The Jackman et al. (2019) list serves as a basis for a supervised machine learning task in which we attempt to classify whether the spacecraft is in the magnetosphere, magnetosheath or solar wind. We explore the predictive value of different sensor measurements sampling the *in situ* magnetic and plasma environment versus the time-based variance of a subset of features, and compare algorithms of varying computational complexity. By utilizing an extensively verified event list as our basis, we can thoroughly examine the context of the algorithm predictions to elucidate whether ML-based approaches can sufficiently "learn" the physics of the system of interest.

2 METHODS

2.1 Data Sets and Problem Setting

In an effort to explore whether machine learning (ML) algorithms may be able to replicate the selection processes of the scientists, we explored classifying segments within Cassini's orbit according to one of three regions - the solar wind (upstream of the bow shock), magnetosheath (between the bow shock and magnetopause) or the magnetosphere (inside of the magnetopause). There are four possible types of crossings as identified by Jackman et al. (2019) - bow shock out (BSO; spacecraft is moving across the bow shock boundary from the magnetosheath to the solar wind), bow shock in (BSI; spacecraft is moving from the solar wind into the magnetosheath), magnetopause in (MPI; spacecraft is moving across the magnetopause from the magnetosheath into the magnetosphere) and magnetopause out (MPO; spacecraft is moving from the magnetosphere into the magnetosheath). Despite the long length of the Cassini mission, there were relatively few crossings - in total Jackman et al. (2019) found approximately 3,300 crossings over a span of twelve years (see **Figure 1** for a depiction of Cassini's orbit path and the locations of the boundary crossings). Structuring the ML approach to identify the three distinct regions in lieu of directly identifying crossings ensured much larger data sets were available for training, validation and testing, enabling a much greater variety of ML algorithms to be utilized. However, as a consequence of this approach, algorithm performance is optimized for identifying the bulk region (i.e., the mean conditions for each region) and can be expected to suffer in the vicinity of boundary transitions.

To identify the regions, we explored various combinations of data from four sensors: 1) the Cassini magnetometer (MAG) (Dougherty et al., 2005); 2) the Ion Mass Spectrometer (IMS) of the Cassini Plasma Spectrometer (CAPS) (Young et al., 2004) instrument suite; and two sensors from the Magnetospheric Imaging Instrument (MIMI) (Krimigis et al., 2004) suite: 3) the Low Energy Magnetospheric Measurement System (LEMMS) and 4) the Charge Energy Mass Spectrometer (CHEMS). For completeness, we briefly describe the instruments and the associated data products used for this study below but more detailed descriptions can be found in the instrument papers.

MAG: MAG consists of a fluxgate magnetometer (MAG) and vector helium magnetometer (VHM) also capable of operating in a scalar mode. For this study, we utilize MAG data interpolated to a one-minute sampling cadence in the KRTP coordinate frame.

CAPS/IMS: The Ion Mass Spectrometer (IMS), measures energy and mass resolved fluxes over an energy-per-charge range of 1 eV/q to 50 keV/q and consists of eight look directions. In this study, we use the ion singles data product averaged over a ten-minute window, and utilize directional data specifically from anode four spanning an energy range of $\approx 0.06\ eV - \approx 46.3 keV$

MIMI/LEMMS: LEMMS is a particle detector with two separate telescopes, a low-energy telescope (LET) and a high-energy telescope (HET). We utilize both LET and HET data as well as pulse height analyzed (PHA) data. The selected subset of LET and HET data capture proton fluxes spanning energy ranges of 27–158,700 keV, while the PHA data spans 25.7–67.7 keV, over which the dominate species present will be protons.

MIMI/CHEMS: CHEMS is an instrument designed to characterize the suprathermal ion population in Saturn's magnetosphere by measuring the charge state, energy, mass and angular distributions of ions (Krimigis et al., 2004). Double and triple count incidence data is utilized for H^+, He^+ and He^{++} over energy ranges of 2.81–220.2 keV.

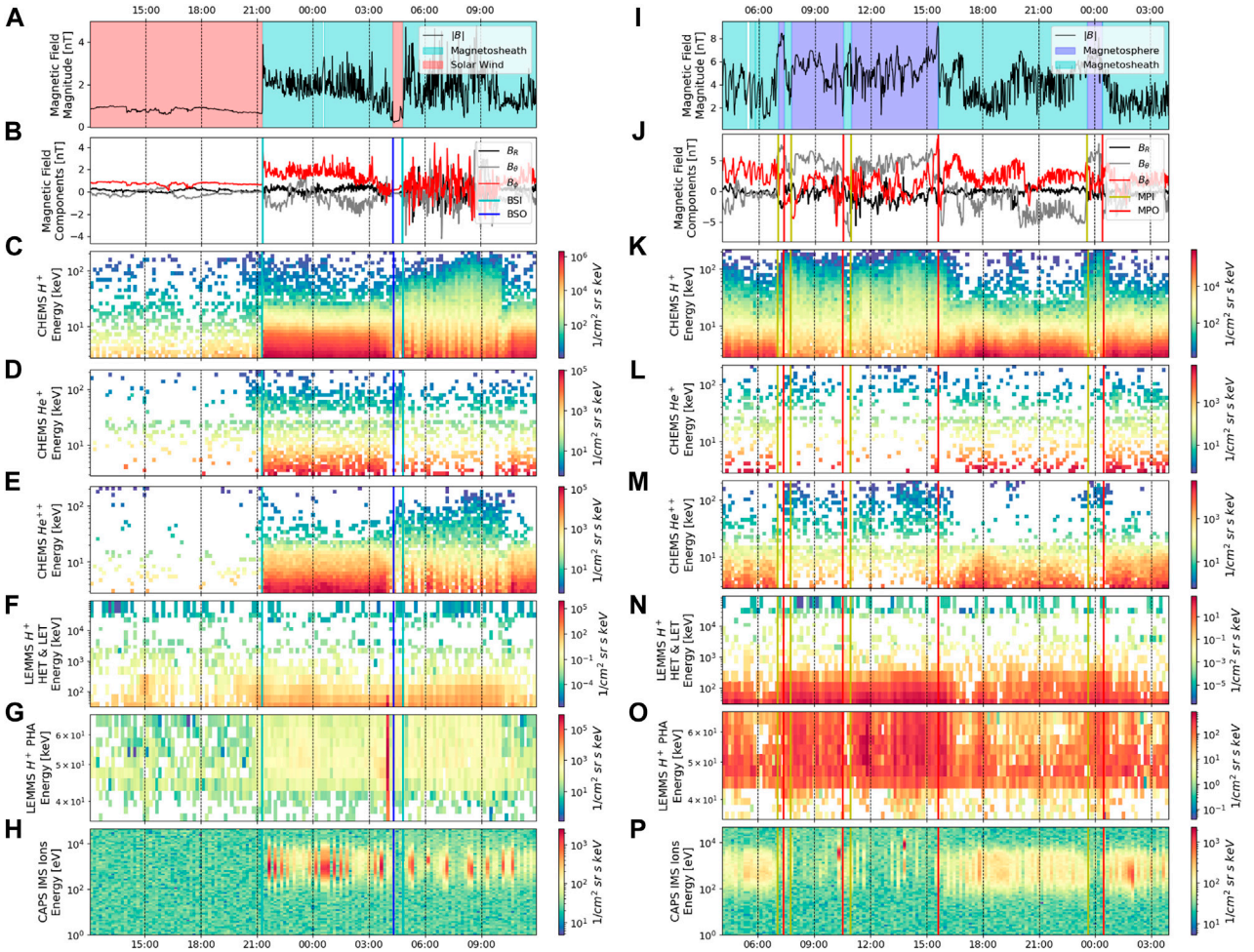

FIGURE 2 | Detailed view of MAG, MIMI CHEMS & LEMMS data, and CAPS/IMS data from two example 24-h periods on 8 March 2008 (left column) and 3 May 2008 (right column) in which the spacecraft passed through the BS and MP, respectively. The regions (magnetosphere, magnetosheath and solar wind) are denoted by shaded bands **(A)** and **(I)**, with vertical lines in panels **(B–H)** and **(J–P)** denoting the labeled crossings. The MAG data are shown on a 1-min interpolated sampling rate, while all other data are shown on a 10-min interpolated sampling rate. The only data shown from CAPS, CHEMS, and LEMMS are those utilized by the ML algorithms, and specifically spans targeted energy ranges of H^+, He^+ and He^{++} ions.

Figure 2 shows two example 24-h periods from 8 March 2008 and 3 May 2008, in which the spacecraft crossed through a bow shock and magnetopause crossing, respectively, with only the selected subsets of data from the MAG, CHEMS, LEMMS and CAPS instruments used in the later algorithm approaches shown. Note that LEMMS data below 35 keV is not shown in **Figure 2** due to known spurious instrument artifacts in those channels, however that data was included in the ML data sets to avoid embedding bias of known instrument performance issues in the training, validation and test data sets. Immediately evident is the rapidity with which the transitions into and out of regions can occur, with the spacecraft briefly transitioning into the solar wind over the span of just an hour (**Figure 2A**, at approximately 04:30) and likewise moving rapidly between the magnetosheath and magnetosphere (**Figure 2I**). We also see that the changes in the running mean and variance in the magnetic field magnitude closely align with the observed crossings as to be expected since the MAG data was used predominately by the scientists when discerning boundary crossings. In addition to the total field magnitude, the components of the magnetic field (shown in **Figure 2B**) can reveal particular characteristics of the regions. For example, the magnetosphere will primarily reflect the orientation of the planetary field, while the solar wind may reveal features such as field rotations associated with the crossings of the heliospheric current sheet (Jackman et al., 2004). Bow shock crossings are generally much clearer in the magnetometer data than magnetopause crossings as the character of the solar wind is typically vastly different to the character of the magnetosheath. In contrast, crossings of the magnetopause may be more or less clear in the magnetometer data depending on the relative orientations of the planetary field (inside the magnetosphere) versus the shocked interplanetary magnetic field (IMF; in the magnetosheath). From the perspective of ion observations, regions can generally be identified based on

different characteristic energies, spectral profiles, and composition. For example, the bulk solar wind ion populations are typically in a narrow range of 1–2 keV and thus below the minimum energy bin of CHEMS. However, the solar wind bulk ion population becomes heated while crossing the bow shock such that H^+ and He^{++} ions are within the energy range of the CHEMS instrument (a few to 10's keV). For magnetosphere to magnetosheath transitions, boundary transitions tend to appear most clearly in magnetometer data (**Figure 2**). For populations near the magnetopause, suprathermal and energetic ions and electrons have much larger gyroradii and lower densities, and so consequently will not always move in the direction of the bulk plasma flow. This can at times result in boundary transitions that appear "fuzzier" (Liou et al., 2021) as compared to low-energy, bulk species (particularly electrons) or the magnetometer data which can demonstrate sharp discontinuities between the various regions.

2.2 Data Set Preprocessing

Initial preprocessing of the data set consisted of applying background subtraction and calibration factors to convert instrument voltages to physical units. Data gaps which were noted in the reference crossing list from Jackman et al. (2019) were excluded from contention. Being sampled at a much higher cadence, the MAG data was interpolated to a 1-min sampling rate, with the other data features (CAPS/IMS, MIMI LEMMS and CHEMS) being interpolated to a 10-min sampling rate. MAG data were formatted in the Kronian Radial-Theta-Phi (KRTP) coordinates, a spherical polar coordinate system. B_R (the radial component) is positive radially outward from Saturn to the spacecraft, B_θ (the meridional component) is positive southward, and B_ϕ (the azimuthal component) is positive in the direction of corotation. Specific combinations of the features were then considered to elucidate feature importance in model prediction capability. Those subsets included (and their abbreviated name):

1) MAG at 1 min cadence
2) MAG at 10 min cadence
3) MAG with subset of CAPS/IMS and MIMI/LEMMS/CHEMS at 10 min cadence (MAG & subset particle)
4) Subset of CAPS/IMS and MIMI/LEMMS/CHEMS at 10 min cadence (Subset particle)
5) All CAPS/IMS and MIMI/LEMMS/CHEMS data at 10 min cadence (Full particle)
6) MAG with all CAPS/IMS and MIMI/LEMMS/CHEMS data at 10 min cadence (MAG & full particle)

For the MAG data, the features used consisted of the MAG field components in the KRTP system (B_R, B_θ, and B_ϕ) as well as the total magnitude of the magnetic field ($|B|$), giving a total of four total features. The specific "subset" of CAPS/IMS and MIMI/LEMMS/CHEMS data chosen were:

1) CAPS/IMS 8.002 eV ions
2) CAPS/IMS 107.654 eV ions
3) CAPS/IMS 16.387 keV ions
4) MIMI/CHEMS 3.78 keV protons
5) MIMI/CHEMS 6.75 keV protons
6) MIMI/LEMMS 44.27 keV protons

With this list of features specifically chosen due to their significantly divergent behavior from one another, and ability to provide the minimum set of representative channels. The "full particle" data set refers to the entire set of species and energy levels—specifically H^+, He^+ and He^{++} ions—as previously mentioned in the instrument descriptions. When all of the MIMI CHEMS, LEMMS and CAPS/IMS data were made available to the machine learning algorithms there were a total of 194 features. When combined with the magnetic field data, there were a total of 198 features. Given that CAPS data is included in all subsets of data featuring particle data, and that the CAPS sensor failed in 2012, all "particle" data sets only span through 2012, while MAG-only data sets span the entirety of the mission.

Spacecraft position data were never used as features within any of the ML approaches, given the sparsity of the space around Saturn through which Cassini flew relative to the entire region under the influence of Saturn's magnetic field. However, spacecraft position data were used to correct for sample imbalance within the three regions, ensuring that there was not an orbit bias to the training, validation or test data sets, and finally to interpret model results. The spacecraft position was calculated in the Kronocentric Solar Magnetospheric (KSM) coordinate system. In KSM coordinates, the X axis is the line from Saturn's center to the Sun, with positive X pointing in the direction of the Sun. The Y axis is the cross product of Saturn's magnetic axis with the X axis, and Z completes the triad. The XYZ KSM coordinates were then converted to spherical polar coordinates (R, θ, and ϕ) and θ was converted to magnetic local time, with noon along the line from Saturn's center to the Sun.

Initial data exploration revealed that there were far more data present within the magnetosphere than within either the solar wind or magnetosheath once the data from before the end of the first capture orbit (i.e., data collected before 1 November 2004) were removed. To correct for the sample discrepancy, regions within the orbit regime that were exclusively within the magnetosphere (and therefore the likelihood of a boundary crossing were zero), were removed from consideration. Those magnetosphere-exclusive regions were restricted to radial locations less than 15.1 Saturn radii (R_S = 60,268 km) and local time regions less than 2.81 h and greater than 20.8 h—corresponding exclusively to the nightside of the planet, far from the flanks and deep in the center of the magnetotail. The radial and magnetic local time thresholds were the location of the minimum radial and minimum/maximum local time positions of magnetopause crossings from our labeled crossings list. While removing these orbit regimes vastly improved the sample imbalance present in the data set, the resulting data set still had more samples from the magnetosphere than from within the magnetosheath or solar wind. Additionally, orbit locations in close proximity to Saturn's moon Titan were excluded. Titan orbits Saturn at a radial distance

of $\approx 20R_S$ which can take it very close to the nominal magnetopause location at certain local times—and the signatures of local field draping near Titan could be misleading for the ML algorithms. A list of Titan close flybys was used, with a buffer period 30 min before and after each event removed from consideration (Simon et al., 2015).

2.3 Machine Learning Algorithms

Two fundamental approaches were undertaken with regards to framing the ML classification problem: 1) classifying the region based on a single time point or 2) using a time series of points to classify the region the spacecraft was in at the last time step. The time series approach was motivated by the observation that the running mean and variance of a time series of features can provide indication of the region the spacecraft is transiting. For instance, in **Figure 2**, we see the total magnetic field magnitude ($|B|$) varies significantly in amplitude and variance in each of the three regions, with the magnetosheath having a very large running variance in $|B|$ while the magnetosphere has a higher mean $|B|$ but lower variance. Similarly, a single time step of data, if rich in features, may provide sufficient information to classify the region. Therefore, the two different approaches can be viewed as assessing the predictive capability of the time-related variance (i.e., gradients) of a small subset of features versus the predictive capability of many features at a single snapshot in time. The two approaches were also motivated by data availability, with only MAG data available at a 1-min cadence, and therefore the only set of features available in sufficient quantities for a deep learning, time-series approach. In contrast, many more features were available at a 10-min cadence, including data from MAG, CAPS/IMS and MIMI CHEMS and LEMMS.

To classify a single time point, several different combinations of algorithms and data sets were used. Algorithms that were tested include the multi-class implementation of logistic regression (LR), linear-kernel support vector machine (SVM), and a random forest (RF). For the LR approach, the multi-class implementation utilized a multinomial loss fit and a L2 norm penalization (Pedregosa et al., 2011). For the SVM model, the multi-class implementation utilized a one-versus-rest methodology in which 3 different one-versus-rest classifiers were trained (one classifier for each region) (Pedregosa et al., 2011). For the RF approach, hyperparameter tuning consisted of iterating on the number of trees in the forest and the minimum number of samples to define a leaf node. All combinations of the data mentioned in **Section 2.2** at a 10-min cadence were utilized. By varying the features that were used in the algorithm development, it was possible to assess whether certain sensor data (or combinations of sensor data) provided more predictive capability.

To classify the last time-point in a time series, a recurrent neural network (RNN) with long short-term memory (LSTM) cells was utilized. Because of the quantities of data required to appropriately train a RNN algorithm, only the 1-min MAG data was utilized with a total of four features—total magnetic field magnitude ($|B|$), and the magnetic field components in KRTP coordinates (B_R, B_θ, and B_ϕ). Variations in the number of LSTM layers (1–4) and the number of neurons per layer were explored, with a dropout layer (with a 50% drop rate) utilized after every LSTM layer. All neural network approaches were implemented *via* the TensorFlow module (Abadi et al., 2015). For multiple RNN layers, the full sequence (i.e., the output from all the neurons in the layer) was returned and passed along to the next layer. The output of the final neuron of the last RNN layer was passed to a dense fully-connected network with three neurons and a softmax activation. The Adam optimizer (Kingma and Ba, 2017) was used to train all variations of the RNN network, with the categorical cross entropy loss function and unweighted classification accuracy used to assess algorithm training progress. An early stopping criteria was implemented to prevent over-fitting, with training stopped if validation loss failed to achieve a minimum decrease of 0.001 over a period of two epochs.

For the time-series-based approach, it was necessary to sample from continuous segments of data, particularly because time-series ML approaches such as the RNNs used here have no concept of time other then the ordering of the samples fed to the algorithm (i.e., time stamps are not supplied). Within continuous segments of data, care was taken to sample the data such that the training, validation and test splits were not biased with regards to orbit location. It was found that reserving large continuous segments of data—such as an entire year—for the validation or test set produced an algorithm that was significantly biased. This is due to the large year-to-year variation in Cassini's orbit, which results in some years being biased towards an orbit scheme that was closer to Saturn (i.e., low R) or a more equatorial orbit scheme (i.e., low latitude). To reduce the bias between the three sets as much as possible, a weekly split was used (depicted in **Figure 3**) in which one week of continuous data was split into 105 h of training data, and 22.5 h each for the testing and validation data sets. A 6 h buffer between each of the sets was then discarded (18 h of data in total), which ensured that there was no overlap between the training, validation or test sets. When splitting the continuous data for the time-series-based approach, different sample lengths (20 versus 40 versus 60 min) were explored. A 5-sample "stride", where "stride" refers to the number of samples skipped over before the next sample is indexed, was used for all iterations. As an example, using a 20 min sample length with a five sample stride, sample one would utilize the data indexed from 0 to 19, while sample two would utilize the data indexed from 5 to 24. Offsetting the samples in this way ensured that there were still enough samples to attempt more data-intensive methods such as RNNs, but that samples were not so closely overlapped that over-fitting was a concern. By analyzing the distribution of spacecraft positions in KSM coordinates for the overall data set as well as across the training, validation and test sets, it could be deduced whether the time-based splicing induced any bias. **Figure 4** shows the spacecraft position histograms for the 1-min-interpolated MAG data that was utilized by the RNN (**a**—**c**) and the 10-min-interpolated data used in the SVM/LR/RF algorithms (**d**—**f**). Generally, the time-based splitting produced a relatively equal distribution across the overall sets and the three subsets for the 1-minute-interpolated data. There does appear to be some slight aliasing in the local time for the three sets (**Figures**

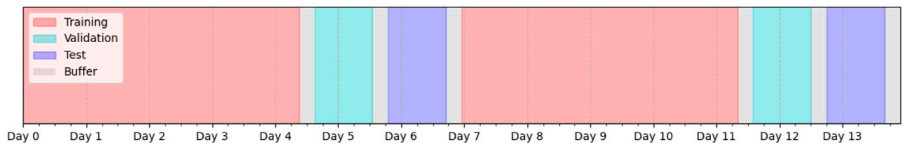

FIGURE 3 | Depiction of time-based splitting of data set into training, validation and test splits. 105 h of continuous data from each week were reserved for the training set, and 22.5 h each for the validation and test sets. A 6 h buffer period between each of the three sets was discarded, ensuring that there was no overlap between the sets.

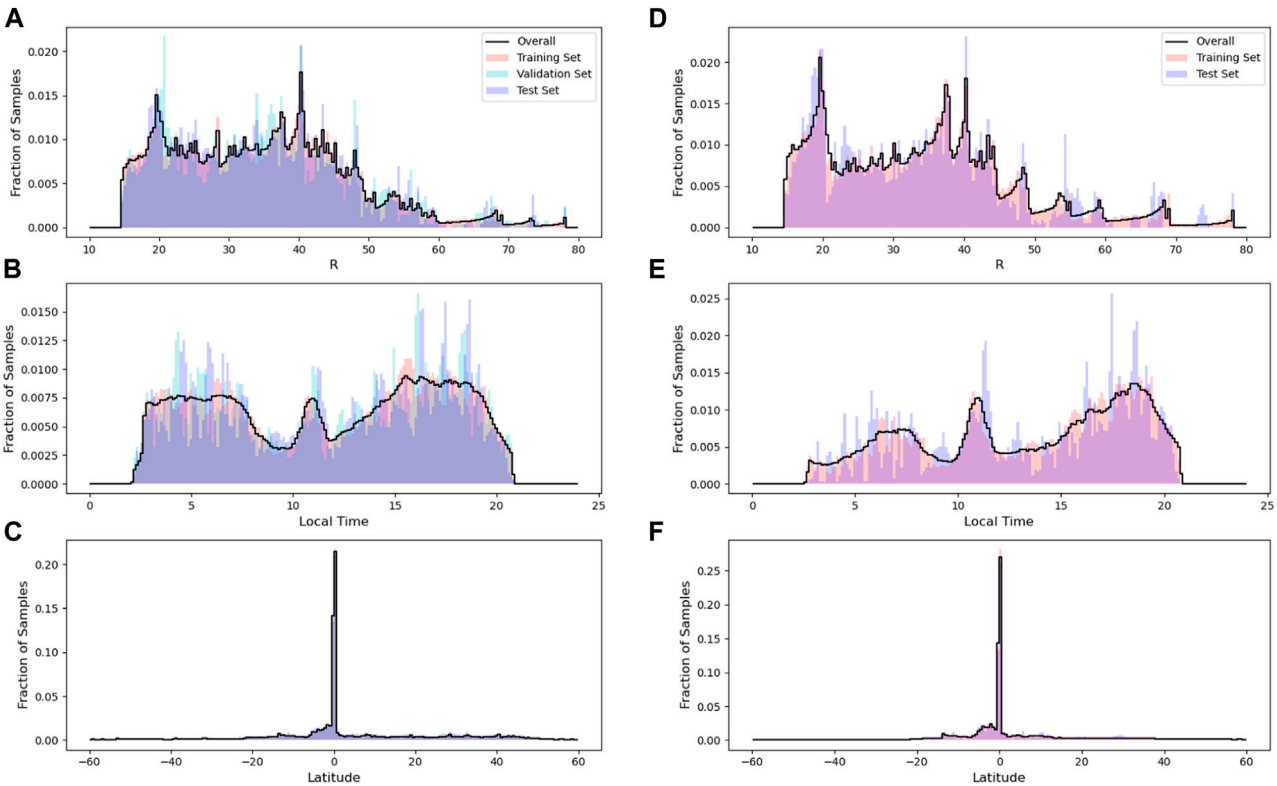

FIGURE 4 | Histogram of spacecraft position in KSM coordinates for Saturn radius [R; **(A)** and **(D)**], local time [**(B,E)**] and latitude [**(C,F)**] for the 1-min interpolated MAG-only data used in the RNN approaches (left panels) and 10-min interpolated MAG and full particle data set used in the RF/SVM/LR approach. For the RNN approach the data is split across the overall data set (black line), training set (red), validation set (cyan) and test set (blue). For the RF approach, a validation set was not used but the time periods of the test and training sets were matched as closely as possible to the RNN data sets for the sake of comparison. The histograms are scaled relative to the total number of samples in each set (for the 1-min interpolated data: N_{train} = 504,338, N_{val} = 109,582, and N_{test} = 109,259; for the 10-min interpolated data N_{train} = 103,665 and N_{test} = 32,475).

4B,E) which is may be related to the periodicity of Cassini's orbit and the week period chosen to do the time splitting.

While all iterations of the RNN approach used the same four features derived from the MAG data, the SVM, LR and RF approaches used solely the 10-min interpolated data sets (previously mentioned in **Section 3.2**) since even at a lower sampling cadence there were still sufficient amounts of training and test data. Both approaches used the time-based splitting procedure previously described, with the sets spanning the same intervals whether at a 1-min or 10-min cadence to allow comparisons across the algorithms. In other words, the same time span used for training a RNN algorithm with the 1-min-interpolated data was used to train the SVM/LR/RF algorithms with 10-min-interpolated data. One deviation between the two approaches was to ignore the validation data for the SVM/LR/RF algorithms since these algorithms do not require epoch-based training.

The final pre-processing step that was completed prior to ML algorithm development was to standardize and scale each of the features independently. This was done using the python scikit-learn "Robust Scaler" algorithm, which operates on each feature independently, removing the median and scaling the data to the range of the 1st (25%) and 3rd quartiles (75%) (Pedregosa et al., 2011). After scaling the features, the training, validation and test

TABLE 1 | Number of samples in each region for the training, validation and test sets. Note that for the 10-min interpolated data sets, which were only used by the SVM, RF, and LR classifier, a validation data set was not used. The time spans of the training, validation and test sets remained as close as possible across the different sets to allow for intracomparison of the model results.

Data Set	Total ($N_{train}/N_{val}/N_{test}$)	Magnetosphere	Magnetosheath	Solar Wind
1-Minute MAG	504300/109500/109200	290692/63265/63851	130017/29821/28024	83591/16414/17325
10-Minute MAG	265300/-/62400	152646/-/36262	68553/-/16195	44101/-/9943
10-Min. Some Particle	142925/-/33245	86584/-/20229	33173/-/8532	23168/-/4484
10-Min. Full Particle	139665/-/32475	84714/-/19795	32484/-/8307	22467/-/4373
10-Min. MAG & Some Particle	142925/-/33245	86584/-/20229	33173/-/8532	23168/-/4484
10-Min. MAG & Full Particle	139665/-/32475	84714/-/19795	32484/-/8307	22467/-/4373

sets were randomly shuffled. The final breakdown of the number of samples in each region for the training, validation and test sets is shown in **Table 1**. As is evident in **Table 1**, there remained approximately three times as many samples from within the magnetosphere than either the magnetosheath or the solar wind even after removing samples within a low radial or near midnight local time position.

2.4 Error Metrics

The algorithms report a confidence in each of the three regions, the maximum of which was taken as the prediction and compared to the accompanying label for the sample. We measured the effectiveness of the various ML models on the unseen test samples using four different metrics: accuracy, balanced accuracy, Matthew's Correlation Coefficient (MCC) and the F1 score. Accuracy is simply the ratio of the number of correct samples to the total number of test samples, where no weighting has been applied to any samples from a particular class. Balanced accuracy, in contrast, accounts for the sample imbalance in the test set and weights samples from a particular class according to the occurrence of that class within the test set. The weighting for a sample from a particular class is simply the fraction of test samples which belong to that class. In this instance, in which magnetosphere samples outnumber the magnetosheath and solar wind samples by a factor of roughly three, it can be expected that the balanced accuracy will give a more appropriate depiction of the model's performance across all the classes.

In the binary case, the F1 score is the harmonic mean of the precision and recall:

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{1}$$

where precision is defined as:

$$precision = \frac{TP}{TP + FP} \tag{2}$$

and can be interpreted as the ability of the model to maximize the detection of true events while minimizing the detection of false events. Recall is defined as:

$$recall = \frac{TP}{TP + FN} \tag{3}$$

and can be interpreted as the ability of the model to correctly identify all the events in the test set. The subcomponents for precision and recall are also best described in the binary case: True Positives (TP) are positive-class samples that have been correctly identified as positive by the model, False Positives (FP) are negative class samples that have been incorrectly identified as positive by the model, with True Negatives (TN) and False Negatives (FN) defined similarly for the negative samples. In the multi-class setting, the F1 score was calculated for each class independently, and then combined into a single metric using a weighted average. The "weight" of a class's F1 was scaled as the ratio of the samples from a particular class to the total number of test samples.

Matthew's Correlation Coefficient (MCC) was derived in the binary case as a means of encompassing the confusion matrix within a singular number (Matthews, 1975). The MCC in the binary case is described by the following equation:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

For a model which has predictions which are perfectly anti-correlated with the labels, MCC will return a value of −1, while for a model in which the predictions are perfectly correlated with the labels MCC will return a value of +1. For a model in which there is no relationship evident between the predictions and the labels (i.e. predictions are equivalent to a random guess), MCC will return a value of 0. MCC was extended to the multi-class setting by Gorodkin (2004), and in such cases the lower limit for anti-correlation may range between 0 and −1, but the maximum remains +1 for perfect correlation. For the sake of brevity, the equation for MCC in the multi-class setting which was used in our model evaluation is not shown here (see (Gorodkin, 2004) for details). Recent evidence has pointed to MCC being a more informative and less misleading metric than F1 or accuracy (Chicco and Jurman, 2020). All of the classification metrics were implemented *via* scikit-learn (Pedregosa et al., 2011).

3 RESULTS

3.1 Single Time Step Classification Results

Table 2 provides a breakdown in the performance of the SVM, LR, and RF models for different combinations of feature sets. We find that across all feature sets, the RF model, when appropriately tuned, performs the best. **Figure 5** illustrates the accuracy of the RF models at predicting the three regions when utilizing different

TABLE 2 | Comparison of SVM, Logistic Regression and RF models using various feature sets as described in the Methods section. Depending on the feature sets used, the amount of training and test data available will change, however all the time intervals used for training and testing are consistent across the different feature sets.

Feature Set	Model Type	Accuracy	Balanced Accuracy	F1	MCC
MAG	SVM	78.45%	72.77%	0.766	0.623
	Logistic	78.35%	73.86%	0.778	0.620
	RF	82.21%	77.22%	0.820	0.686
Some Particle	SVM	69.18%	45.88%	0.625	0.370
	Logistic	66.03%	41.32%	0.575	0.282
	RF	73.99%	56.69%	0.710	0.484
Full Particle	SVM	68.60%	56.82%	0.682	0.409
	Logistic	41.14%	37.10%	0.426	0.054
	RF	86.11%	81.49%	0.858	0.740
MAG & Some Particle	SVM	84.02%	78.97%	0.835	0.708
	Logistic	81.78%	73.91%	0.806	0.657
	RF	87.08%	82.08%	0.869	0.760
MAG & Full Particle	SVM	78.41%	72.81%	0.777	0.603
	Logistic	43.82%	37.68%	0.449	0.053
	RF	91.38%	88.83%	0.912	0.840

combinations of feature sets. We find generally that utilizing the CAPS/IMS and MIMI/CHEMS/LEMMS data alone, without any magnetic field data, leads to a model which over-predicts the magnetosphere region. This is particularly the case when utilizing only the small subset of features (6 total) from the CAPS/IMS and MIMI/LEMMS/CHEMS data set (see **Section 3.2** for a list of features). In contrast, using the magnetic field data alone provides some physical interpretation of the different regions, since the magnitude of the magnetic field acts as a proxy for the radial distance from the planet. However, we see there is still confusion between the adjoining regions - solar wind being confused for magnetosheath or magnetosheath being confused for magnetosphere, and vice versa. The best performance is found when the MAG and full particle data set is used as shown in **Figure 5E**. While the model using this feature set still confuses magnetosheath samples for magnetosphere, we generally see a much improved performance over the models using either only the magnetometer data or only the CAPS/IMS and MIMI/CHEMS/LEMMS data. In contrast to the RF models, the SVM and LR models fail to approach the same accuracy level on the test set predictions, except for when the MAG data is used alone.

When comparing the performance of different input sets it needs to be considered that boundaries and regions can appear different in different measurements. Boundaries can appear generally more gradual in energetic particle data (Mauk et al., 2019; Liou et al., 2021) and show dependencies on particle energy and direction that are still under scientific investigation (Mauk et al., 2016, 2019). Results from particle measurements that disagree from magnetic measurements are therefore not necessarily wrong from the scientific perspective but are a signature of physical processes such as particle escape that effectively soften up boundaries. However, our goal here is not to understand the underlying physics but to find the best defined boundaries, which can be found through magnetic field measurements. We therefore calculate our error measures relative the manually derived list that relied on magnetic field data.

3.2 Time Series Classification Results

Table 3 shows the RNN model performances for varying time sequence lengths along with the hyperparameters for the best-performing model at each time segment length. As mentioned in the Methods section, the number of layers and number of neurons per layer was iterated on to find the best performing model without overfitting. An exhaustive search for the optimal number of layers and neurons per layer was not performed due to the limitations on time and computational resources. However, general trends in test accuracy and test loss were observed by iterating over various combinations of neurons and layers. Overall, it can be observed that the 60-min RNN model provides the best performance on the test set. There was some slight overfitting (as can be seen by comparing the training loss with the validation and test set loss), however, stopping criteria were implemented to prevent substantial over-fitting. It also should be noted that the number of samples for the training, validation and test sets changed slightly between the 20-, 40-, and 60-min models due to the length of the time sample and allowed overlap between samples.

As the length of the time segment increases from 20 to 40–60 min, we see overall accuracy slightly increases, as indicated in **Table 3**. Therefore, it can be reasonably concluded that the gradients of the individual features and amount of variance in the features over the selected time frame is important for correctly classifying the region. Essentially, the longer the time segment, the more contextual information is provided to the model which allows for correct prediction of the region at the last time step. This is even more noticeable when we consider the samples which contain a boundary transition, which are a very small subset of the overall sample set. As shown in **Figure 6**, there is a drastic improvement in the model's accuracy for the small subset of samples containing a listed boundary transition as we increase the length of the sample. The improvement in accuracy is most drastic when moving from a 20-min sample to a 40-min

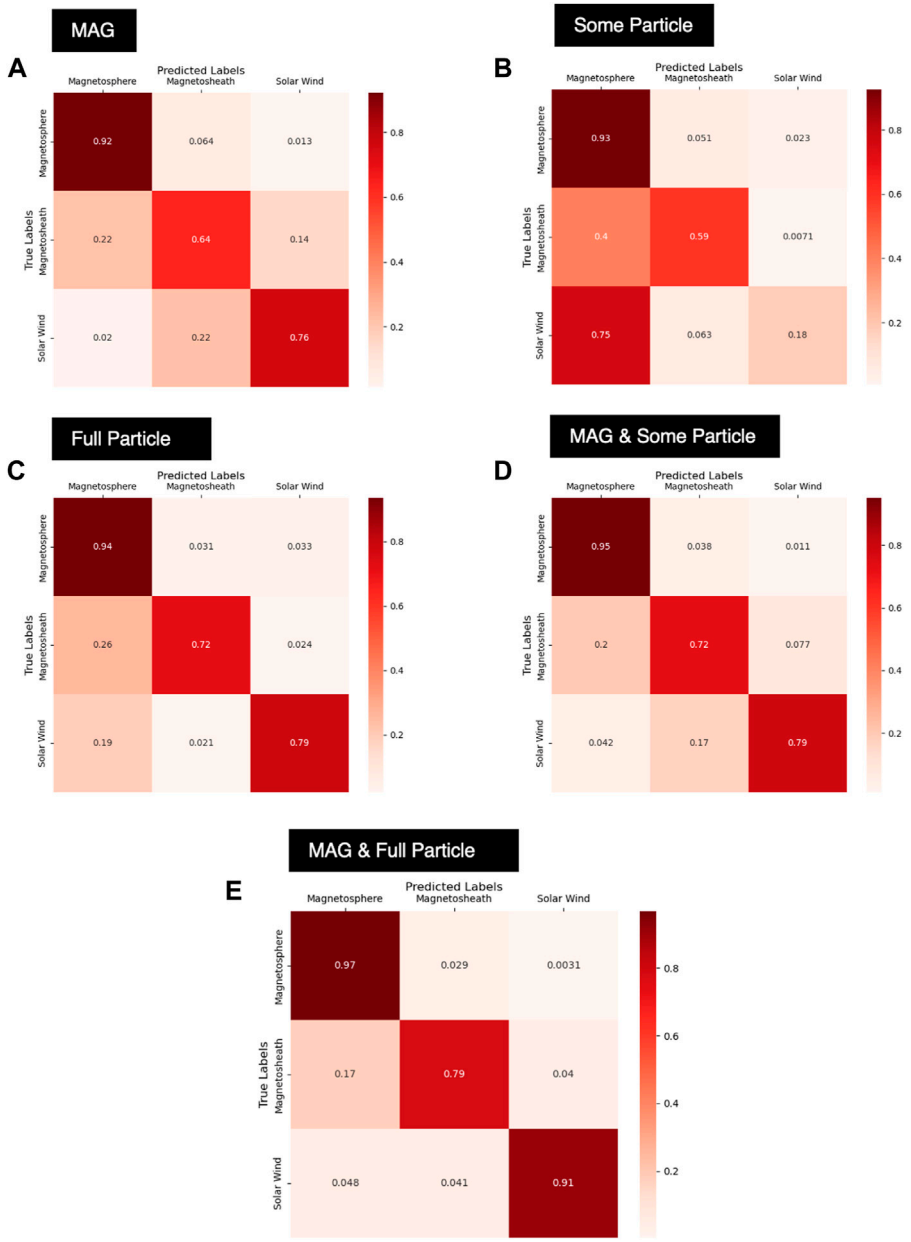

FIGURE 5 | Normalized confusion matrices for different combinations of data, all using the RF model which was the best performing model across all feature sets [**(A)** MAG-only, **(B)** Some particle, **(C)** Full particle, **(D)** MAG & some particle, and **(E)** MAG & full particle]. The comparisons here are shown as normalized confusion matrices in which each row is divided by the number of "true" samples in the class. A perfect model would have all ones on the diagonal and all zeros on the off-diagonal.

TABLE 3 | Comparison of RNN model performance for differing time sequence lengths as well as relevant model parameters.

Parameter	20-Minute Model	40-Minute Model	60-Minute Model
Accuracy	92.25%	93.08%	93.14%
Balanced Accuracy	91.69%	92.76%	93.08%
F1	0.923	0.931	0.932
MCC	0.863	0.877	0.878
Number Layers	2	2	1
Number Neurons	120	120	180
Trainable Parameters	176043	176043	133743

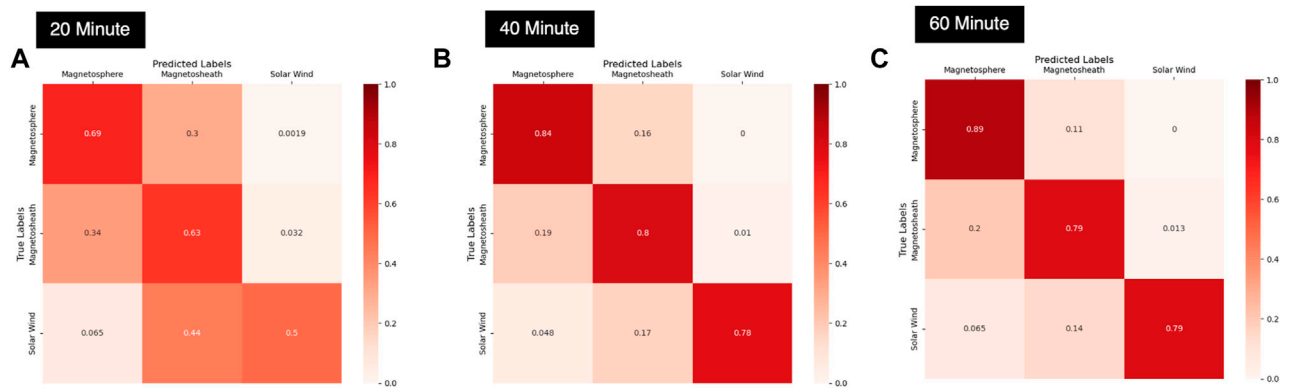

FIGURE 6 | Comparison of length of the time sample used with overall prediction performance on samples containing a boundary transition for the 20-min **(A)**, 40-min **(B)** and 60-min **(C)** RNN models. The comparisons here are shown as normalized confusion matrices.

sample, but incremental improvements are also observed as we increase the sample length from 40 to 60 min. Most notably, the confusion between the magnetosphere and magnetosheath regions decreases, which is where most of the confusion lies for the 20-min model. The number of samples containing a boundary transition is only approximately 1.5% of the total samples in the test set (1,619 samples out of 109,200 test samples for the 60-min model), however we see the improvement in accurately predicting the sample jumps from 62.28% for the 20-min model to 81.03% for the 40-min model to 84.25% for the 60-min model.

3.3 Spatial Errors

All remaining analysis is focused on three models in particular—RF MAG, RF MAG & full particle, and RNN 60-min MAG model. These models were chosen as the best performers for time series classification (RNN 60-min) and single time point classification (RF MAG & full particle). The results from the RF MAG model are shown given that they are the closest comparison in feature space to the RNN model. **Figure 7** shows the spatial discrepancies between the model predictions and the labeled data for three models in particular—the RF model with only MAG data [predictions (b) and difference from actual **(c)**], the RF model with MAG and full particle data set [predictions (d) and difference from actual **(e)**] and the RNN 60-min model with only MAG data [predictions (f) and difference from actual **(g)**]. The data has been binned according to local time and R, with the total number of predictions or labels of a particular region (magnetosphere, magnetosheath or solar wind) in a particular polar bin scaled to the total number of observations across all regions in that bin. The discrepancy plots have been scaled to highlight differences between the actual fraction of a region in a polar bin to the predicted fraction exceeding +/− 0.25. Despite having no information about the spacecraft position, we see in all cases that the models are generally able to correctly discern the physical layering of the problem, with the magnetosphere most commonly predicted radially close to the planet, the solar wind farthest from the planet, and the magnetosheath sandwiched in between.

The discrepancy polar histograms (**Figures 7C,E,G**) show the differences between the binning of the model predictions of particular regions and the binning of the labeled data (a), revealing where the model has under-predicted (in blue) or over-predicted (in red) a particular region. It is clear the RF model utilizing only MAG data performs the worst (as is also evident in comparing it's test accuracy with that of the RF MAG & full particle model and the RNN 60-min model). Utilizing only the MAG data set at a single time step, the model has much greater confusion on the spatial location of the magnetosphere and magnetosheath regions. We see a strong tendency to over-predict the magnetosphere and under-predict the magnetosheath on the dawn side of the planet. This confusion between the magnetosheath and the magnetsphere is then reversed on the dusk side of the planet, where there is a preference to under-predict the magnetosphere and over-predict the magnetosheath. Dawn-side errors could be due to the presence of the foreshock, which, as previously mentioned, causes large perturbations in the solar wind magnetic field. When the full particle data set is added to the RF model, we see that the spatial discrepancies are drastically improved as compared to the MAG data alone. **Figure 7E** shows that instead of the strong dawn/dusk preferences in the model predictions that we see with the MAG-only RF model (c) for the magnetosphere and magnetsheath predictions, that generally the MAG & full particle RF model tends to over-predict the magnetosphere and under-predict the magnetosheath at all radial and local time bins. Finally, the RNN 60-min model demonstrates the best spatial accuracy of the three (**Figure 7G**), with the least amount of spatial discrepancy from the true labels. It can be observed, however, that the 60-min RNN model, using the same feature set as the RF MAG model, again shows the dawn/dusk confusion between the magnetosphere and magnetosheath regions, though to a much lesser degree than the RF MAG model. In particular, there appears to be a very spatially narrow (14:00 to 16:00 h LT and 20–40 R_S) but discernible preference to over-predict magnetosheath and under-predict magnetosphere. Outside of this spatially-narrow region,

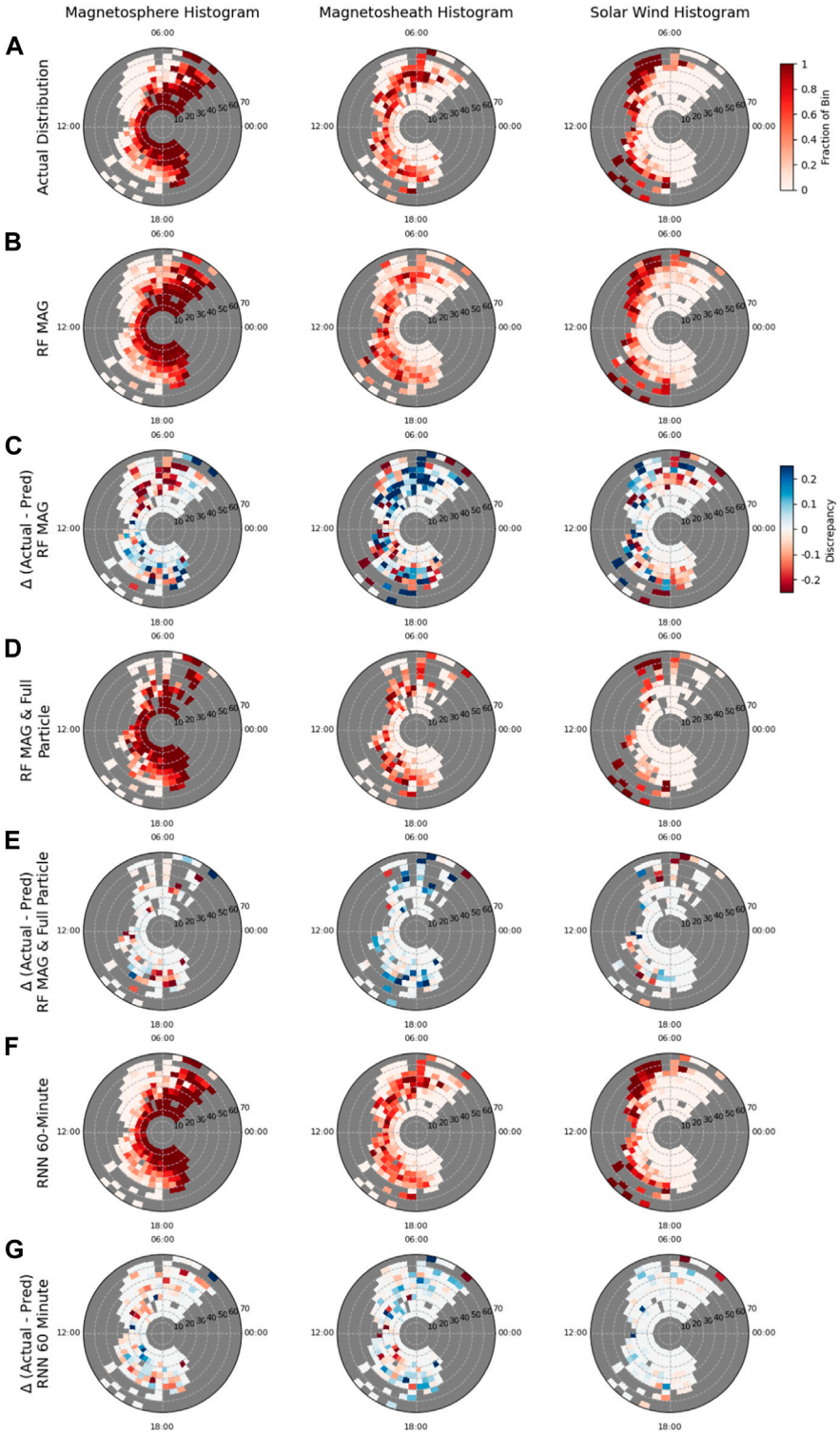

FIGURE 7 | Comparison on the spatial distribution of predictions for the magnetosphere (left column), magnetosheath (middle column) and solar wind (right column), for the true labels **(A)**, RF MAG-only model **(B)**, RF MAG & full particle model **(D)** and RNN 60-min MAG-only model **(F)**. The predictions for a particular region have been binned by local time (0.5 h increment) and radial distance (5 R_S increment), with bins in gray indicating where there is no data. The discrepancy polar histograms **(C,E,G)** shows the difference between the observed fraction of a bin labeled as a particular region **(A)** and the predicted fraction **(B,D,or,F)**, highlighting model errors. Discrepancy bins trending toward red indicate where the model has over-predicted a region, while blue indicates where a model has under-predicted a region. Panels **(B,D,F)** share the colorbar in panel **(A)**, while panels **(E,G)** share the colorbar for panel **(C)**.

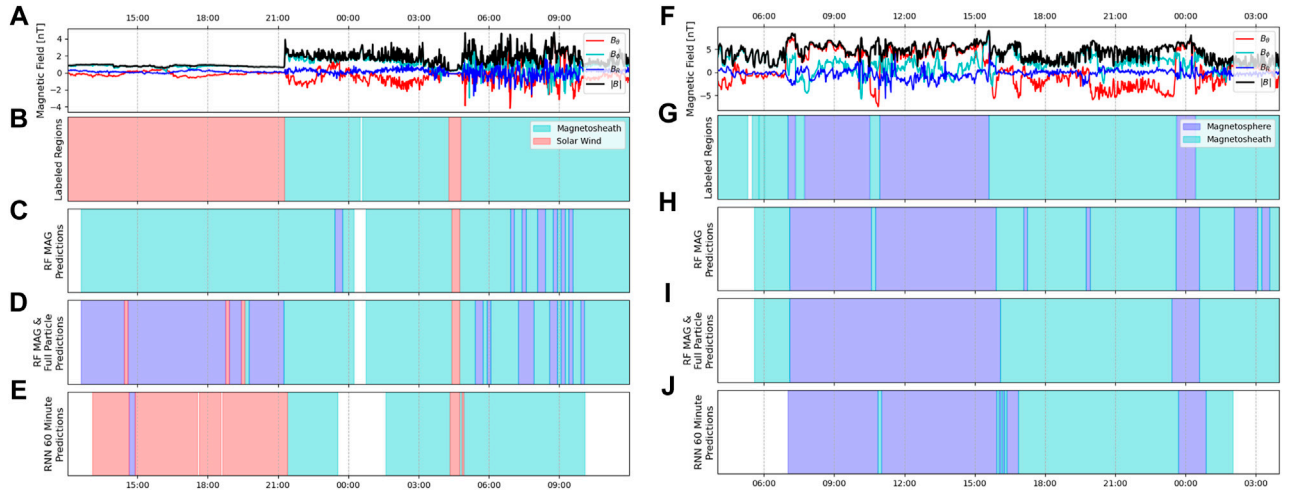

FIGURE 8 | An example BS crossing is shown from 8 March 2008 (left column) and an example MP crossing is shown from 3 May 2008 (right column), coincident with the transitions shown in **Figure 2**. Magnetic field magnitude and KRTP components are shown **(A,F)** along with regions as denoted by the labeled MP and BS crossings **(B,G)** and the model predictions from three selected models - RF MAG **(C,H)**, RF MAG & full ion **(D,I)**, and RNN MAG 60-min model **(E,J)**. Time periods which do not have any predictions due to a lack of data are denoted by white. Ideally the time periods of the model-predicted regions would coincide directly with the labeled regions. Here, we see that the RNN 60-min model comes closest to the labeled test set, while the RF models predict significantly more transitions, and particularly with the BS crossing, they are incorrect in their prediction of the solar wind region.

however, we find that the RNN model tends to underestimate the magnetosheath and overestimate the magnetosphere across all areas, similar to the Mag & full particle RF model. Confusion at low radial positions does not appear to be driven by traversals of the cusp region (see **Supplemental Material** and Figure ??), though previous studies have shown that Saturn's cusp shows a depressed magnetic field relative to the surrounding magnetosphere (Jasinski et al., 2017) and contains magnetosheath plasma (Arridge et al., 2016; Jasinski et al., 2016). In the case of the MAG-only models, a depressed magnetic field may cause the algorithm to predict magnetosheath in lieu of magnetosphere, while the RF MAG & full particle model would likewise predict magnetosheath due to the presence of magnetosheath plasma. The 60-min RNN model demonstrates by far the best spatial accuracy in predictions of the solar wind (**Figure 7G**, right), with virtually no discrepancy from the label set with the exception of a few large radial, dawn side bins where the solar wind is over-predicted. It is important to note that the RF model utilizing the MAG & full particle data set has far fewer samples than the MAG-only RF and RNN models due to the failure of the CAPS sensor in 2012.

3.4 Temporal Errors

To investigate temporal consistency in the model predictions, each continuous segment of testing data (22.5 h of data per week) was individually analyzed. Example segments demonstrating a bow shock and magnetopause crossing are shown in **Figure 8** and are directly corollary to the crossing shown in **Figure 2**, showing the temporal evolution of the predictions for the three models in particular. Within each continuous segment of data the time points in which predictions changed from one region to another can be used to derive the model's predicted crossings. Counting the number of predicted crossings in the continuous time frame and comparing with the labeled crossings over the same segment can thus provide an indication of the temporal consistency of the model's output. The example test segments shown in **Figure 8** is one such example of how a worse-performing model will have much less consistency in its predictions, with the RF models predicting far more transitions than the RNN 60-min model, as well as being largely incorrect in the case of the BS crossing. The numbers of predicted and actual transitions in each weekly test period were counted and summed up to the encompassing month for ease of comparison across the entire length of the mission. The results are shown in **Figure 9** for RF MAG (b), RF MAG & full particle (c) and RNN 60-min MAG (d). Here there are four possible types of transitions (as defined earlier)—BSI, BSO, MPI, and MPO. It is important to note, however, that these inbound/outbound notations simply refer to the spacecraft direction of travel at the time of the boundary encounter, and there is no expectation that the character of the regions on either side of the transition would be biased by the travel direction of the spacecraft.

All the models analyzed drastically over-predicted the number of transitions occurring, with the RF models demonstrating more false boundary transitions than the RNN 60-min model. The MAG-only RF model performs the worst of all, with significantly higher numbers of false transitions predicted at every time interval. Of all the RF models analyzed (see the appendix for all possible feature combinations), we find that the MAG & full particle data set (**Figure 9C**) produces the greatest consistency in region prediction (i.e., least false transitions). The RNN 60-min MAG model performs better yet (**Figure 9D**), while still predicting vastly more transitions than present in the labeled data set.

FIGURE 9 | For the bow shock transitions, comparison of the numbers of crossings from the labeled data set **(A)** and those predicted by the RF MAG model **(B)**, RF MAG & full particle model **(C)** and RNN 60-min model **(D)** are shown in the top grouping of panels. The numbers of crossings were computed per each continuous segment of test data and grouped by month. Panels **(E–H)** show the corollary for the magnetopause crossings, with the number of actual magnetopause crossings shown in **(E)**. The number of actual crossings is then compared with predicted crossings from the RF MAG model **(F)**, RF MAG & full ion **(G)** and RNN 60-min models **(H)**. The vertical scaling of panels **(A,E)** are significantly smaller than the other panels to allow for ease of comparison. All models over-predict the number of crossings, however we find of the three shown that the RNN 60-min model has the best performance, with the least amount of crossings predicted and closest crossing type alignment to the true crossings shown in **(A)** and **(E)**.

TABLE 4 | Performance of the RF MAG, RF MAG & full particle, and 60-min RNN model at correctly detecting labeled boundary crossings. A boundary crossing was considered "matched" if there was the same type of boundary in the model predictions within one hour of the labeled crossing. The mean time offset of the matched boundaries is positive if the detected boundary crossing occurred after the labeled boundary (i.e., the model was delayed in its prediction). Noted is the shorter length of the RF MAG & full particle test set (extending only through 2012) and fewer labeled boundary crossings.

Parameter	Model Type	BSO	BSI	MPO	MPI
Matched Boundaries (% Total)	RF MAG	76 (79.2%)	72 (75.8%)	120 (81.1%)	126 (84.0%)
	RF MAG & Full Part.	25 (73.5%)	30 (75.0%)	91 (85.0%)	87 (79.8%)
	RNN 60-Min	77 (90.6%)	78 (91.8%)	102 (78.5%)	114 (85.7%)
Unmatched Boundaries (% Total)	RF MAG	20 (20.8%)	19 (20.0%)	28 (18.9%)	24 (16%)
	RF MAG & Full Part.	9 (26.5%)	10 (25.0%)	16 (15.0%)	22 (20.2%)
	RNN 60-Min	8 (9.4%)	7 (8.2%)	28 (21.5%)	19 (14.3%)
Mean time offset to matched boundary (median) (min)	RF MAG	+ 7.33 ± 14.62 (+ 7)	− 0.58 ± 13.67 (+ 3)	+ 5.24 ± 18.34 (+ 6.5)	+ 2.97 ± 18.05 (+ 5)
	RF MAG & Full Part.	+ 6.52 ± 10.44 (+ 7)	+ 5.03 ± 9.69 (+ 5)	+ 3.75 ± 14.70 (+ 5)	+ 1.67 ± 15.70 (+ 4)
	RNN 60-Min	+ 11.41 ± 15.23 (+ 8.5)	+ 3.55 ± 13.00 (+ 3.5)	+ 13.93 ± 17.54 (+ 11.5)	+ 7.18 ± 17.48 (+ 7)

However, comparisons between the labeled transitions and the RNN-predicted transitions qualitatively reveal that the general trends of BSI/BSO-dominant periods (such as 2011–2012) versus MPI/MPO-dominant periods (2010) seen in the labeled data set are echoed in the model results, giving confidence that the model is capturing the underlying physics of the system. We also note that the Jackman et al. (2019) list, upon which this supervised learning approach is based, was formulated to capture the clearest and longest duration boundary crossings, and was not optimised to select multiple short-duration (2–3 min) crossings. While the aim of our ML approach is to determine what method best classifies the bulk of the regions, and the models demonstrate proficiency at doing so, the multiple short-duration "false" crossings predicted by the models could be actual phenomena (e.g. boundary-layer dynamics) that are not fully labeled and thus require further investigation (see **Supplemental Material**). In our investigation, the prediction for a particular sample was taken as the maximum of the algorithm confidence in the three regions, as is standard practice in the machine learning community. However, examining the algorithm confidence in the three regions rather than the maximum, as well as the inter-sample variance in the confidence, could eliminate many "false" crossings as well as highlight the need for SITL-intervention in the case where confidence in any one particular region is not high.

3.5 Derived Boundary Crossings

To understand whether the boundary crossings identified by the temporal error analysis aligned with those in our labeled data set, we analyzed each of the model's boundary crossings shown in **Figure 9** to see if they were a "matched" event (coincided with a boundary crossing of the same type identified in the labeled data set), an "unmatched" event (a labeled boundary which did not have a corresponding match in the model's boundary crossings), or a "False Boundary," (FB) i.e., a model boundary without a corresponding match in the labeled boundary list. A model-identified boundary crossing would be considered a "match" if it occurred within an hour before or after a list boundary crossing of the same type. In the case that the model identified multiple boundary crossings of the same time within the +/− hour span surrounding a labeled event, we chose the model crossing that was closest in absolute time. **Table 4** shows the results for the RF MAG, RF MAG & full particle and RNN 60-min MAG model. For labeled boundaries which were matched to a model prediction, the time difference between the model-predicted boundary and the true boundary was calculated, with a positive difference indicating that the model transition occurred after the list transition (i.e., the model was delayed).

In general, we find that all the models perform relatively well at identifying the crossings manually identified by Jackman et al. (2019), however, there were a high number of FBs across all models and all boundary types. The number of FBs was especially pronounced for the RF MAG model, echoing the large amount of spatial and temporal variability in model predictions seen in **Figures 7**,**9**, respectively. The RNN MAG model, which covers the same duration of the mission as the RF MAG model (2004–2016), observes much fewer FBs, particularly of BSO and BSI transitions. The RF MAG & full particle model observes much fewer FBs than the MAG-only RF model, likely as a consequence of the addition of the CAPS/IMS and MIMI/CHEMS/LEMMS data. Relative to the total number of test samples provided to the respective models, the RNN model shows the least FBs by far, indicating it has far more temporal accuracy and consistency than the RF approach. Investigating the RNN performance more closely, an interesting observation is the greater lag observed on outward transitions (BSO and MPO) as opposed to inward transitions (BSI and MPI), as well as a greater lag observed at the magnetopause transitions as opposed to the bow shock transitions. The lag suggests that the model needs at least a few minutes of data from the new region before it is able to shift its prediction, with the running variance and mean of the features within the new region "learned" by the model. Therefore, it can be assumed that RNN-based approaches for predicting region transitions may lag on their exact prediction of the boundary crossing, particularly when the boundary between the regions is only subtly hinted at by the behavior of the features. This is especially the case at the magnetopause boundary, where the transition between the magnetosheath

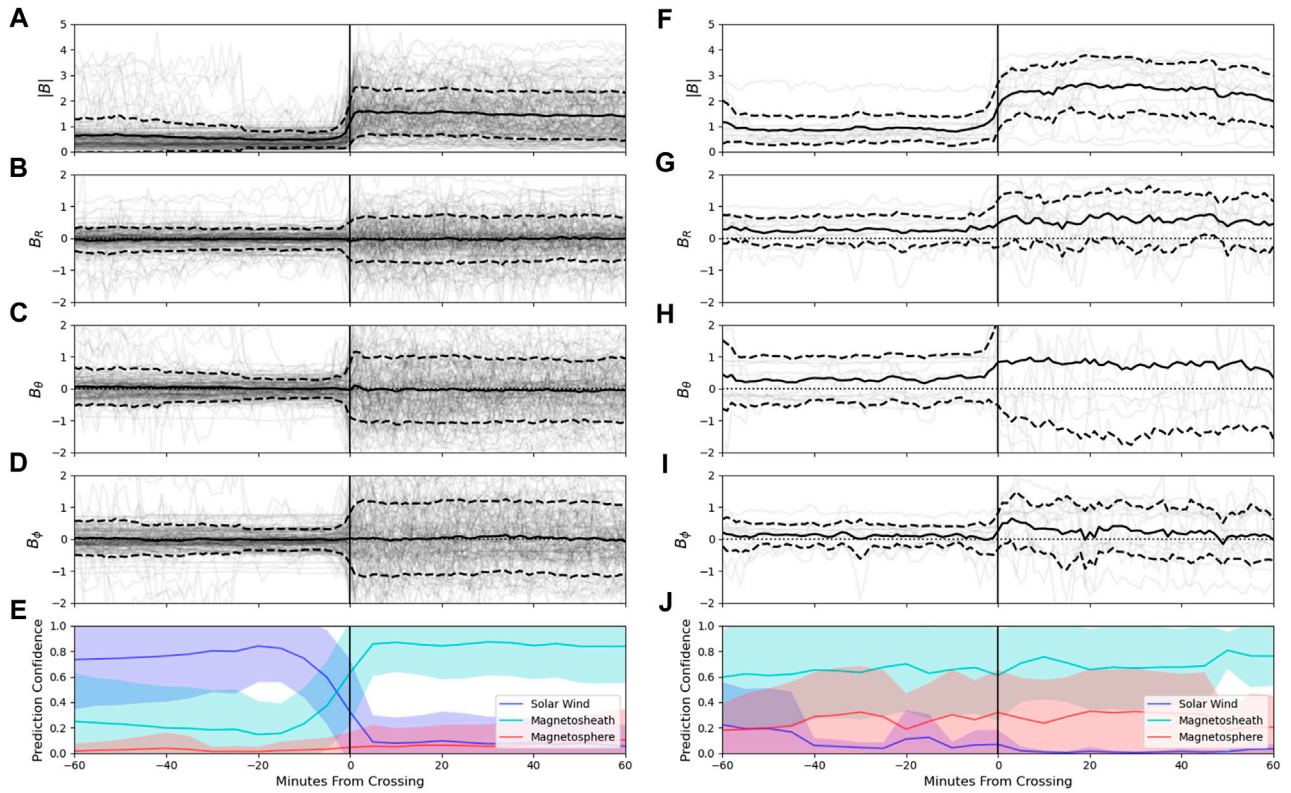

FIGURE 10 | Comparison of the bow shock crossings in the labeled data set which were matched to a predicted crossing **(A-E)**, versus those which were not matched **(F–J)** in the RNN 60-min model predictions. Panels **(A–D)** and **(F–I)** show the magnetic field conditions in a 1-h vicinity surrounding the labeled crossing, with individual instances plotted as transparent black lines. The average conditions ± the standard deviation are shown as thick black solid and dotted lines, respectively. Panels **(E,J)** show the prediction confidence of the model surrounding the labeled crossing for matched **(E)** and unmatched **(J)** crossings, with the average ± the standard deviation shown in the shaded region. Here the inwards and outwards crossings have been overlaid, such that all crossings are oriented in a inwards trajectory. In total there were 155 matched crossings and 15 unmatched crossings (see **Table 4**).

and magnetopause can be somewhat ambiguous when using the MAG data alone for times of small magnetic shear and/or highly turbulent boundary layers. In contrast, a sharp difference between the solar wind and the magnetosheath is typically observed, particularly in the enhancement of the running variance across all MAG field components as we move into the magnetosheath. We see that consequently the BSI transition appears to be the easiest transition for the RNN model to discern, with a lag of only ≈3.6 min.

It should also be noted that the five-minute stride present within the test sample data for the RNN to prevent over sampling means that the "labeled" boundary may be slightly offset from the "true" boundary depending on whether the timing of the "true" boundary falls on the same sample cadence of the test data. In cases where the true boundary timing does not directly coincide with a test sample, the nearest following sample was indicated as the location of the boundary, which was at most 4 min away from the true boundary location. For the RF models, the sampling cadence of 10 min results in the model's first sample within a new region being at most 9 min away from the true boundary. A secondary point to note is how the model results are interpreted, which impacts the determination of predicted boundary crossings. The output of the models is a three component vector, representing the model's confidence in each of the three regions; the maximum of these three components is interpreted as the model's predicted region. The confidence in a particular region would have to exceed 0.33 before it is interpreted as the current region, yet the model's confidence in a region would increase prior to it becoming the dominant region. Therefore, the lag in recognizing a transition may not be as severe as suggested when we only interpret the maximum as the model's prediction, since investigating the individual confidence levels may reveal an increasing trend in a particular region before it overtakes the confidence levels of the other regions and becomes the maximum.

3.5.1 Epoch Analysis

Again focusing only on the RNN 60-min model, **Figures 10**,**11** show the corresponding behavior of the magnetic field features at the bow shock and magnetopause boundary crossings, respectively for both the matched and unmatched boundaries. Outward transitions (i.e., the spacecraft is encountering the boundary on an outward trajectory) and inward transitions are overlaid in the figures, such that all transitions are oriented to be inwards. The sharp division between the solar wind and magnetosheath is present in **Figure 10**, with the cross

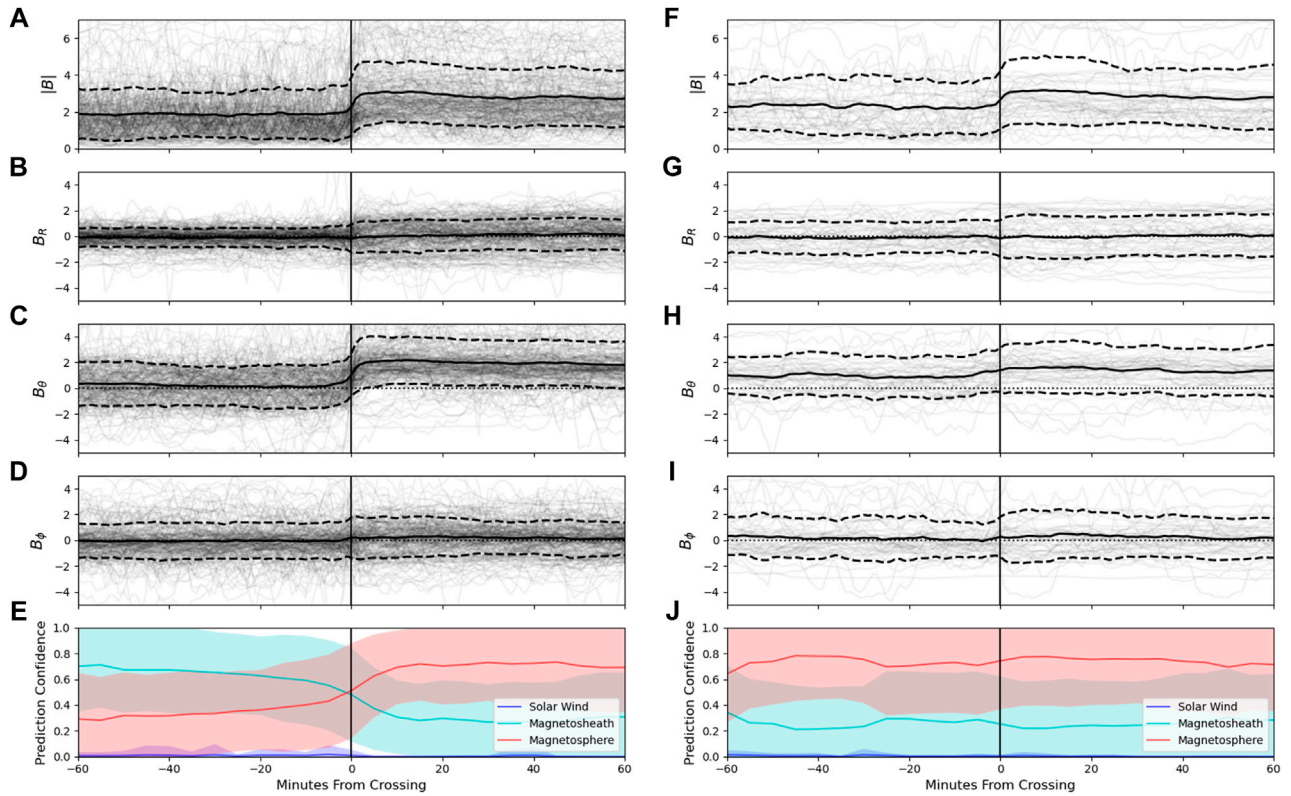

FIGURE 11 | Comparison of the magnetopause crossings in the labeled data set which were matched to a predicted crossing **(A–E)**, versus those which were not matched **(F–J)** in the RNN 60-min model predictions. Panels **(A–D)** and **(F–I)** show the magnetic field conditions in a 1-h vicinity surrounding the labeled crossing, with individual instances plotted as transparent black lines. The average conditions ± the standard deviation are shown as thick black solid and dotted lines, respectively. Panels **(E,J)** show the prediction confidence of the model surrounding the labeled crossing for matched **(E)** and unmatched **(J)** crossings, with the average ± the standard deviation shown in the shaded region. Here the inwards and outwards crossings have been overlaid, such that all crossings are oriented in a inwards trajectory. In total there were 216 matched crossings and 47 unmatched crossings (see **Table 4**).

over into the magnetosheath resulting in a much more variable and higher magnitude magnetic field. As indicated in **Table 4**, we see that the RNN 60-min model is easily able to detect the BS in most cases as revealed by the high confidence levels in the solar wind and magnetosheath before and after the transition, respectively (**Figure 10E**). In the few cases (N = 7 for BSI, N = 8 for BSO) when the BS was missed, we see that it was because the model failed to register it was in the solar wind before the transition, seemingly due to elevated B_θ values.

The boundary between the magnetosheath and the magnetosphere is much more subtle than that between the solar wind and magnetosheath as revealed in **Figure 11**. The subtle nature of the boundary is underscored by the greater percentage of missed MPI (14.3%) and MPO (21.1%) transitions relative to the BSI (9.1%) and BSO (10.4%) transitions (see **Table 4**), and the greater delay in the matched transitions from the timing of the actual boundary crossing and the model detection of the new region. The MP crossings that were successfully identified demonstrate a sharp increase in $|B|$ as the spacecraft moves into the magnetosphere, which is principally driven by an increase in B_θ. The missed MP transitions show a slightly more gradual increase in $|B|$ and particularly in B_θ, with the model failing to recognize the magnetosheath is present before the transition. For all four boundary transition types, we see that the missed transitions exhibit confusion mainly between the magnetosphere and the magnetosheath, even for bow shock boundaries.

4 CONCLUSION

Here we have found that a variety of ML algorithms are capable of producing relatively accurate classifications of the region the spacecraft is inhabiting using only instrument data as the model input. Architecting the problem as a region-classification task instead of attempting to directly classify the boundary crossings afforded a much larger data set for both training and testing, enabling a broader swath of algorithms to be explored. However, as a consequence, assessment of where and how well the model predicted boundary crossings required a more-complicated post processing methodology and ultimately led to a large number of FBs. Daigavane et al. (2020) performed a complementary study in which they attempted to directly detect magnetopause and bow shock crossings in the CAPS-ELS data set

using an anomaly detection methodology. Similar to the results contained herein, they found that bow shock crossings were substantially easier to detect than magnetopause crossings.

Comparing the predictive value of different feature sets, as was possible with the simpler and less data-intensive RF models, we find that the inclusion of more features clearly increases the predictive capability of the model, as expected. It should be noted that the specific subset of features from the plasma data chosen is important for this type of classification scheme. Ultimately, the best models will have inputs derived from the most physically relevant measurements, which given the architecture of this problem would be those features showing distinctly different characteristics in the bulk regions. We find that ultimately a time-series-based approach, as is possible with the RNN LSTM algorithm, produces a model with the greatest accuracy and temporal consistency, indicating that the temporal trends and variances of the MAG data alone provides sufficient predictive capability. This is further underscored by the improvement in the model performance as the length of the time sample fed to the RNN models is increased from 20 to 40 and, finally, 60 min. Though outside the scope of this study, we urge future studies to consider algorithm approaches which can leverage the benefits of both time-variance of the features and a richer feature set encompassing multiple instruments. While the scope of the algorithms explored in this study was relatively limited, other algorithms such as 1-D Convolutional Neural Networks (CNNs) or hybrid CNN-LSTM architectures should be explored given their utility in other sequence classification tasks, such as natural language processing and speech recognition (Sainath et al., 2015; Yin et al., 2017).

We have also shown the necessity of doing a full error analysis of the results and expand beyond the scope of analysis typically done in multi-class ML classification tasks. Blanket accuracy metrics fail to measure the algorithm prediction consistency over temporal or spatial scales. Nor do such metrics capture the feature context leading to model errors, or attempt to elucidate whether model predictions are tied to particular physical phenomena. By investigating the errors on spatial and temporal scales, we have found that models only slightly different in their overall accuracy metrics have demonstrably different performance in terms of temporal or spatial cohesion. The RF models in particular are only slightly worse than the RNN 60-min model in terms of their overall accuracy, and yet their predictions exhibit much more temporal volatility and undesirable patterns in spatial errors.

4.1 Implication for On-Board AI Utilization on Future Space Missions

Given that there was an intentional decision to not apply filtering or smoothing techniques such as a centered running mean to the data prior to implementing the ML methods, the algorithms presented here could be run in a real-time scenario (ignoring the computational limitations of current spacecraft). As such, the instability of the model output could be addressed in real-time by implementing a persistence counter, i.e., a prediction of a different region would have to persist for a set number of continuous samples before the model were to shift its predictions. Such persistence measures are already widely used in spacecraft fault management autonomy systems to prevent outlier measurements from driving operational fault containment measures to the detriment of science or broader mission objectives (Fesq, 2009). Similarly, a threshold on the model's confidence in a particular region needed before shifting the region prediction from one region to another, as would be the case in a boundary crossing, could be implemented. Both of these measures would reduce the rapid, and likely incorrect, false boundary crossings observed here—reducing risk with the side effect of potentially lengthening the lag between the true boundary crossing and the model's recognition of the boundary crossing.

As noted by several studies (Azari et al., 2020; Hook et al., 2020; Theiling et al., 2021; Vandegriff et al., 2021), current missions are already facing severe downlink constraints and more data-intensive sensors. Without increased capabilities in on-board storage and deep space communications, missions may ultimately require the use of on-board autonomy to sift through the deluge of collected data to prioritize the most relevant observations for downlink or optimize the science collection of the sensors for the environment the spacecraft or lander is currently inhabiting. Examples of automated decisions the spacecraft could complete with on-board AI could be changing the sampling rate of an instrument or changing the binning scheme of plasma data. Already, research is being done to optimize data downlink using AI on earth-orbiting missions such as MMS, where only 4% of the high-rate data collected daily can be sent to the ground (Argall et al., 2020). In these cases, where predictions from an on-board AI system could contribute to mission operations, model stability becomes critical else undue risk is embedded in the mission. The results shown here illustrate that while simpler algorithms such as a RF can replicate the overall accuracy of more complicated RNNs and are more apt for on-board application due to their ability to operate in low-Size, Weight and Power (SWaP) embedded applications, they fail to replicate the accuracy and temporal stability of neural network approaches. Therefore, assessments of candidate algorithm performance must not only assess model performance using an unbiased, representative test set but also fully evaluate the context of the predictions.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://pds-ppi.igpp.ucla.edu/mission/Cassini-Huygens.

AUTHOR CONTRIBUTIONS

KY and JV conceptualized the study. KY developed the models and resulting analysis of model predictions. All authors

contributed to the interpretation of the results and writing of the manuscript.

FUNDING

KY, JV, GC, SV, and PK were supported by an JHU APL internal research and development grant, funded by the Civil Space Mission Area within the Space Exploration Sector. KY and JV were additionally supported by NASA Cassini/MIMI mission contract NNN06AA01C. GC was additionally supported by NASA's Cassini Data Analysis program under grant 80NSSC18K1234, which also partially supported KY and SV was additionally supported by NASA grant 80NSSC19K0899. TG's work at the University of Southampton was supported by the Alan Turing Institute and the Science and Technology Facilities Council (STFC). CJ's work at DIAS was supported by the Science Foundation Ireland (SFI) Grant 18/FRL/6199. AS was supported by STFC Consolidated Grant ST/S000240/1 and NERC grants NE/P017150/1 and NE/V002724/1.

ACKNOWLEDGMENTS

We are thankful to the institutions and personnel that made the Cassini-Huygens mission and the associated instruments a success.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fspas.2022.875985/full#supplementary-material

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. doi:10.48550/arXiv.1603.04467

Argall, M. R., Small, C., Piatt, S., Breen, L., Petrik, M., Kokkonen, K., et al. (2020). *Mms Sitl Ground Loop: Automating the Burst Data Selection Process*. doi:10.3389/fspas.2020.00054

Arridge, C. S., Jasinski, J. M., Achilleos, N., Bogdanova, Y. V., Bunce, E. J., Cowley, S. W. H., et al. (2016). Cassini Observations of Saturn's Southern Polar Cusp. *J. Geophys. Res. Space Phys.* 121, 3006–3030. doi:10.1002/2015ja021957

Azari, A. R., Biersteker, J. B., Dewey, R. M., Doran, G., Forsberg, E. J., Harris, C. D. K., et al. (2020). *Integrating Machine Learning for Planetary Science: Perspectives for the Next Decade*. doi:10.48550/arXiv.2007.15129

Bertucci, C., Achilleos, N., Mazelle, C., Hospodarsky, G., Thomsen, M., Dougherty, M., et al. (2007). Low-frequency Waves in the Foreshock of Saturn: First Results from Cassini. *J. Geophys. Res. Space Phys.* 112. doi:10.1029/2006ja012098

Camporeale, E., Carè, A., and Borovsky, J. E. (2017). Classification of Solar Wind with Machine Learning. *J. Geophys. Res. Space Phys.* 122 (10), 910–920. doi:10.1002/2017JA024383

Case, N. A., and Wild, J. A. (2013). The Location of the Earth's Magnetopause: A Comparison of Modeled Position and *In Situ* Cluster Data. *J. Geophys. Res. Space Phys.* 118, 6127–6135. doi:10.1002/jgra.50572

Chicco, D., and Jurman, G. (2020). The Advantages of the Matthews Correlation Coefficient (Mcc) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC genomics* 21, 6–13. doi:10.1186/s12864-019-6413-7

Dougherty, M. K., Achilleos, N., Andre, N., Arridge, C. S., Balogh, A., Bertucci, C., et al. (2005). Cassini Magnetometer Observations during Saturn Orbit Insertion. *Science* 307, 1266–1270. doi:10.1126/science.1106098

Fesq, L. M. (2009). "Current Fault Management Trends in Nasa's Planetary Spacecraft,", 2020 IEEE Aerospace Conference, Big Sky, MT, USA, 7-14 March 2020 (IEEE), 1–9. doi:10.1109/AERO.2009.4839530

Gorodkin, J. (2004). Comparing Two K-Category Assignments by a K-Category Correlation Coefficient. *Comput. Biol. Chem.* 28, 367–374. doi:10.1016/j.compbiolchem.2004.09.006

Guo, R. L., Yao, Z. H., Wei, Y., Ray, L. C., Rae, I. J., Arridge, C. S., et al. (2018). Rotationally Driven Magnetic Reconnection in Saturn's Dayside. *Nat. Astron* 2, 640–645. doi:10.1038/s41550-018-0461-9

Hook, J. V., Castillo-Rogez, J., Doyle, R., Vaquero, T. S., Hare, T. M., Kirk, R. L., et al. (2020). "Nebulae: A Proposed Concept of Operation for Deep Space Computing Clouds,". 2020 IEEE Aerospace Conference, Big Sky, MT, USA, 7-14 March 2020 (IEEE), 1–14. doi:10.1109/AERO47225.2020.9172264

Ivchenko, N. V., Sibeck, D. G., Takahashi, K., and Kokubun, S. (2000). A Statistical Study of the Magnetosphere Boundary Crossings by the Geotail Satellite. *Geophys. Res. Lett.* 27, 2881–2884. doi:10.1029/2000gl000020

Jackman, C., Forsyth, R., and Dougherty, M. (2008). The Overall Configuration of the Interplanetary Magnetic Field Upstream of Saturn as Revealed by Cassini Observations. *J. Geophys. Res. Space Phys.* 113. doi:10.1029/2008ja013083

Jackman, C. M., Achilleos, N., Bunce, E. J., Cowley, S. W. H., Dougherty, M. K., Jones, G. H., et al. (2004). Interplanetary magnetic field at 9 au during the declining phase of the solar cycle and its implications for saturn's magnetospheric dynamics. *J. Geophys. Res. Space Phys.* 109. doi:10.1029/2004JA0106110.1029/2004ja010614

Jackman, C. M., Thomsen, M. F., and Dougherty, M. K. (2019). Survey of Saturn's Magnetopause and Bow Shock Positions over the Entire Cassini Mission: Boundary Statistical Properties and Exploration of Associated Upstream Conditions. *J. Geophys. Res. Space Phys.* 124, 8865–8883. doi:10.1029/2019JA026628

Jasinski, J. M., Arridge, C. S., Coates, A. J., Jones, G. H., Sergis, N., Thomsen, M. F., et al. (2016). Cassini Plasma Observations of Saturn's Magnetospheric Cusp. *J. Geophys. Res. Space Phys.* 121, 12–047. doi:10.1002/2016ja023310

Jasinski, J. M., Arridge, C. S., Coates, A. J., Jones, G. H., Sergis, N., Thomsen, M. F., et al. (2017). Diamagnetic Depression Observations at Saturn's Magnetospheric Cusp by the Cassini Spacecraft. *J. Geophys. Res. Space Phys.* 122, 6283–6303. doi:10.1002/2016ja023738

Jelínek, K., Němeček, Z., and Šafránková, J. (2012). A New Approach to Magnetopause and Bow Shock Modeling Based on Automated Region Identification. *J. Geophys. Res.* 117, a–n. doi:10.1029/2011JA017252

Kanani, S. J., Arridge, C. S., Jones, G. H., Fazakerley, A. N., McAndrews, H. J., Sergis, N., et al. (2010). A New Form of Saturn's Magnetopause Using a Dynamic Pressure Balance Model, Based on *In Situ*, Multi-Instrument Cassini Measurements. *J. Geophys. Res.* 115, a–n. doi:10.1029/2009JA014262

Kingma, D. P., and Ba, J. (2017). *Adam: A Method for Stochastic Optimization*. doi:10.48550/arXiv.1412.6980

Krimigis, S. M., Mitchell, D. G., Hamilton, D. C., Livi, S., Dandouras, J., Jaskulek, S., et al. (2004). *Magnetosphere Imaging Instrument (MIMI) on the Cassini Mission to Saturn/Titan*. Dordrecht: Springer Netherlands, 233–329. doi:10.1007/978-1-4020-2774-1_3

Liou, K., Paranicas, C., Vines, S., Kollmann, P., Allen, R. C., Clark, G. B., et al. (2021). Dawn-dusk Asymmetry in Energetic (> 20 Kev) Particles Adjacent to Saturn's Magnetopause. *J. Geophys. Res. Space Phys.* 126, e2020JA028264. doi:10.1029/2020ja028264

Matthews, B. W. (1975). Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochimica Biophysica Acta (BBA) - Protein Struct.* 405, 442–451. doi:10.1016/0005-2795(75)90109-9

Mauk, B. H., Cohen, I. J., Haggerty, D. K., Hospodarsky, G. B., Connerney, J. E. P., Anderson, B. J., et al. (2019). Investigation of Mass-/Charge-Dependent Escape of Energetic Ions across the Magnetopauses of Earth and Jupiter. *J. Geophys. Res. Space Phys.* 124, 5539–5567. doi:10.1029/2019JA026626

Mauk, B. H., Cohen, I. J., Westlake, J. H., and Anderson, B. J. (2016). Modeling Magnetospheric Energetic Particle Escape across Earth's Magnetopause as Observed by the MMS Mission. *Geophys. Res. Lett.* 43, 4081–4088. doi:10.1002/2016gl068856

Olshevsky, V., Khotyaintsev, Y. V., Lalti, A., Divin, A., Delzanno, G. L., Anderzén, S., et al. (2021). Automated Classification of Plasma Regions Using 3d Particle Energy Distributions. *J. Geophys. Res. Space Phys.* 126, e2021JA029620. doi:10.1029/2021ja029620

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Pilkington, N. M., Achilleos, N., Arridge, C. S., Guio, P., Masters, A., Ray, L. C., et al. (2015). Internally Driven Large-Scale Changes in the Size of Saturn's Magnetosphere. *J. Geophys Res. Space Phys.* 120, 7289–7306. doi:10.1002/2015JA021290

Sainath, T. N., Vinyals, O., Senior, A., and Sak, H. (2015). Convolutional, Long Short-Term Memory, Fully Connected Deep Neural Networks , 4580–4584. doi:10.1109/icassp.2015.7178838

Simon, S., Roussos, E., and Paty, C. S. (2015). The Interaction between Saturn's Moons and Their Plasma Environments. *Phys. Rep.* 602, 1–65. doi:10.1016/j.physrep.2015.09.005

Sulaiman, A. H., Masters, A., and Dougherty, M. K. (2016). Characterization of Saturn's Bow Shock: Magnetic Field Observations of Quasi-Perpendicular Shocks. *J. Geophys. Res. Space Phys.* 121, 4425–4434. doi:10.1002/2016ja022449

Theiling, B., Brinckerhoff, W., Castillo-Rogez, J., Chou, L., Poian, V. D., Graham, H., et al. (2021). Non-robotic Science Autonomy Development. *Bull. AAS* 53. doi:10.3847/25c2cfeb.ee4e6b64

Vandegriff, J., Smith, B., Yeakel, K., Vines, S., Ho, G., Clark, G., et al. (2021). *Developing Smarter Techniques to Deal with the Heliophysics Science Data Flood.*

Went, D. R., Hospodarsky, G. B., Masters, S., Hansen, K. C., and Dougherty, M. K. (2011). A New Semiempirical Model of Saturn's Bow Shock Based on Propagated Solar Wind Parameters. *J. Geophys. Res. Space Phys.* 116. doi:10.1029/2010ja016349

Xu, F., and Borovsky, J. E. (2015). A New Four-Plasma Categorization Scheme for the Solar Wind. *J. Geophys. Res. Space Phys.* 120, 70–100. doi:10.1002/2014JA020412

Yin, W., Kann, K., Yu, M., and Schütze, H. (2017). *Comparative Study of Cnn and Rnn for Natural Language Processing*. doi:10.48550/arXiv.1702.01923

Young, D. T., Berthelier, J. J., Blanc, M., Burch, J. L., Coates, A. J., Goldstein, R., et al. (2004). *Cassini Plasma Spectrometer Investigation*. Dordrecht: Springer Netherlands, 1–112. doi:10.1007/978-1-4020-2774-1_1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Astronomy and Space Sciences

ORIGINAL RESEARCH
published: 25 May 2022
doi: 10.3389/fspas.2022.869740

Revisiting the Ground Magnetic Field Perturbations Challenge: A Machine Learning Perspective

Victor A. Pinto[1], Amy M. Keesee[1,2], Michael Coughlan[2], Raman Mukundan[2], Jeremiah W. Johnson[3], Chigomezyo M. Ngwira[4] and Hyunju K. Connor[5,6]*

[1]Institute for the Study of Earth, Oceans and Space, University of New Hampshire, Durham, NH, United States, [2]Department of Physics and Astronomy, University of New Hampshire, Durham, NH, United States, [3]Department of Applied Engineering and Sciences, University of New Hampshire, Manchester, NH, United States, [4]Orion Space Solutions, Louisville, CO, United States, [5]NASA Goddard Space Flight Center, Greenbelt, MD, United States, [6]Geophysical Institute, University of Alaska Fairbanks, Fairbanks, AK, United States

Forecasting ground magnetic field perturbations has been a long-standing goal of the space weather community. The availability of ground magnetic field data and its potential to be used in geomagnetically induced current studies, such as risk assessment, have resulted in several forecasting efforts over the past few decades. One particular community effort was the Geospace Environment Modeling (GEM) challenge of ground magnetic field perturbations that evaluated the predictive capacity of several empirical and first principles models at both mid- and high-latitudes in order to choose an operative model. In this work, we use three different deep learning models-a feed-forward neural network, a long short-term memory recurrent network and a convolutional neural network-to forecast the horizontal component of the ground magnetic field rate of change (dB_H/dt) over 6 different ground magnetometer stations and to compare as directly as possible with the original GEM challenge. We find that, in general, the models are able to perform at similar levels to those obtained in the original challenge, although the performance depends heavily on the particular storm being evaluated. We then discuss the limitations of such a comparison on the basis that the original challenge was not designed with machine learning algorithms in mind.

Keywords: geomagnetically induced currents, deep learning, ground magnetic disturbance, space weather, neural network

OPEN ACCESS

Edited by:
Peter Wintoft, Swedish Institute of Space Physics, Sweden

Reviewed by:
Simon Wing, Johns Hopkins University, United States
Stefano Markidis, KTH Royal Institute of Technology, Sweden

****Correspondence:***
Victor A. Pinto
victor.pinto@gmail.com

Specialty section:
This article was submitted to Space Physics, a section of the journal Frontiers in Astronomy and Space Sciences

Received: *04 February 2022*
Accepted: *04 May 2022*
Published: *25 May 2022*

Citation:
Pinto VA, Keesee AM, Coughlan M, Mukundan R, Johnson JW, Ngwira CM and Connor HK (2022) Revisiting the Ground Magnetic Field Perturbations Challenge: A Machine Learning Perspective. Front. Astron. Space Sci. 9:869740. doi: 10.3389/fspas.2022.869740

1 INTRODUCTION

Horizontal magnetic field variations (dB_H/dt) derived from ground magnetometer recordings have been utilized commonly as a proxy for evaluating the risk that geomagnetically induced currents (GIC) present in different regions (e.g., Viljanen et al., 2001; Pulkkinen et al., 2015; Ngwira et al., 2018). GICs occur in ground-level conductors following an enhancement of the geoelectric field on the ground, usually in association with active geomagnetic conditions (Ngwira et al., 2015; Gannon et al., 2017), and have been known to cause damage to power transformers, corrode pipelines, and interfere with railway signals (Pirjola, 2000; Boteler, 2001; Pulkkinen et al., 2017; Boteler, 2019). As our society continues to become more "technology dependent" and as we enter a new cycle of intense geomagnetic activity during the ascending and maximum phases of solar cycle 25, having the appropriate tools to assess the risk GICs pose to different regions becomes urgently relevant (Oughton et al., 2019; Hapgood et al., 2021).

GIC levels are dependent on the characteristics of the system they affect as well as the environmental conditions, and unfortunately, measured GIC data are rarely available to the scientific community as they are either not monitored or the measurements restricted by power operator and therefore not made public. For this reason, variations of the measured ground magnetic field are commonly used as proxy to estimate the risk of GIC occurrence (Viljanen, 1998; Viljanen et al., 2001; Wintoft, 2005; Dimmock et al., 2020). These variations can be utilized to calculate the geoelectric field in regions where the ground conductivity profile is available (Love et al., 2018; Lucas et al., 2020; Gil et al., 2021).

In the past, many attempts have been made to forecast dB_H/dt with different degrees of success, using first-principles and empirical models (e.g., Tóth et al., 2014; Wintoft et al., 2015). However, comparisons are rarely made between models, in part because most models are not meant to be deployed for operational purposes, but also because models have different general forecasting objectives. The Geospace Environment Modeling (GEM) challenge (Pulkkinen et al., 2013) that ran during the years 2008–2012 tried to provide a direct comparison between models and to choose a model for real-time forecasting. It involved the entire space weather community in order to come up with a standardized method to test models against each other, and from there select a model to be transitioned into operation at NOAA (Pulkkinen et al., 2013).

Recently, machine learning empirical models have become more common thanks in part to the increased availability of data for training and the improvement of open-source machine learning tools (e.g. Keesee et al., 2020). Machine learning models present the advantage that, once trained, execution time is extremely low, and as such, they are able to deploy for real-time forecasting with extremely low computational cost. But while machine-learned models are able to forecast dB_H/dt or even GICs when data is available to different degrees of success, few attempts have been made to evaluate them on the grounds of established benchmarks. It is within that framework that we attempt to evaluate a series of machine learning models with the same metrics used by the GEM Challenge. In **Section 2** we describe the GEM challenge in detail as well as the datasets we utilized and the models we developed. **Section 3** presents the results of our models in the context of the GEM challenge metrics. In **Section 4** we discuss the main challenges and lessons from our model development and comparisons. Finally, **Section 5** presents our summary and conclusions.

2 DATA AND METHODOLOGY

The Geospace Environment Modeling (GEM) ground magnetic field perturbations challenge ("the GEM challenge") consisted of a multi-year community effort that ran roughly between 2008 and 2011 with the objective of testing, comparing, and eventually delivering a model to be used at National Oceanic and Atmospheric Administration (NOAA) Space Weather Prediction Center (SWPC). The final results, description and evaluations of the different models that participated in the challenge are described in depth by Pulkkinen et al. (2013). The purpose of this study is to evaluate our machine learning based models using the same conditions and test on the same benchmarks, only deviating when an exact replication is not possible. The GEM challenge (and therefore the work presented here) consisted of forecasting the 1-min resolution of the horizontal component of ground magnetic field perturbations at several mid- and high-latitude stations. The horizontal component H is defined by

$$\frac{dB_H}{dt} = \sqrt{\left(\frac{dB_N}{dt}\right)^2 + \left(\frac{dB_E}{dt}\right)^2} \tag{1}$$

where E represents the east-west component, and N the north-south component in magnetic coordinates. The choice of forecasting the horizontal fluctuations is based on the assumption that it is the most important component for GIC occurrence (Pirjola, 2002). Although the GEM challenge involved a total of 12 different ground magnetometer stations during its different stages, the final evaluation presented in Pulkkinen et al. (2013) was performed only on 6 of them. Because the published scores are only available for those six stations, they will be the focus of this study. **Table 1** lists the ground magnetometer stations, their code name and their magnetic latitude and longitude. Note that SNK replaced PBQ after 2007, so those data serve as a single location.

The GEM challenge proposed a unique and interesting evaluation mechanism. The models forecast four known geomagnetic storms during the testing period, and two extra storms were added as "surprise events" during the final evaluation. **Table 2** presents the six storms used in the evaluation of the models. Our first deviation from the original challenge is that we are not evaluating our models on unknown storms—we have only calculated the final scores of the six storms after our training of the models was complete, and therefore we did not perform tuning of the models after the evaluation. The model output is the 1-min resolution horizontal component dB_H/dt predicted 1 minute ahead of time. This is counted from the time of arrival of the solar wind to the bow-shock nose, which involves a propagation from the L1 monitors. Once the forecast is done, the 1-min resolution predictions are reduced to obtain the maximum dB_H/dt value every 20 min. Each 20-min window prediction is then evaluated against four different thresholds set up at 18, 42, 66, and 90 nT/min. This approach turns the challenge into a classification problem, and a contingency table can be made for each of the thresholds counting true positives (hits), true negatives (no crossings), false positives (false alarms) and false negatives (misses). From this contingency table the values of probability of detection, probability of false detection, and the Heidke Skill Score are calculated. The definitions can be found in Pulkkinen et al. (2013). To obtain each model performance, the contingency tables are added by grouping the mid-latitude stations together (NEW, OTT, WNG) and the high-latitude stations together (ABK, PBQ/SNK, YKC) for each of the events and each of the thresholds.

TABLE 1 | Ground magnetometer stations used in this study and their location. Stations PBQ and SNK (in bold) are complementary as one replaces the other after the year 2007.

Station name	Code	Geomagnetic latitude	Geomagnetic longitude
Abisko	ABK	65.74	101.7
Newport	NEW	54.65	–54.82
Ottawa	OTT	54.98	2.52
Poste-de-la-Baleine	**PBQ**	**65.01**	**0.2**
Sanikiluaq	**SNK**	**66.31**	**–1.92**
Wingst	WNG	50.15	86.75
Yellowknife	YKC	69.42	–56.85

TABLE 2 | Storms used for model evaluation.

Storm start date (UT)	Storm end date (UT)	Minimum D_{st} (nT)
2001-08-31 00:00	2001-09-01 00:00	–40
2003-10-29 06:00	2003-10-30 06:00	–353
2005-08-31 10:00	2005-09-01 12:00	–131
2006-12-14 12:00	2006-12-16 00:00	–139
2010-04-05 00:00	2010-04-06 00:00	–73
2011-08-05 09:00	2011-08-06 09:00	–113

2.1 Datasets and Pre-processing

For our study we have used the OMNI dataset obtained from the CDAWeb repository (https://cdaweb.gsfc.nasa.gov/pub/data/omni/omni_cdaweb/) at 1-min resolution. The OMNI database provides solar wind measurements obtained mostly from spacecraft located at the L1 Lagrangian point (~ $235R_E$ sunward of Earth) and then time-shifted to the magnetosphere's bow shock nose (King and Papitashvili, 2005). We train our models to forecast 1-min ahead of the current time on the OMNI dataset, however, this is equivalent to a 20–40 min lead time if we were using real-time data, depending on the solar wind speed. The benefits of using the OMNI dataset for training is that it is a well validated dataset that is readily available for anyone to use with minimal work involved, and as such, it increases the reproducibility of the results. For our study, we used data between (and including) January 1995 and December 2019.

The OMNI dataset provides both plasma and magnetic field parameters, as well as some derived physical quantities. It suffers from having significant gaps which amount to around 20% of missing data in the plasma parameters and around 7% of missing data in the magnetic field. Further exploration of the data shows that most of the gaps are relatively small, and therefore we have performed a linear interpolation in the magnetic field parameters for gaps of up to 10 min, and we have performed a linear interpolation with no limit on time of the plasma parameters, to fill any possible gap. The remaining gaps, as determined by the missing magnetic field data, are dropped from the training dataset.

The ground magnetic field perturbations from the six different stations were obtained from the SuperMAG 1-min resolution database (https://supermag.jhuapl.edu/) with baseline removed (Gjerloev, 2012). The data availability is high for all the studied stations, although there are some significant gaps in the SNK/PBQ set around the time of the replacement in 2007–2008. We have decided not to perform any interpolation in the magnetic field components and therefore all missing data points are excluded from the training. For training, we use the N and E components to obtain dB_H/dt (**Eq. 1**) and also the MLT position of the observatories from the SuperMAG data.

Given the nature of the system we are trying to predict, one of the issues we have encountered is that the magnetic field fluctuations are heavily biased towards 0 nT/min. That is, during quiet times, the fluctuations are relatively low, and they amount for a sizable portion of the available dataset. On the contrary, during active times, the fluctuations can easily go up to the hundreds of nT/min at least for high-latitude stations. To reduce the bias, we have decided to reduce our training samples to only those times in which a geomagnetic storm is occurring. To do this, we have identified all geomagnetic storms in the 1995–2018 period with SYM-H < -50 nT and we have selected for training the period between ±12 h around the minimum SYM-H value. **Figures 1A,B** show a visual representation of the effect of using only storm-time data. As can be appreciated for both the mid-latitude NEW station and the high-latitude YKC station, the restriction to storm-time only reduces the training dataset to ~ 10% of its original size eliminating mostly small fluctuations. From the histogram, it can also be observed that—especially at high-latitudes—some strong fluctuations do occur outside of the storm-time. Those cases can prove interesting for analysis in the future, but will not be further discussed in the context of this work. It is important to note that the six storms considered for testing have been removed from the storm dataset. A list with the storm dates can be found in the **Supplementary Material**.

To train the models we have decided to use the following solar wind parameters: solar wind speed (Vx, Vy, Vz), interplanetary magnetic field (B_T, B_y, B_z), proton density, solar wind dynamic pressure, reconnection electric field ($-VB_z$), and proton temperature. **Figure 1C** shows the absolute value of the maximum correlation coefficient between dB_H/dt and the different solar wind parameters for the previous 60 min (i.e., max correlation of dB_H/dt(t) with param(t), param (t-1), etc). The symbol corresponds to the average correlation over the six stations used in this study, and the bar corresponds to the range of correlations. Here it is important to note that some parameters are most likely contributing significantly more to the training process than others. We have decided to keep them all on the basis that the models can support the amount of input parameters.

2.2 Models

For the evaluation of the GEM Challenge scores we used three different deep learning models: a feed-forward fully connected artificial neural network (ANN), a long short-term memory recurrent neural network (LSTM) and a convolutional neural network (CNN). The election of those particular models offers a continuation to our previous modelling attempts of dB_H/dt using neural networks (ANN + LSTM) (Keesee et al., 2020) as well as to test the capabilities of convolutional neural networks after they

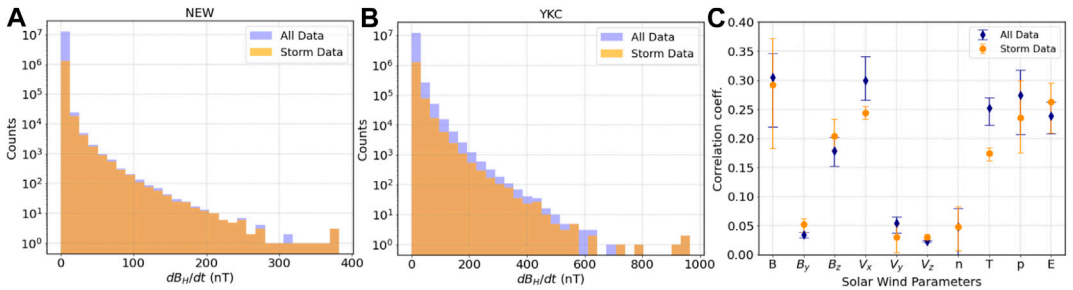

FIGURE 1 | Histogram of dB_H/dt for all data between 1995 and 2019 (blue) and storm-only data (orange) for **(A)** the mid-latitude station NEW and **(B)** the high-latitude station YKC. **(C)** Correlation coefficient between different solar wind parameters and dB_H/dt for all data (blue) and storm-only data (orange). The symbol indicates the average of all six stations, and bars represent the range of the individual stations.

have shown promise for time series forecasting in different Space Weather applications (e.g., Collado-Villaverde et al., 2021; Siciliano et al., 2021; Smith et al., 2021). The development and training of the models was done using the TensorFlow-Keras framework for *Python* (Abadi et al., 2016) as well as the scikit-learn toolkit (Pedregosa et al., 2011). All models used in this study were trained by minimizing the mean square error. This optimization was done in each case using the Adam optimization algorithm. Further description of each model is given in the next sections.

2.2.1 Artificial Neural Network

Fully-connected feed-forward neural networks can capture temporal behavior (similar to a recurrent neural network) if the time history is embedded as a set of new features. In our case, we have built a 50-min time history of the selected solar wind parameters by creating new features (columns) in our dataset corresponding to the time-history of each parameter $t - 1, \ldots, t - 50$ min. The time history length was determined purely by our maximum computational capabilities. This has resulted for our final model in an input array of 513 features. The network architecture contains four layers of 320–160–80–40 nodes. The activation function is the rectified linear unit (ReLU). To avoid overfitting, a dropout rate of 0.2 was added between the first and the second, and then between the second and third layers. The training ran for 300 epochs with the possibility of early stopping after 25 epochs of no improvement.

A consequence of embedding the time-history as extra features is that an independent array exists for each training point, and therefore we have trained our ANN model using a random 0.7/0.3 split, as opposed to the sequential split of the data that would be needed with a recurrent neural network. We have reasonably determined that the random split does not introduce data leakage to the model in our testing and that it resolves the bias introduced by the effect of different solar phases in the system. In this case, a more complex manual split of the data or a k-folds technique did not offer substantial improvement over the random split, which increased performance by ~ 20% compared to a sequential split.

2.2.2 Long-Short Term Memory

The Long-Short Term Memory (LSTM) neural network (Hochreiter and Schmidhuber, 1997) was developed as an alternative to solve the gradient vanishing problem of traditional recurrent networks by adding a "long memory." This "memory" refers to the network's ability to "remember" the state of previous cell states as well as previous outputs. The LSTM does this by using a series of gates, the first of which is the *forget gate*. The forget gate uses a sigmoid activation function, which varies between 0 and 1, to decide how much of the output from the previous cell output $(t - 1)$ to feed to the next cell state (t). The *input gate* follows the forget gate and, as its name implies, determines what new information the cell state will receive. The first part of this gate consists of a tanh function, which uses a linear combination of the previous cell output and new input to the current cell, as well as a weight and bias factor. Another sigmoid function is then used to determine how much of the information from the tanh function will be input to the current cell state. The final gate used in the LSTM cell is the *output gate*, which uses another sigmoid function to determine how much information should be passed onto the next cell.

In our model, we used 100 cells in our LSTM layer, followed by two hidden dense layers using 1,000 and 100 nodes respectively. Each dense layer used ReLU activation. Dropout layers with weights of 0.2 were placed in between the hidden layers, and in between the final hidden layer and the output layer, to help prevent overfitting. The training ran for 100 epochs with the possibility of early stopping after 25 epochs of no improvement, and processed data with 60 min (determined by computational limitations) of time history embedded using the method described in **Section 2.2.1**.

2.2.3 Convolutional Neural Network

Convolutional Neural Networks (CNNs) were initially proposed as a method of detecting handwritten digits. They have since proved extraordinarily successful in a variety of image analysis problems (LeCun et al., 2015), and in recent years have shown promise in space weather forecasting (e.g., Collado-Villaverde et al., 2021; Siciliano et al., 2021; Smith et al., 2021). The CNN reads in a matrix all at once, and thus is not explicitly fed the time series information like the LSTM. The dimensions of CNN input

array are (N, height, width, channels), where N is the number of sequences available for training, the height corresponds to the time history, and the width, the number of input features. The CNN is capable of analyzing multiple arrays in the same step. The channels dimension corresponds to the number of arrays to be analyzed at the same time, typically three for RGB color images. For this study we just have the CNN analyze one array per time step, so we set the number of channels equal to one. To keep some consistency between the LSTM and the CNN we used the same input parameters, time history, training data, and training/validation splits, so the input array has dimensions of (N, 60, 13, 1).

The CNN layer functions by using a matrix window called a kernel, which is smaller in size than the 2D input array being analyzed by the layer at step t. The kernel performs a matrix multiplication between a weight matrix the size of the kernel and a segment of the input array of the same size. The output is then put through the activation function (here ReLU), and the kernel window repeats the operation after moving to the next segment of the image. The length that it moves is defined by the stride. In this study, a kernel of size (1,2) and stride of one were used, resulting in overlapping kernel windows between parameters, but not between t and $t - 1$ for the same parameter. Padding, which is the process of adding columns of zeros to the ends of the array image to retain the initial image size, was used. A Pooling layer was used to reduce computational time in the models. The Pooling layer is a method of using a kernel window to move over the output of a CNN layer. Unlike the CNN layer, it does not perform a matrix multiplication using a weight matrix, it only extracts the maximum value in the kernel for the MaxPool, or the average in the kernel for the AveragePool. In this case a MaxPooling layer was used, the maximum value in the kernel window is taken, and the dimensions of the resulting image are reduced. In our case, the output of the CNN layer was of size (60, 13, 1). A 2×2 kernel window and a stride of (2,2) were used, and the resulting dimensions of the output array were (30, 6, 1). The flatten layer was used, which stacks the resulting 2D output from the Pooling layer into a 1D array that can be used as input to the Dense layers. Following the MaxPooling layer were two Dense Layers with 1,024 and 128 nodes, respectively, and dropout of 0.2 in between to help prevent overfitting. The model was trained for 100 epochs and early stopping was used after 25 epochs of no improvement.

3 RESULTS

The results presented in this section correspond to those obtained with the "best" version of each model. Our process of optimization involved testing the use of different solar wind parameters, lengths of the solar wind time series, scalers, splits, loss functions, etc. However, a formal hyper-parameter tuning process such as a Grid Search was not performed. Since model optimization is a never-ending task, we expect to continue it in the future.

Each model (for each station) was trained to output 1-min resolution dB_H/dt values. The final evaluation of those models was done on the six different storms listed in **Table 2**. **Figure 2** shows two of the six storms: 14 December 2006 (left) and 5 April 2010 (right). The rest of the storms can be found in the **Supplementary Material**. Panels (a-d) in **Figure 2** show the main parameters of the solar wind for each storm: SYM-H index, solar wind speed (V_x) component, proton density and interplanetary magnetic field (IMF) B_z. Both geomagnetic storms are driven by interplanetary coronal mass ejections, with a sharp increase in solar wind speed associated with the arrival. It is somewhat expected that most chosen storms correspond to coronal mass ejections as the sudden storm commencement has been associated with larger fluctuations on the ground (e.g., Kappenman, 2003; Fiori et al., 2014; Rogers et al., 2020; Smith et al., 2021). Beyond that, both storms are significantly different in strength and in their proton density and IMF profiles. **Figures 2E–J** panels show the 1-min dB_H/dt measurement from the six different stations considered for this study (black). The three top stations (e-g) correspond to the mid-latitude stations while the bottom three (h-j) are the high-latitude stations. It can be seen that, in general, dB_H/dt spikes tend to scale with the strength of the storm, although peaks can significantly differ in timing and magnitude for stations at similar latitudes depending on their magnetic local time (MLT).

The predictions in the lower panels are shown in red for the ANN, blue for the CNN and green for the LSTM. Those colors will remain associated with the respective models throughout the text. A quick overview of the predictions shown in **Figure 2** indicates that the models are able to somewhat follow the trend of the enhanced activity, while missing most of the variability and spikes in dB_H/dt. A consequence of this is that all models severely under-predict the values unless the real measurements are relatively low. All three models do capture some of the spikes, or the overall increase of dB_H/dt during the storm-period. This is somewhat promising and let us speculate that the models can indeed follow the general evolution of the disturbance strength. At the moment, this is only true for certain stations and certain storms and further studies would be required to improve and evaluate the timing accuracy of the predictions.

Figure 3 shows the root mean square error (RMSE; smaller is better) and the coefficient of determination (R^2; bigger is better) for each of the stations for the same storms shown in **Figure 2**. The rest of the storms can be found in the **Supplementary Material**. By itself, RMSE doesn't allow us to evaluate the quality of the predictions. As can be clearly seen, different stations present markedly different results, with mid-latitude stations having lower RMSE than high-latitude stations due to the significantly lower magnetic fluctuations measured during geomagnetic storms. We can see in **Figure 3** that RMSE values for the different models tend to obtain similar scores at mid-latitudes. At high-latitudes the CNN model performs slightly better than the other two models (by up to 10% depending on the station and the storm). The LSTM tends to perform similarly to the ANN in most of the stations for both storms, although the LSTM performance is slightly better, approaching and even surpassing the CNN performance on a few evaluations. The coefficient of

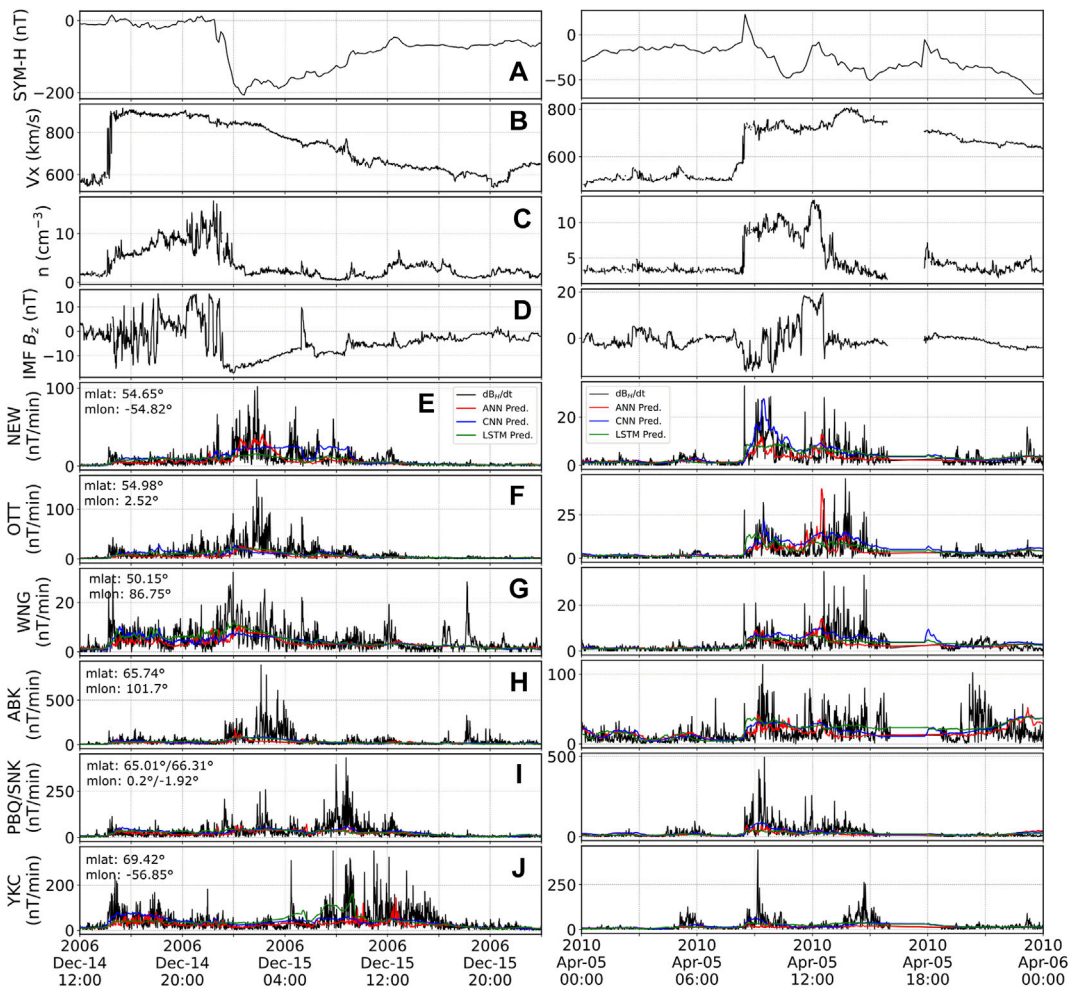

FIGURE 2 | Solar wind parameters **(A–D)** and ground magnetometer dB_H/dt fluctuations as well as our model predictions **(E–J)** for all selected stations during the 14 December 2006 (left) and the 5 April 2010 (right) geomagnetic storms. Panels show **(A)** SYM-H index, **(B)** V_x, **(C)** proton density, **(D)** IMF B_z. Panels **(E–J)** show for each of the labeled stations the 1-min dB_H/dt fluctuations (black), and predictions from the ANN (red), CNN (blue) and LSTM (green) models.

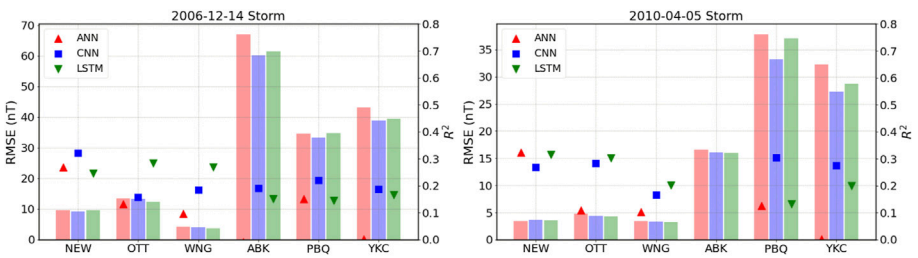

FIGURE 3 | Root mean square errors (bars, left axis) and coefficient of determination R^2 (symbols, right axis) for each model and each station for the 14 December 2006 (left) and the 5 April 2010 (right) geomagnetic storms.

determination (R^2) parameter is less dependent on the magnitude of the fluctuations, and the results are relatively similar across all stations, suggesting that the models may have similar performance based on their solar wind inputs. From the figure, LSTM scores slightly better at mid-latitudes, while CNN performs better at high-latitudes. Still, the overall R^2 values are relatively low (0.1–0.3) and thus is hard to speculate on which model is better just from the pair of metrics shown.

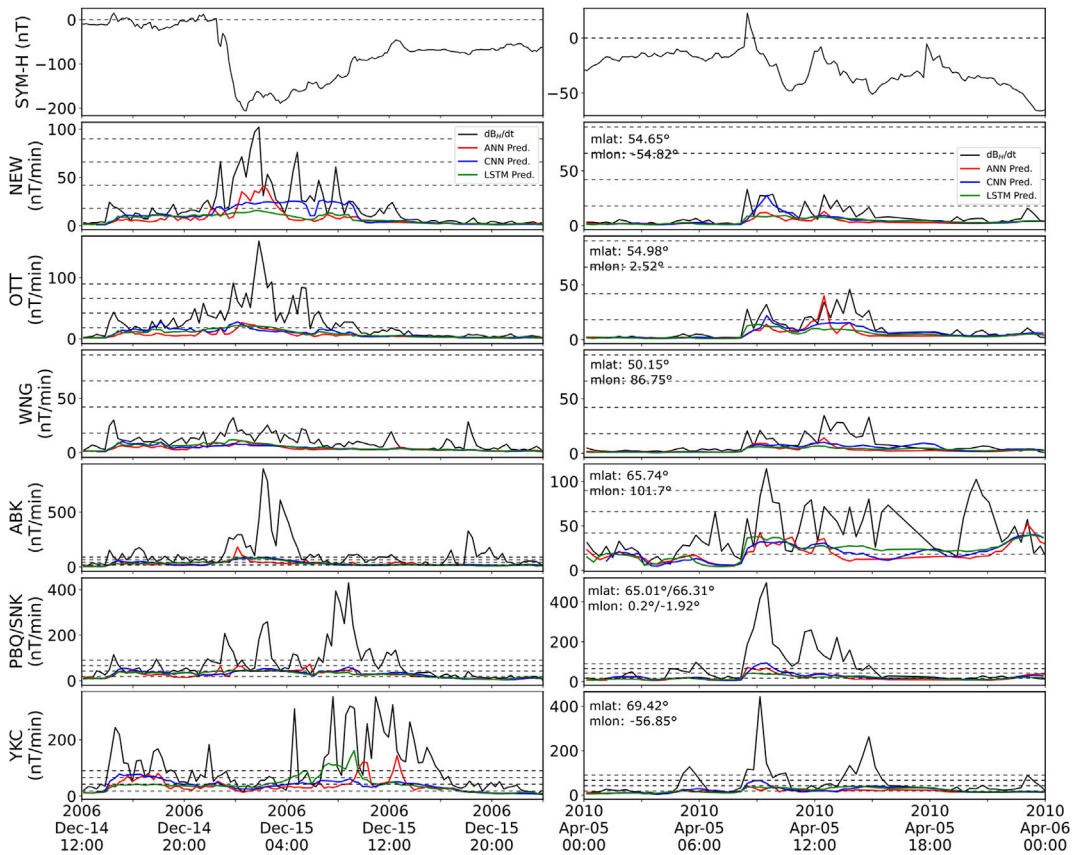

FIGURE 4 | Maximum ground magnetic fluctuations every 20 min (black) and maximum predictions every 20 min for ANN (red), CNN (blue) and LSTM (green). Dashed black lines indicate the thresholds of prediction at 18, 42, 66, and 90 nT/min. From top to bottom, SYM-H index and the six magnetometer stations. The 14 December 2006 storm is shown to the left and the 5 April 2010 storm is shown to the right.

The 1-min resolution forecast proves similarly difficult for our models as it did in the original GEM challenge for the models that were evaluated (Pulkkinen et al., 2013). Therefore, a risk-assessment approach was introduced to evaluate whether the models would predict crossing at different thresholds using the maximum value of the predicted and real data every 20 min. **Figure 4** shows the result of that transformation, with black indicating the real values, and colors indicating the prediction of the different models. Thresholds are drawn at 18, 42, 66 and 90 nT/min (dashed lines) and were selected following the requirements imposed on the models during the GEM Challenge (Pulkkinen et al., 2013). In the figure, the constant under-prediction of the models gets magnified by the drawing of the "upper envelope" of the fluctuations. This can be clearly seen in the 14 December 2006 results where the peak values at most stations are a factor of 10 or more higher than the predictions. This figure, however, does not necessarily indicate that the models perform poorly in the risk-assessment approach; as with the threshold evaluation, it is only important whether or not both the model and the original measurement cross a certain value. The relevant question for the metrics is whether both model and measurements are on the same side of the threshold or not. To do this, a contingency table is created for each storm, station, and threshold and the true positives (hits, H), true negatives (no crossing, N), false positives (false alarms, F), false negatives (missed crossing, M) are recorded.

Following Pulkkinen et al. (2013) we transform the contingency table into probability of detection POD = $H/(H + M)$, probability of false detection POFD = $F/(F + N)$ and the Heidke Skill Score given by

$$\mathrm{HSS} = \frac{2(HN - MF)}{(H + M)(M + N) + (H + F)(F + N)}. \quad (2)$$

The Heidke Skill Score weights the proportion of correct predictions obtained by the model against those that would be obtained purely by randomness. A positive score therefore indicates that the model performs better than chance. **Figure 5** and **Figure 6** show the probability of detection, probability of false detection and Heidke skill scores obtained at each station for the storms discussed in the previous figures. **Figure 5** shows the values for the threshold of 18 nT/min. Despite the general under-prediction of the models, the probability of detecting the crossings at high-latitudes (ABK, PBQ, YKC) is > 0.5 for all models in the 2006 storm and only slightly lower in the 2010 storm. At mid-latitudes the probability of detection is significantly lower for all stations, yet we see again a

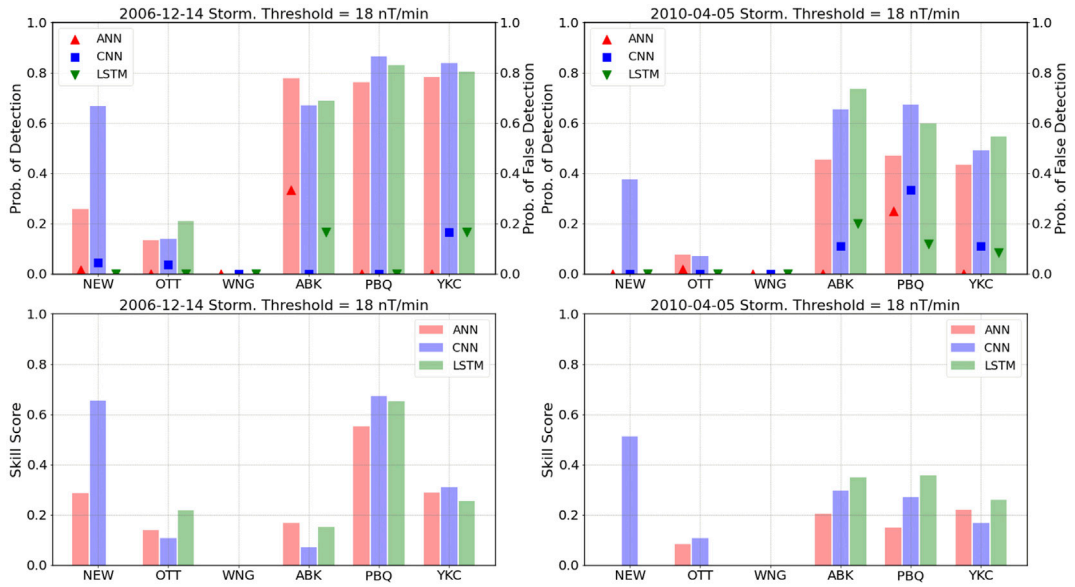

FIGURE 5 | Top panels: Probability of detection (bars, left axis), probability of false detection (symbols, right axis). Bottom panels: Heidke skill score, calculated for the 18 nT/min threshold for each model and each station for the 14 December 2006 (left) and the 5 April 2010 (right) geomagnetic storms.

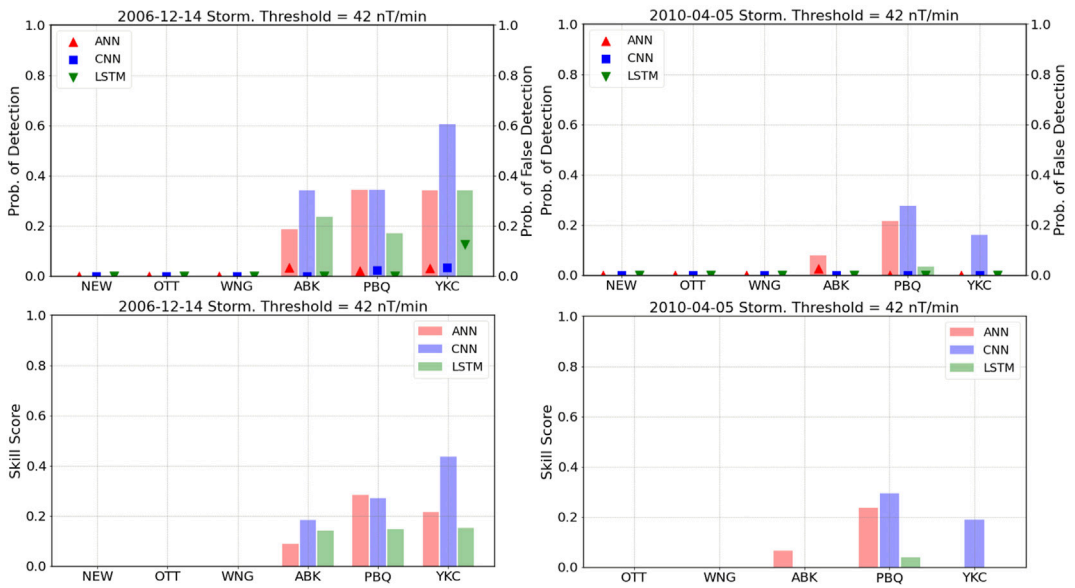

FIGURE 6 | Top panels: Probability of detection (bars, left axis), probability of false detection (symbols, right axis). Bottom panels: Heidke skill score, calculated for the 42 nT/min threshold for each model and each station for the 14 December 2006 (left) and the 5 April 2010 (right) geomagnetic storms.

dominance of the CNN model for these particular cases. The probability of false detection is generally low at all stations and storms, although it is not quantified in this figure if that occurs because of the lack of real crossings over the threshold in that particular storm or not. Still, given the models' consistent under-prediction problem, it is not reasonable to expect a significant number of false positives to contribute to this score.

The Heidke Skill Score shows a larger spread even at the same station for different models, but consistently with the other metrics it seems to indicate a better performance of the models at high latitudes. A particularly interesting result is the extremely poor performance of the models in the station WNG, where none of the three models can get a single correct detection. This seems to be at least in part driven by

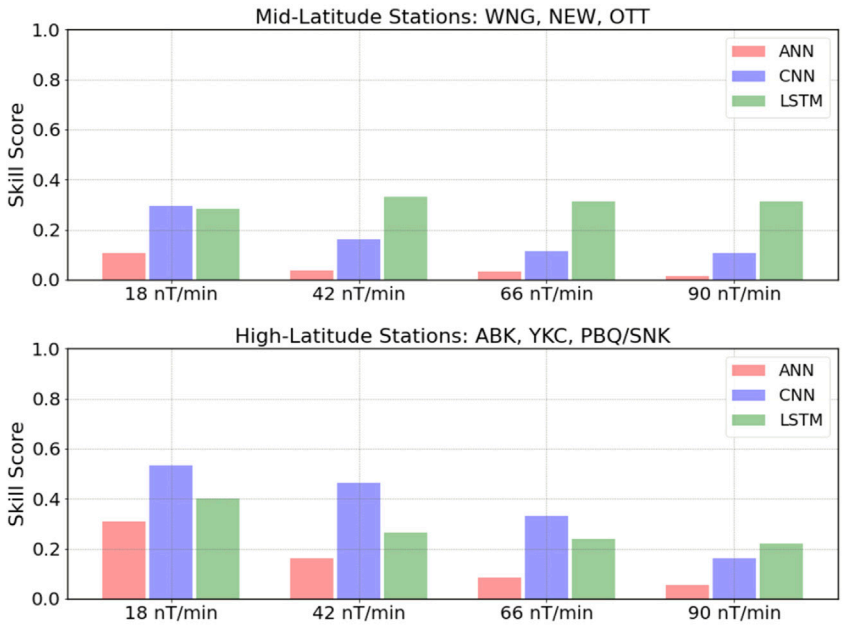

FIGURE 7 | Cumulative Heidke Skill Score "GEM score" calculated by adding the contingency tables of all storms and all stations at mid (top) or high (bottom) latitudes for all three models and four thresholds.

the very small dB_H/dt values measured at that station for those storms.

Figure 6, which shows the values for the threshold of 42 nT/min, shows a similar trend as **Figure 5**. The performance at high latitudes is varied depending on the station and the model, with the CNN model still outperforming the other two, but with results that are (at the very best) moderately good. The lack of a significant number of real crossings of the 42 nT/min threshold at mid-latitude stations makes evaluation of the models very difficult. Though a few crossings do occur, the models miss them. For that same reason we are not showing the individual results for the 66 nT/min and the 90 nT/min thresholds, although they are included in the **Supplementary Material** for the sake of completeness.

To properly compare with the GEM Challenge, we calculated the Heidke Skill Score by aggregating all the geomagnetic storms for all mid-latitude stations (WNG, NEW, OTT) and doing the same for the high-latitude stations (ABK, YKC, PBQ/SNK). This results in two scores for each threshold, one at high latitudes and one at mid-latitudes. **Figure 7** shows the results obtained by each of the models at mid-latitudes (top panel) and high latitudes (bottom panel). From the figure, we can note that the final scores are generally consistent with the individual scores obtained in the previous figures (and with those not shown in the manuscript). It is clear that the model that uses a CNN outperforms the other two consistently at high-latitudes, for the first three thresholds. However, at mid-latitudes it is the LSTM model that performs the best, even holding some predictive power (i.e., HSS positive) at the 90 nT/min threshold. A comparison against the models shown by Pulkkinen et al. (2013) would indicate that the CNN and LSTM models outperform all the GEM challenge models at high latitudes for the lowest two thresholds but do a bit worse than the top performer (Space Weather Modeling Framework-SWMF) for the highest two. At mid-latitudes, however, even the LSTM model is outperformed by most of the GEM Challenge models, indicating that our models do present a different behavior at mid-latitudes and high latitudes, even beyond the differences in the scores which can be attributed to many causes.

4 DISCUSSION

The development of machine-learning models to forecast 1-min ground magnetometer fluctuations (dB_H/dt) and our benchmark against the set of metrics previously used in similar models during the GEM Challenge for ground magnetic perturbations presented several interesting challenges, and therefore we have learned important lessons from the process. In the next sections we discuss a few of the most important points regarding the evaluation of the models and the improvements that need to be made moving forward.

4.1 The 30 October 2003 Storm

Out of the events selected for evaluation, perhaps the most interesting is the storm that occurred on 30 October 2003. This storm is the third largest storm recorded in the high resolution OMNI dataset (1995-present). It is reasonable to expect modelers to test the models on such extreme event. This storm, however, presents a series of challenges for our models, the most important being that there are no high resolution plasma parameters available during most of the storm due to a saturation of the instrument on-board the ACE

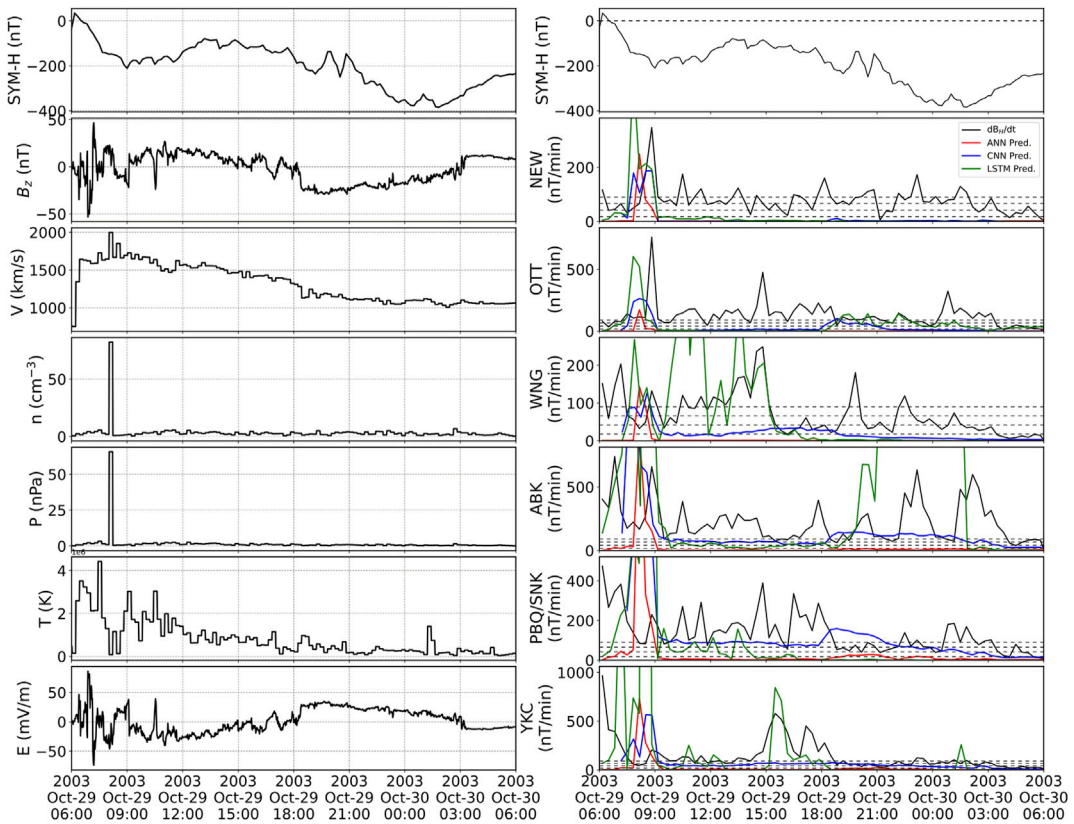

FIGURE 8 | Solar wind parameters (left) and 20-min window ground magnetometer data and predictions (right) for the 30 October 2003 geomagnetic storm. From top to bottom (left) SYM-H index, IMF B_z, proton speed, proton density, dynamic pressure, temperature and electric field. From top to bottom (right) SYM-H index, and all six ground magnetometer stations showing 20-min window maximum values of real measurements (black), ANN model (red), LSTM model (green), and CNN model (blue).

spacecraft (Skoug, 2004). For our model evaluation of the 2003 storm, we used a procedure similar to that described by Pulkkinen et al. (2013) which involved the use of low-resolution (1-h) ACE data to reconstruct the plasma parameters and 4-s resolution data for the interplanetary magnetic field. The data was then propagated to the bow-shock nose to make it consistent with OMNI data. **Figure 8** (left) shows the reconstructed solar wind data re-sampled at 1-min resolution. The only data that could not be reconstructed are solar wind speed V_y and V_z which are shown as straight lines connecting last known values (linear interpolation).

Figure 8 (right) shows the prediction of the models for the different stations during the 30 October 2003 storm. Here, one of the main difficulties when training machine-learning models becomes evident: their poor ability to extrapolate to unseen data. It can be seen that the models behave in strange different ways. All three of the models respond to the sudden increase in proton density at the beginning of the evaluated timespan, but the models' predictions differ significantly afterwards. For example, the ANN predictions go to zero following the initial spike, thus missing most of the strong fluctuations. The CNN model, although troubled to produce a strong prediction, seems to at least be robust enough to follow a pattern of prediction similar to what it would predict in different storms. Finally, the LSTM model predicts huge spikes in at least two stations. Fine-tuning a model to get good predictions on extreme (and unseen) data was not among the goals we set for this work, but it is something that we will consider moving forward.

4.2 Metrics

The Heidke Skill Score (HSS) was the main metric used here for comparison with the GEM challenge. The main reason for its use was that it was also their metric of choice, and as such was the simpler choice. We believe that the use of only one metric to evaluate a model is restrictive, as it provides only a glimpse into the strengths and weaknesses of that model. For example, the HSS (**Equation 2**) contains a series of products or sums between elements of the contingency table. This requires a variety of table elements to produce a meaningful score. During the process of model evaluation, the most intense storm in the testing suite, the 2003 Halloween storm, had a large percentage of missing data, meaning the model evaluation was only done on a portion of the storm where data was available. This portion of storm data was completely above the lowest (18 nT/min) threshold. The model, recognizing the intensity of the storm, predicted over the threshold for the same time period. This resulted in the H

(hits) element of the table being the only one populated, as all of the predictions and real data were over the lowest threshold, ideally a perfect model. However, because only one element of the table was nonzero, we get zeros in both the numerator and denominator of the HSS, producing a NaN value in our evaluation. Similarly, in the evaluation of the 2003 storm, the PBQ station had an almost perfect prediction in terms of being all hits for the 18 nT/min threshold. However, while the proportion of hits was very high, there was one false negative. Because there were only two elements of the table represented, but they are in different terms of the numerator, we get a result of zero for the HSS. A score of zero is supposed to be akin to 50–50 random chance model; however, with a hit-to-false-negative proportion of 13:1 for this particular storm, that is obviously not the case, showing that the HSS does not do justice to the skill of the model. Thus, it is important to consider multiple metrics when validating or comparing models. Liemohn et al. (2021) provides an overview of numerous metrics, and Welling et al. (2018) recommends adding a Frequency Bias metric to those used by Pulkkinen et al. (2013) for assessment of ground magnetic field perturbations.

It is also important to consider that out of the six storms evaluated for the six ground magnetometer stations, the 30 October 2003 storm is the only storm that provides a high number of crossings above the higher three thresholds. This is also discussed in Pulkkinen et al. (2013) because it heavily impacts the overall HSS score of a model depending on whether the model can effectively predict fluctuations that are large enough to cross over those thresholds. In our case, the ANN model that fails to predict the 2003 storm at all, sees its HSS tremendously affected when compared against the other two models, even if they are all similar in performance for the remaining of the geomagnetic storms evaluated.

4.3 Training and Testing

One of the reasons to replicate an existing community effort is that we wanted to benchmark our model results against known baseline models. In doing so, we have made choices that may or may not be the optimal choices for a machine learning model. A good example is the 2003 storm, which would be ideally used for training instead of for testing given its unique nature in the existing dataset (and that we will use when the models move into operational real-time forecast). As mentioned before, it is understandable that modelers may want to test using extreme events, as opposed to machine-learning practices where extreme events can help models perform better. However, in the future, it may be worth exploring new events for testing, such as those already proposed by Welling et al. (2018).

Another important aspect not addressed in detail here is the choice of the target parameter. Following the GEM challenge we focused on the 1-min resolution dB_H/dt values, and then reprocessed those predictions to obtain the maximum value every 20-min, which is what was finally used for the actual evaluation. While a 20 or 30 min window of prediction is probably a reasonable timespan in which to raise warnings when a model is operational, the way the model was proposed, it is not actively creating predictions that far into the future but rather 1-min ahead (plus the time of propagation from L1), which can lead to confusion. In the future, we plan to try different types of forecasts, such as doing a direct prediction of the maximum value of the fluctuations over a determined time window.

5 SUMMARY AND CONCLUSION

We have revisited the ground magnetic field perturbations challenge "GEM Challenge" using deep learning models for our evaluation: a feed forward neural network (ANN), a convolutional neural network (CNN) and a long short-term memory recurrent network (LSTM). We followed the same procedure set by the original challenge, including the forecast of 1-min resolution dB_H/dt values, followed by a conversion to a "maximum of" in 20-min windows. We then evaluated our models by creating a contingency table for thresholds of 18, 42, 66 and 90 nT/min. The metrics created from these contingency tables were probability of detection, probability of false detection and the Heidke Skill Score, which we used to evaluate our models at six ground magnetometer stations, three mid-latitude and three high-latitude, over six different geomagnetic storms. We finally calculated an overall score by aggregating storms at mid-latitude stations and also at high-latitude stations.

Overall, we found that the machine-learning models we developed tend to perform similarly or slightly worse compared against the models presented by Pulkkinen et al. (2013), with scores that would situate them roughly in the middle of all the models they tested. Pulkkinen et al. (2013) does not present exact numbers, so those need to be inferred from their figures. For example, our models perform poorly for the 18 nT/min threshold at mid-latitudes compared to all models discussed there. On the other hand, two of our models (CNN, LSTM) outperform all but the two top models at high-latitude for the same threshold. At the 42 nT/min threshold, our models (LSTM at mid-lat, CNN at high-lat) would outperform all but the top model presented there. There are several reasons for such results, including difficulties in predicting the 30 October 2003 geomagnetic storm, which is a unique and extreme case that causes machine learning training to predict poorly. Out of the three models we tested, the CNN did consistently better than the other two.

The machine-learning models we used here have a few advantages over traditional simulations such as the minimal computational requirements they need for training, and to be run in real-time. Most of our models have been trained in machines of moderate computational power, and more importantly can provide real-time predictions on a desktop computer. This allows for great flexibility in the design of models and quick iteration between different algorithms as they become available. Here we used an LSTM, CNN, and even an ANN model for their capability to capture the time-history of the time series used as an input. We consider that any machine learning model capable of capturing the temporal evolution of the target parameter is worth exploring and could be used in the future. We plan in the future to continue exploring models of this type, with the intention of moving into real-time forecasting.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: OMNI Dataset: https://cdaweb.gsfc.nasa.gov/pub/data/omni/SuperMAg Dataset: https://supermag.jhuapl.edu/ACE Dataset: http://www.srl.caltech.edu/ACE/ASC/level2/index.html.

AUTHOR CONTRIBUTIONS

VP contributed to conception and design of the study, data preparation, model development and analysis, interpretation of results, and writing. AK contributed to design of the study, interpretation of results, data preparation, writing and general guidance. MC and RM contributed to model development, analysis, interpretation, and assisted with writing. JJ contributed to model development and methodology design. CN and HC contributed with design of the study and overall discussion. All authors contributed to manuscript revision and read and approved the submitted version.

FUNDING

This work is supported by NSF Award 1920965. CN was supported through NASA Grant Award 80NSSC-20K1364 and NSF Grant Award AGS-2117932.

ACKNOWLEDGMENTS

We thank all members of the MAGICIAN team at UNH and UAF that participated in the discussions leading to this article. We also thank the OMNIWeb, SuperMAG and ACE teams for providing the data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fspas.2022.869740/full#supplementary-material

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv:1603.04467 [cs].

Boteler, D. H. (2019). A 21st Century View of the March 1989 Magnetic Storm. *Space Weather* 17, 1427–1441. doi:10.1029/2019SW002278

Boteler, D. H. (2001). "Space Weather Effects on Power Systems," in *Geophysical Monograph Series* (Washington, D. C.: American Geophysical Union), 347–352. doi:10.1029/GM125p0347

Collado-Villaverde, A., Muñoz, P., and Cid, C. (2021). Deep Neural Networks with Convolutional and LSTM Layers for SYM-H and ASY-H Forecasting. *Space Weather* 19, e2021SW002748. doi:10.1029/2021SW002748

Dimmock, A. P., Rosenqvist, L., Welling, D. T., Viljanen, A., Honkonen, I., Boynton, R. J., et al. (2020). On the Regional Variability of d *B*/d *t* and its Significance to GIC. *Space Weather* 18, e2020SW002497. doi:10.1029/2020SW002497

Fiori, R. A. D., Boteler, D. H., and Gillies, D. M. (2014). Assessment of GIC Risk Due to Geomagnetic Sudden Commencements and Identification of the Current Systems Responsible. *Space Weather* 12, 76–91. doi:10.1002/2013SW000967

Gannon, J. L., Birchfield, A. B., Shetye, K. S., and Overbye, T. J. (2017). A Comparison of Peak Electric Fields and GICs in the Pacific Northwest Using 1-D and 3-D Conductivity. *Space Weather* 15, 1535–1547. doi:10.1002/2017SW001677

Gil, A., Berendt-Marchel, M., Modzelewska, R., Moskwa, S., Siluszyk, A., Siluszyk, M., et al. (2021). Evaluating the Relationship between Strong Geomagnetic Storms and Electric Grid Failures in Poland Using the Geoelectric Field as a GIC Proxy. *J. Space Weather Space Clim.* 11, 30. doi:10.1051/swsc/2021013

Gjerloev, J. W. (2012). The SuperMAG Data Processing Technique. *J. Geophys. Res. Space Phys.* 117, A09213. doi:10.1029/2012ja017683

Hapgood, M., Angling, M. J., Attrill, G., Bisi, M., Cannon, P. S., Dyer, C., et al. (2021). Development of Space Weather Reasonable Worst-Case Scenarios for the UK National Risk Assessment. *Space Weather* 19, e2020SW002593. doi:10.1029/2020SW002593

Hochreiter, S., and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.* 9, 1735–1780. doi:10.1162/neco.1997.9.8.1735

Kappenman, J. G. (2003). Storm Sudden Commencement Events and the Associated Geomagnetically Induced Current Risks to Ground-Based Systems at Low-Latitude and Midlatitude Locations. *Space Weather* 1, 1016. doi:10.1029/2003sw000009

Keesee, A. M., Pinto, V., Coughlan, M., Lennox, C., Mahmud, M. S., and Connor, H. K. (2020). Comparison of Deep Learning Techniques to Model Connections between Solar Wind and Ground Magnetic Perturbations. *Front. Astron. Space Sci.* 7, 550874. doi:10.3389/fspas.2020.550874

King, J. H., and Papitashvili, N. E. (2005). Solar Wind Spatial Scales in and Comparisons of Hourly Wind and ACE Plasma and Magnetic Field Data. *J. Geophys. Res.* 110, A02104. doi:10.1029/2004JA010649

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep Learning. *Nature* 521, 436–444. doi:10.1038/nature14539

Liemohn, M. W., Shane, A. D., Azari, A. R., Petersen, A. K., Swiger, B. M., and Mukhopadhyay, A. (2021). RMSE is Not Enough: Guidelines to Robust Data-Model Comparisons for Magnetospheric Physics. *J. Atmos. Sol.-Terr. Phys.* 218, 105624. doi:10.1016/j.jastp.2021.105624

Love, J. J., Lucas, G. M., Kelbert, A., and Bedrosian, P. A. (2018). Geoelectric Hazard Maps for the Mid-Atlantic United States: 100 Year Extreme Values and the 1989 Magnetic Storm. *Geophys. Res. Lett.* 45, 5–14. doi:10.1002/2017GL076042

Lucas, G. M., Love, J. J., Kelbert, A., Bedrosian, P. A., and Rigler, E. J. (2020). A 100-Year Geoelectric Hazard Analysis for the U.S. High-Voltage Power Grid. *Space Weather* 18, e2019SW002329. doi:10.1029/2019SW002329

Ngwira, C. M., Pulkkinen, A. A., Bernabeu, E., Eichner, J., Viljanen, A., and Crowley, G. (2015). Characteristics of Extreme Geoelectric Fields and Their Possible Causes: Localized Peak Enhancements. *Geophys. Res. Lett.* 42, 6916–6921. doi:10.1002/2015GL065061

Ngwira, C. M., Sibeck, D., Silveira, M. V. D., Georgiou, M., Weygand, J. M., Nishimura, Y., et al. (2018). A Study of Intense Local dB/dt Variations during Two Geomagnetic Storms. *Space Weather* 16, 676–693. doi:10.1029/2018SW001911

Oughton, E. J., Hapgood, M., Richardson, G. S., Beggan, C. D., Thomson, A. W. P., Gibbs, M., et al. (2019). A Risk Assessment Framework for the Socioeconomic Impacts of Electricity Transmission Infrastructure Failure Due to Space Weather: An Application to the United Kingdom. *Risk Anal.* 39, 1022–1043. doi:10.1111/risa.13229

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

Pirjola, R. (2000). Geomagnetically Induced Currents during Magnetic Storms. *IEEE Trans. Plasma Sci.* 28, 1867–1873. doi:10.1109/27.902215

Pirjola, R. (2002). Review on the Calculation of Surface Electric and Magnetic Fields and of Geomagnetically Induced Currents in Ground-Based Technological Systems. *Surv. Geophys.* 23, 71–90. doi:10.1023/A:1014816009303

Pulkkinen, A., Bernabeu, E., Eichner, J., Viljanen, A., and Ngwira, C. (2015). Regional-Scale High-Latitude Extreme Geoelectric Fields Pertaining to

Geomagnetically Induced Currents. *Earth Planet Sp.* 67, 93. doi:10.1186/s40623-015-0255-6

Pulkkinen, A., Bernabeu, E., Thomson, A., Viljanen, A., Pirjola, R., Boteler, D., et al. (2017). Geomagnetically Induced Currents: Science, Engineering, and Applications Readiness. *Space Weather* 15, 828–856. doi:10.1002/2016SW001501

Pulkkinen, A., Rastätter, L., Kuznetsova, M., Singer, H., Balch, C., Weimer, D., et al. (2013). Community-Wide Validation of Geospace Model Ground Magnetic Field Perturbation Predictions to Support Model Transition to Operations. *Space Weather* 11, 369–385. doi:10.1002/swe.20056

Rogers, N. C., Wild, J. A., Eastoe, E. F., Gjerloev, J. W., and Thomson, A. W. P. (2020). A Global Climatological Model of Extreme Geomagnetic Field Fluctuations. *J. Space Weather Space Clim.* 10, 5. doi:10.1051/swsc/2020008

Siciliano, F., Consolini, G., Tozzi, R., Gentili, M., Giannattasio, F., and De Michelis, P. (2021). Forecasting SYM-H Index: A Comparison between Long Short-Term Memory and Convolutional Neural Networks. *Space Weather* 19, e2020SW002589. doi:10.1029/2020SW002589

Skoug, R. M. (2004). Extremely High Speed Solar Wind: 29-30 October 2003. *J. Geophys. Res.* 109, A09102. doi:10.1029/2004JA010494

Smith, A. W., Forsyth, C., Rae, I. J., Garton, T. M., Bloch, T., Jackman, C. M., et al. (2021). Forecasting the Probability of Large Rates of Change of the Geomagnetic Field in the UK: Timescales, Horizons, and Thresholds. *Space Weather* 19, e2021SW002788. doi:10.1029/2021SW002788

Tóth, G., Meng, X., Gombosi, T. I., and Rastätter, L. (2014). Predicting the Time Derivative of Local Magnetic Perturbations. *J. Geophys. Res. Space Phys.* 119, 310–321. doi:10.1002/2013JA019456

Viljanen, A., Nevanlinna, H., Pajunpää, K., and Pulkkinen, A. (2001). Time Derivative of the Horizontal Geomagnetic Field as an Activity Indicator. *Ann. Geophys.* 19, 1107–1118. doi:10.5194/angeo-19-1107-2001

Viljanen, A. (1998). Relation of Geomagnetically Induced Currents and Local Geomagnetic Variations. *IEEE Trans. Power Deliv.* 13, 1285–1290. doi:10.1109/61.714497

Welling, D. T., Ngwira, C. M., Opgenoorth, H., Haiducek, J. D., Savani, N. P., Morley, S. K., et al. (2018). Recommendations for Next-Generation Ground Magnetic Perturbation Validation. *Space Weather* 16, 1912–1920. doi:10.1029/2018SW002064

Wintoft, P. (2005). Study of the Solar Wind Coupling to the Time Difference Horizontal Geomagnetic Field. *Ann. Geophys.* 23, 1949–1957. doi:10.5194/angeo-23-1949-2005

Wintoft, P., Wik, M., and Viljanen, A. (2015). Solar Wind Driven Empirical Forecast Models of the Time Derivative of the Ground Magnetic Field. *J. Space Weather Space Clim.* 5, A7. doi:10.1051/swsc/2015008

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Astronomy and Space Sciences

MINI REVIEW
published: 03 June 2022
doi: 10.3389/fspas.2022.865880

Statistical Methods Applied to Space Weather Science

*Daniele Telloni**

National Institute for Astrophysics, Astrophysical Observatory of Torino, Pino Torinese, Italy

Space Weather is receiving more and more attention from the heliophysical scientific community, as it is now well established that an adequate capability of monitoring any Earth-directed heliospheric event and forecasting the most severe perturbations produced by solar activity and their impact on the geo-spatial environment is crucial, given the human increasing reliance on space-related technologies and infrastructures. Predicting how the Sun affects life on Earth and human activities in the short term relies on establishing empirical laws to forecast not only the arrival time on Earth of potentially geo-effective solar drivers, but also, and more importantly, the intensity of induced geomagnetic disturbance (if any). Scientific studies performed on a statistical basis are the key to providing such empirical laws and analytically relating solar-wind properties to geomagnetic indices. This paper summarizes the results achieved by the author in the last few years in the context of Space Weather science, and based on statistical analyses of interplanetary and geomagnetic data.

Keywords: methods: statistical, solar-terrestrial relations, Sun: activity, Sun: coronal mass ejections (CMEs), solar wind, turbulence, interplanetary medium, magnetohydrodynamics (MHD)

OPEN ACCESS

Edited by:
Bala Poduval,
University of New Hampshire,
United States

Reviewed by:
Munetoshi Tokumaru,
Nagoya University, Japan
Robert James Leamon,
University of Maryland, United States

***Correspondence:**
Daniele Telloni
daniele.telloni@inaf.it

Specialty section:
This article was submitted to Space Physics, a section of the journal Frontiers in Astronomy and Space Sciences

Received: *30 January 2022*
Accepted: *20 May 2022*
Published: *03 June 2022*

Citation:
Telloni D (2022) Statistical Methods Applied to Space Weather Science. Front. Astron. Space Sci. 9:865880. doi: 10.3389/fspas.2022.865880

1 INTRODUCTION

The Sun influences conditions in the near-Earth environment, including the magnetosphere, ionosphere and thermosphere, and can pose a persistent hazard in the form of damaging radiation to both space- or ground-based stations and human health. More specifically, Space Weather (as commonly referred to the science dealing with the complex Sun-Earth interaction and forecasting of potentially geo-effective events) covers the geo-space disturbances caused by the release of solar energy into the Earth's magnetosphere during geomagnetic storms, and all related phenomena. Sun-related environmental impacts include a potential slowdown and orbital decay of the low-Earth-orbiting satellites (due to an additional aerodynamic drag force induced by solar activity), induction of very harmful electric currents in power transmission grids and pipelines, disruption of satellite signal propagation with severe implications for positioning systems, and unrecoverable failures of electronics onboard spacecraft. The ionosphere reflectivity can also be altered by the arrival of solar energetic particles, impairing radio communication systems. Finally, Space Weather deals with radiation produced by solar storms that can endanger the astronauts' health.

Interplanetary counterparts of Coronal Mass Ejections (ICMEs, large eruptions of magnetized plasma from the Sun into interplanetary space Webb and Howard, 2012), which occur much more frequently at solar maximum than at minimum, and Corotating Interaction Regions (CIRs, forming at the interface between high- and low-speed streams), which are instead typical of low activity phases of the solar cycle, are the largest interplanetary manifestations of the solar activity (Gosling et al., 1990; Tsurutani et al., 1995; Gonzalez et al., 1999; Yermolaev et al., 2005, 2012). These

interplanetary structures, which can be seen as propagating regions of space of enhanced density and magnetic field strength, are characterized by an intense and long-lasting South-directed magnetic field, which thus magnetically reconnects with the oppositely (North-)oriented Earth's magnetic field, according to the scenario first proposed by Dungey (1961). This process allows a net transfer of energy from the solar wind to Earth, triggering "*de facto*" the most severe geomagnetic disturbances (Russell and McPherron, 1973; Gonzalez et al., 1994; Baker et al., 1996). However, the old-fashioned paradigm that the level of geomagnetic storming depends primarily on how pronounced in the southern direction the interplanetary magnetic field is (Fairfield and Cahill, 1966) is not quite correct. Other solar wind-related parameters, such as the dynamic pressure (e.g., Burton et al., 1975), the transported kinetic/magnetic energy (Telloni et al., 2020), and turbulence (e.g., D'Amicis et al., 2020), play a crucial role in driving the geomagnetic activity.

Although one-to-one studies have been often performed so far, a statistical approach is needed for forecasting Space Weather phenomena, with particular reference to predicting the geomagnetic response to the impact of geo-effective solar structures, the relativistic electron flux (which may cause irreparable damage to the geosynchronous satellites, Forsyth et al., 2020), the occurrence of solar flares, the propagation time of CMEs, the transit of high-speed streams to Earth, and the crossing of the heliospheric current sheet (the latter two also being sources of geomagnetic disturbances, though to a lesser extent). In fact, by means of the analysis of a large amount of solar, interplanetary, and geomagnetic data of past events, it is possible to establish empirical laws, a sort of analytical functions relating the different quantities involved, that allow the prediction of the onset of new solar events and/or their effects on the Earth's magnetosphere.

Most forecasting methods rely on remote-sensing observations of solar phenomena, i.e., CMEs, causing geomagnetic storms, and can be roughly divided into three main classes, namely, physics-based, event-based, and drag-based models, depending on the approach used to provide expectations of CME arrival times. Physics-based models rely on photospheric magnetic field observations to initiate numerical MagnetoHydroDynamic (MHD) simulations of the eruption of the CME and its propagation from Sun to Earth. Predictions of the CME transit time can be thus provided. These numerical codes require the use of supercomputers to run efficiently. In addition, their reliability obviously depends on a correct representation of the physical processes within the models, i.e., the understanding (unfortunately not yet full) of the physics of the corona and the solar wind. The MHD models currently used for operational Space Weather predictions are the well-known Enlil (Odstrcil, 2003) and the EUropean Heliospheric FORecasting Information Asset (EUHFORIA, Pomoell and Poedts, 2018). Simpler and much less computationally expensive (but no less reliable) event-based (or empirical) models rely on statistical studies of past CMEs and essentially relate the CME Sun-Earth transit times to their propagation speeds, as inferred from coronagraphic images. This allows the establishment of empirical laws, say analytical functions, that (assuming that past observations are analogous to future ones, i.e., that CMEs share common kinematic characteristics) allow prediction of the impact time on Earth of a new CME, once its coronagraphic speed is measured (e.g., Manoharan et al., 2004; Schwenn et al., 2005; Vršnak and Žic, 2007). Similar empirically-derived relations to forecast the geo-effectiveness of CMEs are also available (e.g., Dumbović et al., 2015). Observational evidence for an adjustment of the CME propagation speed to the background solar wind and its interpretation in terms of aerodynamic drag, stimulated the development of the so-called drag-based models (e.g., Vršnak and Gopalswamy, 2002; Vršnak et al., 2013), which basically assume that the CME propagation in the heliosphere is governed by aerodynamic drag (one of the most refined drag-based model is 3D COronal Rope Ejection (3DCORE) introduced by Möstl et al., 2018). That is, the dynamics/kinematics of the CME can be analytically described through a pretty simple equation of motion, which can thus provide real-time prediction of the CME arrival time and impact speed at Earth (in spite of various drawbacks associated with the approximations intrinsic to this approach).

Regardless of the pros and cons of the different approaches (whose discussion is beyond the scope of this paper, but the interested reader is refereed to Verbeke et al., 2019), all these methods provide alerts 1–4 days in advance of the geomagnetic storm, although the predictions are significantly model-dependent and affected by large uncertainties. On the other hand, expectations of CME arrival and storminess level based on *in-situ* solar wind data at the Lagrangian point L1, i.e., Space Weather now-casting methods, are not widely used, although they could provide much more accurate warnings. This is essentially due to the difficulty of identifying CMEs locally in the interplanetary medium with *in-situ* measurements. Interplanetary scintillation (IPS), which is scattering phenomenon of solar wind density irregularities, serves as a remote sensing method for observing the solar wind. Thus, IPS observations have the potential to bridge a gap between the Sun and the near-Earth solar wind. Some efforts to improve CME arrival time predictions already have been performed using IPS observations (e.g., Iwai et al., 2021). However, in *in-situ* data, many of the CME distinctive properties (i.e., higher magnetic fields and lower plasma densities/temperatures with respect to the ambient solar wind in which they propagate, Burlaga et al., 1981) are common to a variety of other interplanetary structures, such as high-speed streams. What really distinguishes them is a rotation of the magnetic field vector in the plane perpendicular to the direction of propagation. This is due to the presence of a flux rope (which generally all CMEs embed and carry during their expansion, Vourlidas, 2014), a helical structure that can be revealed as a region of space with high magnetic helicity (an MHD quantity that measures the degree of twisting of magnetic field lines). However, unlike the shock front of a CME, which provides a prompt signal, in order for the flux rope-related magnetic field rotation to be detected, the CME must have entirely passed the spacecraft orbiting at L1, thus drastically reducing forecasting capabilities: indeed, at least for the largest

structures (which can have a radial extension at Earth of even 0.25 au, Klein and Burlaga, 1982, much larger than the L1 point-Earth distance of only 0.01 au), the front may have already impacted Earth by the time diagnostic codes based on magnetic helicity measurements identified the presence of a CME at L1. This issue, along with difficulties associated with measuring the magnetic helicity (the reader is referred to Telloni et al., 2012, 2013, for a detailed discussion) have limited the development of now-casting methods based on *in-situ* L1 measurements of the solar wind for Space Weather purposes. This lack motivated the works by Telloni et al. (2019, 2020, 2021) (hereafter Papers I, II, and III), all appeared on The Astrophysical Journal and based on statistical surveys of solar wind and geomagnetic data (the period analyzed in each case covers more than one solar cycle) with the aim of obtaining statistical relationships, i.e., empirical laws, that can be used in the Space Weather framework to predict the onset and the time evolution of geomagnetic storming.

The present paper summarizes the results obtained in the aforementioned three papers. Ideally following the title, **section 2** reports the statistical results obtained in the three studies, while **section 3** discusses their applications to Space Weather science. **section 4** is devoted to future developments of the application of statistical methods and machine learning in Space Weather science in the framework of the Space Weather Service Network (SWESNET) project.

2 STATISTICAL RESULTS

Paper I addressed the detection, characterization, and geo-effectiveness likelihood of ICMEs. The localization of ICMEs in the near-Earth space environment was accomplished by comparing some MHD quantities measured at L1 with those typical of the unperturbed solar wind plasma. Specifically, ICMEs were identified as structures with a large magnetic helicity content (representative of the embedded flux rope) that also have a total (thermal plus magnetic) internal pressure higher than the medium in which they propagate (Gosling et al., 1994). The potential geo-effectiveness of the so identified ICMEs was ascertained by looking at their energy budget: only those ICMEs carrying an amount of kinetic and/or magnetic energy far exceeding that characteristic of the quiet solar wind (thus ensuring a remarkable energy transfer to the Earth's magnetosphere during magnetic reconnection processes) were in fact defined as able of inducing geomagnetic perturbations. In the 12-year period from 2005 to 2016, 106 likely geo-effective ICMEs were thus revealed in the *Wind* spacecraft data by the *in-situ* data-based tool developed in Paper I. The actual geomagnetic disturbances driven by those ICMEs were verified by inspecting the Earth's magnetospheric activity through the Dst (disturbance storm time) and Ap indices, both indicative (albeit at different ground latitudes) of the intensification of ring current systems caused by solar storms. Specifically, sustained periods of either Dst < -50 nT (Cander and Mihajlovic, 1998) or Ap larger than the value reflective of the quiet configuration of the magnetosphere, identified the ICME-driven geomagnetic perturbations. On the one hand, this allowed the estimation of the efficiency in identifying at L1 CMEs potentially geo-effective. It turned out that the efficiency increases with the storminess level: from 86% for the weakest geomagnetic disturbances (-50 nT $>$ Dst > -100 nT), through 94% for moderate perturbations (-100 nT $>$ Dst > -250 nT), to as high as 100% for the most severe ones (Dst < -250 nT). On the other hand, it allowed quantitation of the time between the *in-situ* detection of the CME and the onset of the related increase in geomagnetic activity. The distribution of the waiting times is shown in panel (A) of **Figure 1**: on average, this time delay is about 4 h and 20 min (vertical dashed red line). Overlaid is a log-normal distribution (blue dashed curve), which allows estimation of the confidence interval for the waiting time: it results that in 98% of instances this waiting time is between 2 and 8 h.

Paper II extended the study to any likely geo-effective solar event (not just CMEs), focusing, in the same 2005–2006 interval, on the relationship existing between solar wind energy and geomagnetic activity. Panel (B) of **Figure 1** displays the 2D histogram of a dimensionless measure of the total (kinetic plus magnetic) energy E carried by the solar wind (see Telloni et al., 2020, for more details on how E was derived from *Wind in-situ* data) and the Dst index. Superimposed are the Dst most likely value (black solid line) and 68% probability range (gray shaded area) for each energy bin. It appears evident that a clean statistical correlation exists between the energy content of the solar wind impacting Earth and the perturbation level of the magnetospheric current system: that is, the larger the energy stored in the solar wind plasma, the more severe the induced geomagnetic perturbations. It follows that in the solar wind-magnetosphere coupling, energy is to be thus regarded as a crucial parameter in solar-terrestrial interactions.

Finally, unlike the above two papers that addressed the topic of what triggers the geomagnetic storms, Paper III dealt with the study of their recovery phase and specifically what determines a slow restoration of the Earth's magnetosphere to its equilibrium conditions. Specifically, the aim was to establish, on a statistical basis, the relationship between long recovery phases and sustained periods of Alfvénic plasma streams that follow the solar event (either recurrent, such as CIRs, or non-recurrent, such as CMEs) driving the geomagnetic disturbance. By defining thresholds for the magnetospheric quiet state and Alfvénicity (i.e., the level of correlation between magnetic and velocity fluctuations, Grappin et al., 1982, measured by the *Wind* spacecraft), it was possibile to quantify the extent of magnetospheric recovery phases (through inspection of the SYM-H geomagnetic index, which is essentially the same as the Dst index, but provided at a higher time resolution, $\Delta t_{\mathrm{SYM\text{-}H}}$) and concurrent Alfvénic solar wind flows ($\Delta t_{\rho_{vb}}$). Their statistical correlation was thus proved on a period covering 16 years from 2005 to 2021: the results are shown in panel (C) of **Figure 1** as blue open circles, where they are fitted with a linear function (red dashed line), which provides a high correlation coefficient ρ of 0.83. It thus clearly emerges that Alfvénic fluctuations counteract the processes involved in a rapid restoration of the magnetospheric ring current system to its pre-storm equilibrium condition.

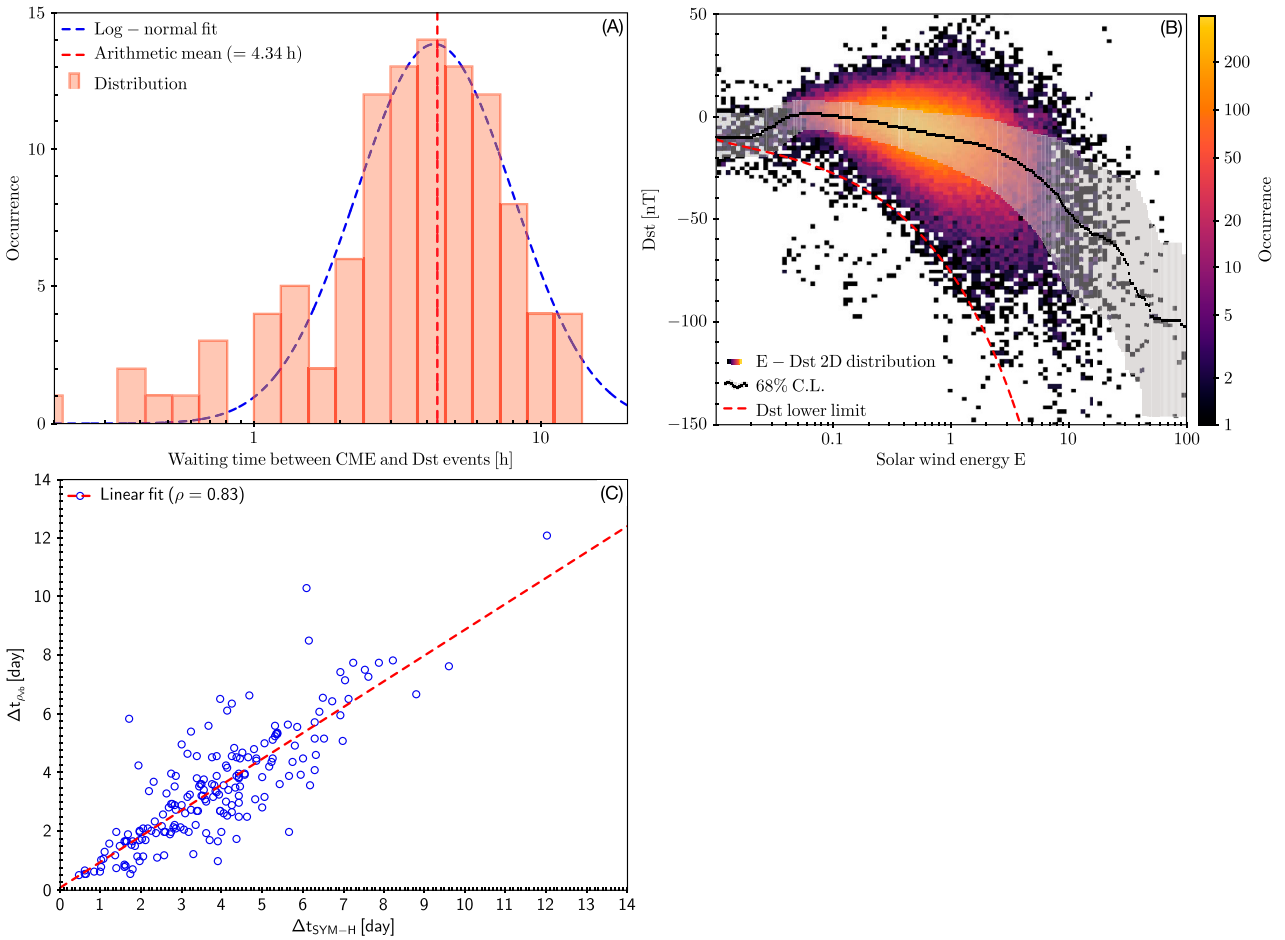

FIGURE 1 | (A) Distribution of the waiting times between the detection of the geo-effective CME and the onset of the induced geomagnetic storm; the arithmetic mean (vertical dashed red line) and the best-fitting lognormal function (blue dashed curve) are also displayed. **(B)** *E* – Dst 2D distribution; the most probable Dst value and the lower Dst threshold are marked, as a function of *E*, as a black solid and red dashed line, respectively; the gray dashed area denotes the 68% probability interval of Dst per energy bin. **(C)** $\Delta t_{SYM-H} - \Delta t_{\rho_{vb}}$ scatter-plot (blue open circles) fitted by a linear function (red dashed line) with correlation coefficient $\rho = 0.83$. Adapted from Figure 4 of Telloni et al. (2021).

3 APPLICATIONS TO SPACE WEATHER SCIENCE

Studies of the solar wind and its effects on Earth performed on a statistical basis allow both a deeper understanding of the physical processes underlying the Sun–Earth relations and an advanced capability in forecasting geomagnetic storm events within the framework of Space Weather science. Panel (A) of **Figure 1** sheds light, for instance, on the time delay between the CME passage and the onset of its magnetospheric effect. This waiting time is the combination of the time interval the CME needs to travel the distance between L1 and Earth with the time required for the CME to trigger the geomagnetic storm, perturbing the magnetospheric current system. As a conclusion, Paper I clearly pointed out that, once detected at L1, 98% of CMEs take between 2 and 8 h to initiate the geomagnetic disturbance, with an average time of about 4 h. This piece of information is particularly important in Space Weather perspective. Subtracting from this delay the CME transit time to Earth (about 30 min (1 h) for the fastest (slowest) CMEs), it appears that the complex (and not yet fully understood) processes involved in intensifying the magnetospheric ring currents take on average 3 −3 h and a half to lead the magnetosphere out of its equilibrium configuration. This result can be easily extended to any solar event, because it can be argued that the processes involved in the response of the Earth's magnetosphere to the Sun's activity do not depend on the particular type of solar driver triggering the geomagnetic storm.

Another crucial question for Space Weather is: once destabilized by a solar event how long does it take the magnetosphere to recover its equilibrium condition? The answer, by no means straightforward since the recovery phase is governed by multiple and competing restoring forces, is nevertheless of paramount importance for all those ground or space-based facilities that, in addition to being affected by the episodic and abrupt magnetospheric reconfiguration due to the impact (in most cases, but not only) of CMEs on Earth, are

equally affected by time-integrated effects throughout the whole storm. Paper III established that a correlation between long recovery phases of geomagnetic storms and the presence of Alfvénic turbulent plasma flows exists on a statistical basis (Panel (C) of **Figure 1**). Specifically, the duration of the recovery phase, when controlled by Alfvénic fluctuations, is 0.88 (which is the slope inferred from the $\Delta t_{\mathrm{SYM-H}} - \Delta t_{\rho_{vb}}$ scatter-plot) times the time length of the Alfvénic stream. Implications for Space Weather science thus stem from the possibility of forecasting the passage (and extent) of Alfvénic solar wind streams (either due to their recurring nature during solar minima or by means of the most advanced models for simulating and predicting the Parker-spiral solar wind, such as Enlil (Odstrcil, 2003) or EUHFORIA (Pomoell and Poedts, 2018)) and, through this, the duration of the recovery phase of any geomagnetic event eventually arising prior to the Alfvénic flow.

However, the most important capability that any forecasting model must have is to predict the likelihood for the solar events to impact the Earth and, if so, the intensity of the resulting geomagnetic storm. Paper III provided in this regard a useful Space Weather diagnostic tool. From the measurement at L1 of the energy load of the incoming solar wind, it is indeed possible to assess not only what will be the most likely geomagnetic activity (with the required confidence interval, black solid curve and gray shaded area in panel (B) of **Figure 1**), but also and especially the maximum response the Earth's magnetosphere could have. In fact, it is clear from the figure that the E − Dst distribution is bounded on the bottom side (red dashed curve). From a physical perspective this means that the perturbations of the ring current system are limited and strictly related to the energy input from the Sun. From a more predictive perspective, it instead allows the assessment of what will be the most severe geomagnetic disturbance that can be expected from the interaction with the magnetosphere of a solar wind carrying an energy E; or, otherwise, whether there is no need to provide an alert. Based on the above considerations, and because the measurement of solar wind energy can be performed in quasi real-time, any alert might be provided, with a confidence level of 98%, between 2 and 8 h in advance of the likely geomagnetic event.

The application of statistical methods to data acquired *in situ* from space missions orbiting L1 in the Space Weather science is being further explored and exploited in the ongoing SWESNET project of the European Space Agency, which involves about 50 research institutes/universities throughout Europe. A brief introduction of SWESNET and the author's tasks in delivering novel statistically-driven services/tools is provided in the following section.

4 OUTLOOK: THE SWESNET PROJECT

The Space Weather Service Network (SWESNET) project aims at the further development of the Space Weather services provided by the European Space Agency (ESA), drawing on the results of the Space Situational Awareness (SSA) Program. Activities include the delivery of Space Weather products and toolkits, for a timely, reliable and accurate monitoring, prediction and dissemination of Space Weather conditions and influences, via the dedicated ESA portal (https://swe.ssa.esa.int/web/guest/), which is the main resource for Space Weather in Europe. The Heliospheric Weather Expert Service Centre (ESC) is one of the five ESCs (along with Solar Weather, Space Radiation, Ionospheric Weather, and Geomagnetic Conditions) contributing to the network and deals with the effects on the Earth's environment of solar wind-related events, such as high-speed streams, CIRs, and CMEs. Characterizing, tracking, and predicting all of these interplanetary structures is vitally important for promptly reacting to the impacts of Space Weather events, thereby protecting critical infrastructures and mitigating their potentially deleterious effects.

The Solar Physics group, at the Astrophysical Observatory of Turin, part of the National Institute for Astrophysics, is one of the expert groups involved in the SWESNET Heliospheric Weather ESC and is in charge of developing several tools/prototype services for real-time analysis of space data to provide results of interest to the ESA-SSA SWESNET program and the end users. Based on the results of the three papers reviewed in this article, the author will lead the implementation into SWESNET of three new services: 1) development of diagnostic code for automatic detection and characterization of ICMEs at L1 with *in-situ* data acquired from near-Earth space observatories (arising from Paper I); 2) development of algorithm for predicting the likely geo-effectiveness of ICMEs based on local estimation of their energy content with *in-situ* data provided by spacecraft orbiting at L1 (arising from Paper II); 3) development of a tool for predicting the length of the recovery phase of the geomagnetic storm and thus estimating the time-integrated effects of sustained periods of albeit low geomagnetic activity (arising from Paper III). In addition, a preliminary investigation for the design of a machine learning-based tool for real-time prediction of geomagnetic events from solar wind measurements acquired *in situ* at L1 will be carried out, thus approaching the challenging field of machine learning techniques for Space Weather, which has received a significant boost in recent years (e.g., Camporeale, 2019).

As a conclusion, this paper reports on the statistical approach necessary to study the magnetospheric response to any solar driver and the benefits this approach may have in Space Weather studies. Only through statistical analyses it is indeed possible to ascertain which solar and geomagnetic parameters are correlated (and to what extent) in the complex solar wind-magnetosphere interaction and, specifically, to establish empirical laws useful for Space Weather purposes, with the final aim to improve the prediction capabilities and increase the robustness of the ESA-SSA SWESNET forecasting service system.

DATA AVAILABILITY STATEMENT

The results summarized in this paper refer to publicly available solar wind and geomagnetic data, stored at the NASA's Space Physics Data Facility (https://cdaweb.gsfc.nasa.gov/index.html/).

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor to this paper, having conceived and written it, and approved it for publication.

FUNDING

The author was partially supported by the Italian Space Agency (ASI) under contract 2018–30-HH.0.

ACKNOWLEDGMENTS

The author is grateful to R. D'Amicis for helpful discussions in conceiving this work.

REFERENCES

Baker, D. N., Pulkkinen, T. I., Angelopoulos, V., Baumjohann, W., and McPherron, R. L. (1996). Neutral Line Model of Substorms: Past Results and Present View. *J. Geophys. Res.* 101, 12975–13010. doi:10.1029/95JA03753

Burlaga, L., Sittler, E., Mariani, F., and Schwenn, R. (1981). Magnetic Loop behind an Interplanetary Shock: Voyager, Helios, and IMP 8 Observations. *J. Geophys. Res.* 86, 6673–6684. doi:10.1029/JA086iA08p06673

Burton, R. K., McPherron, R. L., and Russell, C. T. (1975). An Empirical Relationship between Interplanetary Conditions and Dst. *J. Geophys. Res.* 80, 4204–4214. doi:10.1029/JA080i031p04204

Camporeale, E. (2019). The Challenge of Machine Learning in Space Weather: Nowcasting and Forecasting. *Space Weather.* 17, 1166–1207. doi:10.1029/2018SW002061

Cander, L. R., and Mihajlovic, S. J. (1998). Forecasting Ionospheric Structure during the Great Geomagnetic Storms. *J. Geophys. Res.* 103, 391–398. doi:10.1029/97JA02418

D'Amicis, R., Telloni, D., and Bruno, R. (2020). The Effect of Solar-Wind Turbulence on Magnetospheric Activity. *Front. Phys.* 8, 541. doi:10.3389/fphy.2020.604857

Dumbović, M., Devos, A., Vršnak, B., Sudar, D., Rodriguez, L., Ruždjak, D., et al. (2015). Geoeffectiveness of Coronal Mass Ejections in the SOHO Era. *Sol. Phys.* 290, 579–612. doi:10.1007/s11207-014-0613-8

Dungey, J. W. (1961). Interplanetary Magnetic Field and the Auroral Zones. *Phys. Rev. Lett.* 6, 47–48. doi:10.1103/PhysRevLett.6.47

Fairfield, D. H., and Cahill, L. J. (1966). Transition Region Magnetic Field and Polar Magnetic Disturbances. *J. Geophys. Res. Space Phys.* 71, 155–169. doi:10.1029/JZ071i001p00155

Forsyth, C., Watt, C. E. J., Mooney, M. K., Rae, I. J., Walton, S. D., and Horne, R. B. (2020). Forecasting GOES 15 >2 MeV Electron Fluxes from Solar Wind Data and Geomagnetic Indices. *Space weather.* 18, e02416. doi:10.1029/2019SW002416

Gonzalez, W. D., Joselyn, J. A., Kamide, Y., Kroehl, H. W., Rostoker, G., Tsurutani, B. T., et al. (1994). What Is a Geomagnetic Storm? *J. Geophys. Res. Space Phys.* 99, 5771–5792. doi:10.1029/93JA02867

Gonzalez, W. D., Tsurutani, B. T., and Clúa de Gonzalez, A. L. (1999). Interplanetary Origin of Geomagnetic Storms. *Space Sci. Rev.* 88, 529–562. doi:10.1023/A:1005160129098

Gosling, J. T., Bame, S. J., McComas, D. J., and Phillips, J. L. (1990). Coronal Mass Ejections and Large Geomagnetic Storms. *Geophys. Res. Lett.* 17, 901–904. doi:10.1029/GL017i007p00901

Gosling, J. T., Bame, S. J., McComas, D. J., Phillips, J. L., Scime, E. E., Pizzo, V. J., et al. (1994). A Forward-Reverse Shock Pair in the Solar Wind Driven by Over-expansion of a Coronal Mass Ejection: Ulysses Observations. *Geophys. Res. Lett.* 21, 237–240. doi:10.1029/94GL00001

Grappin, R., Frisch, U., Pouquet, A., and Leorat, J. (1982). Alfvenic Fluctuations as Asymptotic States of MHD Turbulence. *Astron. Astrophys.* 105, 6–14.

Iwai, K., Shiota, D., Tokumaru, M., Fujiki, K., Den, M., and Kubo, Y. (2021). Validation of Coronal Mass Ejection Arrival-Time Forecasts by Magnetohydrodynamic Simulations Based on Interplanetary Scintillation Observations. *Earth, Planets Space* 73, 9. doi:10.1186/s40623-020-01345-5

Klein, L. W., and Burlaga, L. F. (1982). Interplanetary Magnetic Clouds at 1 AU. *J. Geophys. Res. Space Phys.* 87, 613–624. doi:10.1029/JA087iA02p00613

Manoharan, P. K., Gopalswamy, N., Yashiro, S., Lara, A., Michalek, G., and Howard, R. A. (2004). Influence of Coronal Mass Ejection Interaction on Propagation of Interplanetary Shocks. *J. Geophys. Res. Space Phys.* 109, A06109. doi:10.1029/2003JA010300

Möstl, C., Amerstorfer, T., Palmerio, E., Isavnin, A., Farrugia, C. J., Lowder, C., et al. (2018). Forward Modeling of Coronal Mass Ejection Flux Ropes in the Inner Heliosphere with 3DCORE. *Space Weather.* 16, 216–229. doi:10.1002/2017SW001735

Odstrcil, D. (2003). Modeling 3-D Solar Wind Structure. *Adv. Space Res.* 32, 497–506. doi:10.1016/S0273-1177(03)00332-6

Pomoell, J., and Poedts, S. (2018). EUHFORIA: European Heliospheric Forecasting Information Asset. *J. Space Weather. Space Clim.* 8, A35. doi:10.1051/swsc/2018020

Russell, C. T., and McPherron, R. L. (1973). Semiannual Variation of Geomagnetic Activity. *J. Geophys. Res. Space Phys.* 78, 92. doi:10.1029/JA078i001p00092

Schwenn, R., dal Lago, A., Huttunen, E., and Gonzalez, W. D. (2005). The Association of Coronal Mass Ejections with Their Effects Near the Earth. *Ann. Geophys.* 23, 1033–1059. doi:10.5194/angeo-23-1033-2005

Telloni, D., Bruno, R., D'Amicis, R., Pietropaolo, E., and Carbone, V. (2012). Wavelet Analysis as a Tool to Localize Magnetic and Cross-Helicity Events in the Solar Wind. *Astrophys. J.* 751, 19. doi:10.1088/0004-637X/751/1/19

Telloni, D., Perri, S., Bruno, R., Carbone, V., and D'Amicis, R. (2013). An Analysis of Magnetohydrodynamic Invariants of Magnetic Fluctuations within Interplanetary Flux Ropes. *Astrophys. J.* 776, 3. doi:10.1088/0004-637X/776/1/3

Telloni, D., Antonucci, E., Bemporad, A., Bianchi, T., Bruno, R., Fineschi, S., et al. (2019). Detection of Coronal Mass Ejections at L1 and Forecast of Their Geoeffectiveness. *Astrophys. J.* 885, 120. doi:10.3847/1538-4357/ab48e9

Telloni, D., Carbone, F., Antonucci, E., Bruno, R., Grimani, C., Villante, U., et al. (2020). Study of the Influence of the Solar Wind Energy on the Geomagnetic Activity for Space Weather Science. *Astrophys. J.* 896, 149. doi:10.3847/1538-4357/ab91b9

Telloni, D., D'Amicis, R., Bruno, R., Perrone, D., Sorriso-Valvo, L., Raghav, A. N., et al. (2021). Alfvénicity-related Long Recovery Phases of Geomagnetic Storms: A Space Weather Perspective. *Astrophys. J.* 916, 64. doi:10.3847/1538-4357/ac071f

Tsurutani, B. T., Gonzalez, W. D., Gonzalez, A. L. C., Tang, F., Arballo, J. K., and Okada, M. (1995). Interplanetary Origin of Geomagnetic Activity in the Declining Phase of the Solar Cycle. *J. Geophys. Res. Space Phys.* 100, 21717–21734. doi:10.1029/95JA01476

Verbeke, C., Mays, M. L., Temmer, M., Bingham, S., Steenburgh, R., Dumbović, M., et al. (2019). Benchmarking CME Arrival Time and Impact: Progress on Metadata, Metrics, and Events. *Space Weather.* 17, 6–26. doi:10.1029/2018SW002046

Vourlidas, A. (2014). The Flux Rope Nature of Coronal Mass Ejections. *Plasma Phys. Control. Fusion* 56, 064001. doi:10.1088/0741-3335/56/6/064001

Vršnak, B., and Gopalswamy, N. (2002). Influence of the Aerodynamic Drag on the Motion of Interplanetary Ejecta. *J. Geophys. Res. Space Phys.* 107, 1019. doi:10.1029/2001JA000120

Vršnak, B., and Žic, T. (2007). Transit Times of Interplanetary Coronal Mass Ejections and the Solar Wind Speed. *Astron. Astrophys.* 472, 937–943. doi:10.1051/0004-6361:20077499

Vršnak, B., Žic, T., Vrbanec, D., Temmer, M., Rollett, T., Möstl, C., et al. (2013). Propagation of Interplanetary Coronal Mass Ejections: The Drag-Based Model. *Sol. Phys.* 285, 295–315. doi:10.1007/s11207-012-0035-4

Webb, D. F., and Howard, T. A. (2012). Coronal Mass Ejections: Observations. *Living Rev. Sol. Phys.* 9, 3. doi:10.12942/lrsp-2012-3

Yermolaev, Y. I., Yermolaev, M. Y., Zastenker, G. N., Zelenyi, L. M., Petrukovich, A. A., and Sauvaud, J. A. (2005). Statistical Studies of Geomagnetic Storm Dependencies on Solar and Interplanetary Events: a Review. *Planet. Space Sci.* 53, 189–196. doi:10.1016/j.pss.2004.09.044

Yermolaev, Y. I., Nikolaeva, N. S., Lodkina, I. G., and Yermolaev, M. Y. (2012). Geoeffectiveness and Efficiency of CIR, Sheath, and ICME in Generation of Magnetic Storms. *J. Geophys. Res. Space Phys.* 117, A00L07. doi:10.1029/2011JA017139

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Astronomy and Space Sciences

ORIGINAL RESEARCH
published: 23 June 2022
doi: 10.3389/fspas.2022.896632

Towards the Identification and Classification of Solar Granulation Structures Using Semantic Segmentation

S. M. Díaz Castillo[1,2]*, *A. Asensio Ramos*[3,4], *C. E. Fischer*[5] and *S. V. Berdyugina*[1,2]

[1]Leibniz-Institut für Sonnenphysik (KIS), Freiburg, Germany, [2]Physikalisches Institut, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany, [3]Instituto de Astrofísica de Canarias (IAC), San Cristóbal de La Laguna, Spain, [4]Departamento de Astrofísica, Universidad de La Laguna, San Cristóbal de La Laguna, Spain, [5]National Solar Observatory (NSO), Boulder, CO, United States

OPEN ACCESS

Edited by:
Bala Poduval,
University of New Hampshire,
United States

Reviewed by:
Reinaldo Roberto Rosa,
National Institute of Space Research (INPE), Brazil
Herbert Muthsam,
University of Vienna, Austria
Jerome Ballot,
UMR5277 Institut de Recherche en Astrophysique et Planétologie (IRAP), France

***Correspondence:**
S. M. Díaz Castillo
smdiazcas@leibniz-kis.de

Specialty section:
This article was submitted to Astrostatistics, a section of the journal Frontiers in Astronomy and Space Sciences

Received: 15 March 2022
Accepted: 23 May 2022
Published: 23 June 2022

Citation:
Díaz Castillo SM, Asensio Ramos A, Fischer CE and Berdyugina SV (2022) Towards the Identification and Classification of Solar Granulation Structures Using Semantic Segmentation. Front. Astron. Space Sci. 9:896632. doi: 10.3389/fspas.2022.896632

Solar granulation is the visible signature of convective cells at the solar surface. The granulation cellular pattern observed in the continuum intensity images is characterised by diverse structures e.g., bright individual granules of hot rising gas or dark intergranular lanes. Recently, the access to new instrumentation capabilities has given us the possibility to obtain high-resolution images, which have revealed the overwhelming complexity of granulation (e.g., exploding granules and granular lanes). In that sense, any research focused on understanding solar small-scale phenomena on the solar surface is sustained on the effective identification and localization of the different resolved structures. In this work, we present the initial results of a proposed classification model of solar granulation structures based on neural semantic segmentation. We inspect the ability of the *U-net* architecture, a convolutional neural network initially proposed for biomedical image segmentation, to be applied to the dense segmentation of solar granulation. We use continuum intensity maps of the IMaX instrument onboard the *Sunrise I* balloon-borne solar observatory and their corresponding segmented maps as a training set. The training data have been labeled using the multiple-level technique (MLT) and also by hand. We performed several tests of the performance and precision of this approach in order to evaluate the versatility of the *U-net* architecture. We found an appealing potential of the *U-net* architecture to identify cellular patterns in solar granulation images reaching an average accuracy above 80% in the initial training experiments.

Keywords: solar physics, solar granulation, photosphere–convection, dense segmentation, deep learning–artificial neural network

1 INTRODUCTION

The solar photosphere is the lowest visible layer of the solar atmosphere, where the solar plasma changes from almost completely opaque to almost completely transparent, forming the so-called solar surface (Stix, 2002). Continuum intensity images of this layer reveal the existence of the solar granulation. It covers most of the solar surface and is characterized by a recurrent and dynamical cellular pattern. Individual elements are called granules, which are relatively small and bright bubble-like structures with horizontal scales in the order of megameters (10^3 km) evolving on timescales of minutes (Nordlund et al., 2009). Solar granules are evidence of the overturning convection process

occurring at the solar interior, where hot plasma rises at their centre, then cools down and sinks downward at the edges (Stix, 2002). An *intergranular region* forms when the granule's cool plasma drives down into the solar interior. This relatively darker narrow lane surrounding the granules is another identifiable structure at the solar surface. (Nordlund et al., 2009).

Detailed studies of small-scale phenomena on the solar surface have shown specific and systematic morphological patterns in the granulation. A type of pattern that has been studied extensively is the so-called Exploding granules. They were first described by Carlier et al. (1968) as special types of granules with sizes 2–3 times bigger than regular ones, being a product of their rapid horizontal expansion. Based on their morphology, exploding granules are characterized by a reduction in the continuum intensity in their centre, generating a "dark dot", which eventuality evolves by fragmenting (Kitai and Kawaguchi, 1979; Namba, 1986; Hirzberger et al., 1999). Several observational and numerical studies revealed that exploding granules have a close relationship with mesogranular dynamics (Domínguez Cerdeña, 2003; Roudier et al., 2003; Roudier and Muller, 2004), small-scale magnetic field diffusion and concentration (Roudier et al., 2016; Malherbe et al., 2018), and small-scale magnetic flux emergence (De Pontieu, 2002; Palacios et al., 2012; Rempel, 2018; Guglielmino et al., 2020). Another extensively studied pattern are Bright points, point-like bright elements localized within intergranular lanes and which can be clearly identified in certain photospheric spectral bands such as the Fraunhofer's G band (Muller and Roudier, 1984). Those are mostly related with magnetic field elements, being perfect tracers of high magnetic field concentrations in intensity images ((Bellot Rubio and Orozco Suárez, 2019) and references therein). More recently, Granular lanes have been reported as another subgranular pattern of interest (Steiner et al., 2010). Those are arch-like signatures moving from the boundary of a granule into the granule itself. In general, they do not completely cover the granules and are associated with a linear polarisation signal, which corresponds to the emergence of horizontal magnetic fields (Fischer et al., 2020). Granular lanes were described in simulations as signatures of underlying tubes of vortex flow with their axis oriented parallel to the solar surface (Steiner et al., 2010).

The capabilities of the new and upcoming solar telescopes (Daniel K. Inouye Solar Telescope–DKIST (Rimmele et al., 2020) or Balloon-borne telescope Sunrise III (Solanki et al., 2017)) will provide us with large amounts of unprecedented high-resolution images, which could reveal the next level of complexity of granulation. The statistical study of photospheric plasma dynamics at this level of resolution will rely on the correct identification, classification and localization of systematic structures. For this specific task, automatic solutions can be implemented, for instance, Machine Learning techniques (ML) have demonstrated promising results in classification tasks on solar images (Armstrong and Fletcher, 2019; Love et al., 2020; Baek et al., 2021; Chola and Benifa, 2022). The demonstrated effectiveness of those algorithms in pattern identification tasks has motivated us toward the exploration of Deep Learning (DL) in semantic segmentation tasks, i.e., producing automatically labelled maps at the pixel level in order to rapidly distinguish diverse granulation patterns, such as described previously.

Machine Learning techniques have acquired high popularity in resolving diverse problems in daily life during the last decade. For instance, giving computers the ability to learn representations without being directly programmed for a specific task has been extensively leveraged in computer vision (Sebe et al., 2005). Convolutional Neural Networks (CNNs) were particularly developed for image recognition tasks (Le Cun et al., 1997; Krizhevsky et al., 2012). Inspired by biological visual perception, CNNs are trained to react to specific image features, starting from simple forms, as lines or edges, and then detecting more complex and abstract patterns in subsequent layers (Ghosh et al., 2020). Sequentially combining layers inside the network to progressively extract higher-level features is the main line of the DL success (Aloysius and Geetha, 2017). Taking advantage of large amounts of data, this approach may achieve unprecedented performance on several perception tasks, e.g., instance classification (Simonyan and Zisserman, 2015; Huang et al., 2017), object detection (Girshick, 2015) or optical flow estimation (Ilg et al., 2017).

Another task that saw an important push forward with DL was dense prediction, i.e., prediction at a pixel level in images, such as semantic segmentation (Shelhamer et al., 2017; Chen et al., 2018), which solves the classification problem working at pixel resolution. More specifically, the aim is to group the pixels of an image into categories, providing precise localization of labeled structures. Additionally, semantic segmentation seeks to partition the image into semantic meaningful parts (Szeliski, 2011). This paradigm has been successfully addressed using Encoder-Decoder architectures (Badrinarayanan et al., 2017; Yanli and Pengpeng, 2021). Leveraging the properties of CNNs, this type of architecture is capable of producing spatially consistent classification maps, thus providing precise localization of objects of interest.

In this work, we propose to train supervisedly and evaluate the performance of a CNN to carry out solar granulation segmentation. To this end, we apply an encoder-decoder architecture called *U-net* (Ronneberger et al., 2015). This architecture was developed for biomedical image segmentation tasks, and it is especially interesting for our objectives since it has been successfully applied to cellular pattern segmentation. It can work with few training images, and it achieves high levels of accuracy in the localization of specific structures (Ronneberger et al., 2015).

2 METHODS

2.1 *U-Net* Architecture

A *U-net* is composed of fully CNN layers organized in an encoder-decoder architecture (Ronneberger et al., 2015)[1]. The encoder part (left side of **Figure 1**) is responsible for producing a low

[1]More information can be found at https://lmb.informatik.uni-freiburg.de/people/ronneber/u-net/

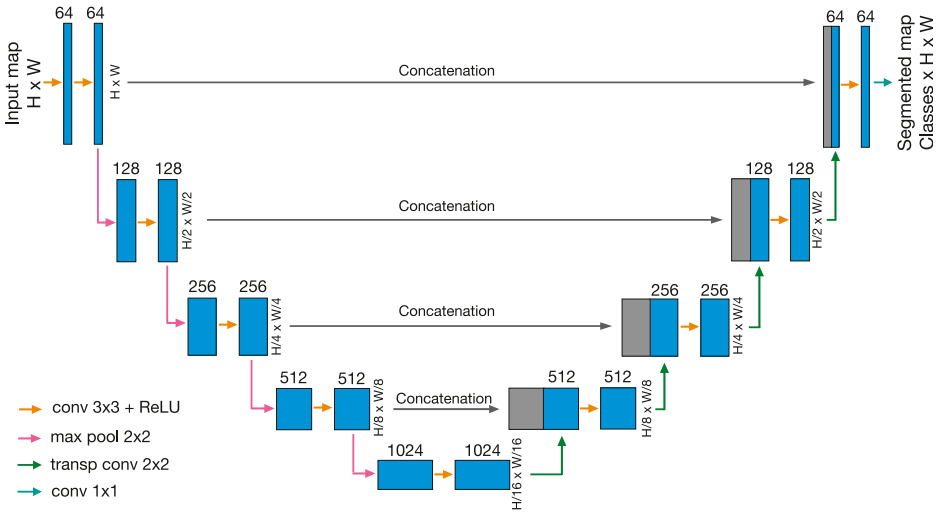

FIGURE 1 | Schematic sketch of the *U-net* architecture used, going sequentially from left to right. Each blue box corresponds to a feature map. The depth of the map is denoted on top of the box. Gray boxes represent copied feature maps. The arrows denote the different operations. We use 64 kernels (feature information) for the initial kernel set, whose number increases sequentially in the contraction levels and decreases in the expansion levels as seen in the feature maps depth values. Based on the intrinsic properties of a fully convolutional neural network, the height (H) and width (W) of the input images can be arbitrary numbers but must be equal (squared maps). For the training procedure, we use input maps of sizes of 128 × 128. Sketch modified from (Ronneberger et al., 2015).

dimensionality dense representation of an image. In the initial stage, the contracting network receives an image of a specific size H (height) × W (width), which is downsampled by sequential layers, each composed by the following operations:

1. Two 3 × 3 padded convolution operations each followed by a rectified linear unit (entire operation represented by orange arrows in the **Figure 1**). As the fundamental components of the convolutional neural network, 2D convolution operations transform the image into feature maps using a set of filters or *kernels*. Those resulting feature maps then pass through a non-linear activation function. We use a set of 64 3 × 3 kernels generating features maps of 64 (depth) × H × W in the initial operation. Then for the subsequent one, each kernel will have a depth dimension, which corresponds to the feature map depth previously generated. We used padded operations in order to not change the size of the input map during the convolution operations (Dumoulin and Visin, 2016). *U-net* convolutional operations use a rectified linear unit (ReLU) as the default activation function, which gives the non-linear character to the network. This function is characterized by being linear for input positive values and zero for input negative values. It is well-behaved and converges fast when using the stochastic gradient descent algorithm. Consequently, it is commonly used as an activation function in deep neural networks (Schmidhuber, 2014).
2. A 2 × 2 max-pooling operation with stride 2, which reduces the dimensions of the input map by computing the maximum value of each successive 2 × 2 pixel set to produce a downsampled map (pink arrows in **Figure 1**). During this process, the spatial information is reduced by a factor of two, while the feature information is increased by a factor of two.

When the lower level is reached, the lower feature map is then expanded by upsampling sequential layers in the decoder part (right side of the **Figure 1**), which is responsible for recovering the initial spatial dimension. This expanding network consists of upsampled layers, each composed by:

1. Two 3 × 3 padded convolution operations, each followed by a rectified linear unit (orange arrows in the **Figure 1**) equivalent to the operations of the encoder part.
2. A 2 × 2 transposed convolution with stride 2 as the upsampling operation (green arrows in **Figure 1**). Those seek to reverse the encoder downsampling operations, while broadcasting input elements via a set of 2 × 2 kernels, thereby producing an output that is larger than the input (Dumoulin and Visin, 2016). During this process, the spatial information is increased by a factor of two, while the feature information is reduced by a factor of two.

As the outstanding component of the *U-net* architecture, each expansion layer is concatenated with high-resolution features from the encoder path at the same level (see grey blocks in **Figure 1**), giving the network the capacity to localize with precision. We use five levels of contraction and expansion, like in the original *U-net* model, giving it its characteristic symmetrical shape. Finally, at the end of the sequence, a 1 × 1 convolution operation produces probability maps per class as output with the same sizes as the original map. Using the configuration shown in **Figure 1**, our model employs around 31 million trainable parameters or hyperparameters.

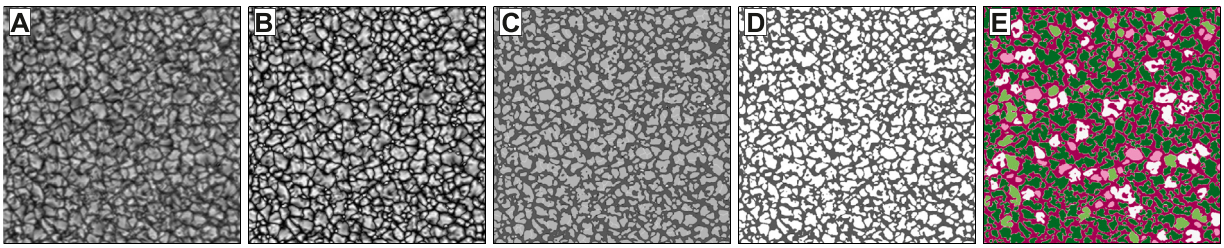

FIGURE 2 | Frame example of IMaX/Sunrise during the labeling procedure. **(A)** Reconstructed continuum intensity map, **(B)** Initial segmentation results using 25 descending detection thresholds, **(C)** Merging and shrinking results, **(D)** Result map differentiating integranular lanes and granules cells as single units, **(E)** Manual selection into defined categories: intergranular lane (dark violet), uniform-shaped granules (pink), granules with a dot (white), granules with a lane (light green) and complex-shaped granules (dark green).

2.2 Ground-Truth Data

2.2.1 Observational Dataset: IMaX/Sunrise

In order to train the model using supervised training, we should provide it with a suitable set of data for the segmentation process, i.e., a ground-truth. We are interested in classifying specific patterns in observational images of the solar surface seen in the continuum. Based on our requirements of high-spatial-resolution data, we select the products of the Imaging Magnetograph eXperiment (IMaX) (Martínez Pillet et al., 2011), the filterpolarimeter onboard the Sunrise I balloon-borne telescope during its flight in June 2009 (Barthol et al., 2011; Solanki et al., 2010). IMaX was tuned to the Fe I at 525.02 nm highly Zeeman-sensitive line, and provided measurements of the local continuum intensity near this line. After phase diversity reconstruction, each map reached a spatial resolution of around 0".15 ≈ 100 km over the solar surface (pixel scale 0".05) and a field of view (FOV) of 50" × 50" ≈ 35 × 35 Mm (Martínez Pillet et al., 2011). Those data products are freely accessible on https://star.mps.mpg.de/sunrise/. We select a time series of 56 min with a cadence of 30 s resulting in 113 individual frames. We selected the frames with the highest quality and spread out in time to obtain as much as possible a diverse data set. Due to the apodization needed for image reconstruction, a portion of the edges was lost. Taking all the above in consideration, our ground-truth dataset is composed of eight frames of 768 × 768 pixels each, with a FOV of 38" × 38" (see one frame as example in **Figure 2** map A).

2.2.2 Labeling Structures

In a supervised learning approach, we need to provide an initial truth segmentation of our selected dataset in order to properly train the model. In previous studies, the identification and tracking of specific granular structures have been done mostly manually with the help of intensity multi-threshold algorithms (Javaherian et al., 2014; Ellwarth et al., 2021). For our experiment, we select a common multi-threshold algorithm MLT4 (Bovelet and Wiehr, 2001; Bovelet and Wiehr, 2007) used for segmenting photospheric structures for the initial granular identification (see for instance (Riethmüller et al., 2008; Fischer et al., 2019; Kaithakkal and Solanki, 2019)) which is freely available[2]. We adopt this approach to assess the extent to which user intervention affects the training process on the network. In particular, for labeling our structures at a pixel level we follow a procedure composed of two phases:

1. Semi-automatic granules identification: Using the Multiple-Level Pattern Recognition algorithm–MLT4 (Bovelet and Wiehr, 2001; Bovelet and Wiehr, 2007), we segregated the intergranular regions and the granular pattern. This is a top-down segmentation technique of brightness structures based on a sequence of descending detection thresholds (Bovelet and Wiehr, 2007). The algorithm uses the reconstructed and normalized continuum intensity maps as input (see one frame as an example in **Figure 2** map A). The procedure starts with an initial segmentation of features at equidistant intensity levels as shown in map B of **Figure 2**, and then the pixel brightness level is normalized within each cell to its maximum value. Consecutively, a semi-automatic procedure of merging over-segmented cells and (4) shrinking these brightness-normalized cells to features of adequate sizes is performed, resulting in maps such as map C of **Figure 2**. Regarding the setup parameters used, we selected 25 descending thresholds, 0.47 as a normalized reference for merging and 0.38 as the unitary cut threshold for shrinking. The unitary cut threshold controls the final size of the recognized features, initially derived from a normalized brightness histogram for the full sample of recognized cells, which was then tuned by visual inspection. The rest of input parameters were set to their default value (Bovelet and Wiehr, 2001). The resulting maps are composed of several hundred individual cells (granules) separated from the intergranular space as shown in map D of the **Figure 2**.
2. Fully manual granules classification: Based on the basic instantaneous morphological features of the granular phenomena that we seek to classify, we propose an initial set of granule classes characterised by the presence of a central dot signatures or an arch-like lane signatures. For completeness, we include two categories that refer to extreme levels of complexity in granules: 1) morphologies with low complexity, i.e. uniform and clean morphologies with circular or ellipsoid shapes, and 2) morphologies with high complexity, i.e. abnormal granules having combinations of dark spots or lanes. In that sense and using the map products

[2]All code and documentation can be found at http://wwwuser.gwdg.de/astronom/

of the previous procedure, we classified manually the set of individual cells into four categories: granules with a dot (cells with dark point-shape features close to the centre of the cell), granules with a lane (cells with a dark arch-like lane following a bright rim mark inside the cell structure), uniform-shaped granules (cells with uniform intensity distribution and with elliptical or circular shapes) and complex-shaped granules (all remaining cells). The map E of the **Figure 2** shows the results of the manual classification, where each colour corresponds to a specific classification, equivalently to the ground-truth maps in the **Figure 4A**; **Figure 5A**. During the selection via visual inspection, we pursue to classify all individual granules per class unequivocally to avoid ambiguity.

We perform this two-step procedure for all pixels of the eight selected frames, generating one ground-truth labeled map for each continuum intensity map. We are interested in evaluating independently and unbiased the performance of our model and simultaneously providing it with as many training examples as possible, thus we split our dataset in such a way that seven frames are used for the training set and one is used for the validation/test set. As an example, **Figure 2** shows the intermediate steps in the complete labeling procedure for the validation/test map.

2.3 Training Strategy

Although the *U-net* architecture has demonstrated a good performance even with a few per-class training examples, it is essential to provide it with a large and diverse set of training data. In that sense, we divided the full FOV of all available maps into several sub-maps of a fixed size. As we are interested in predicting the class of each granule, we select sub-maps of the size 128×128 pixels ($\sim 6.5" \times 6.5"$) as input, with the aim of covering entire granules (see **Figure 1**). In addition, we applied a process of data augmentation, including random rotations, random perspective transformations and warping.

We identified a severely skewed class distribution in the labeled data, where 85% of the pixels of all available maps are associated to two classes (intergranular lane 40% and complex-shaped granules 45%) and the remaining 15% of the pixels belong to the underrepresented classes (granules with a dot 8%, granules with a lane 3% and uniform-shaped granules 4%). This is a known difficulty that affects all classification machine learning algorithms because the metrics used for training assume an equivalent proportion of examples of each class. This assumption decreases the performance of the model for underrepresented classes (He and Garcia, 2009; Fernández et al., 2018). Many strategies have been developed to overcome this issue in computer vision paradigms [see, e.g., (He et al., 2008; He and Garcia, 2009; He and Ma, 2013; Huang et al., 2016; Khan et al., 2017; Oksuz et al., 2020)], however, it is still an active topic research in semantic segmentation tasks [see, e.g., (Havaei et al., 2017; Olã Bressan et al., 2022; Zou et al., 2021)].

We addressed the imbalance-class issue in this work by including a stratified random sub-map sampling previous to the augmentation procedure as follows. 1) We defined weighted pixel maps for each full image, in which the greater weights were given to areas where underrepresented classes were localized. 2) We applied a softmax function to compute probability distribution maps. Those probability distributions were included in a weighted random choice function, which returned sub-maps centred on underrepresented classes regions. With this method, we increased the pixel proportion of the underrepresented classes to 22% in our training dataset. We noticed that this proportion has an upper limit due to the size of the sub-maps. The reason is that the surface covered by underrepresented classes is smaller than the size of the sub-maps, which is mostly covered by the background classes (i.e., intergranular regions and several complex-shaped granules).

An additional strategy towards solving class imbalance is an appropriate selection of the loss function. Neural networks applications learn via optimization, which requires a suitable cost/loss functions to calculate the model error. The iterative process of hyperparameter tuning is controlled by the loss function minimization, which, at the end of the training, ideally provides the best model setup for the assigned task. In particular, metrics for semantic segmentation have been historically dominated by global approaches, like the Cross-Entropy loss (Aggarwal, 2018). Defined as $CE = -\log(p_t)$, where p_t corresponds to the estimated probability for a correct classification for a specific class t, the cross-entropy loss evaluates the overall proportion of the correctly classified pixels as the precision measurement. However, these scores are dominated by the background classes in skewed datasets. Typically, the addition of a cost-sensitive weighting factor α is used in cross-entropy, known as α-balance variant. This seeks to balance the importance of well-classified over the wrong-classified examples in cases of skewed datasets. For several classes, the α factor can be considered as a weight vector with values inversely proportional to the frequency of each class (Lin et al., 2017).

For our experiments, we test the accuracy and effectiveness of two different loss functions during the network training, which are commonly used for imbalanced data problems:

1. The Focal Loss was developed for addressing the unbalance-class problem by adding a modulating factor $(1 - p_t)^{\gamma}$ to the cross-entropy loss. It uses the tunable focusing parameter $\gamma \geq 0$, which adjusts the rate at which background examples are down-weighted. This modification downplays the importance of the background classes, making the training to focus on learning the hard examples, i.e. weakly represented classes (Lin et al., 2017). The use of α-balance variant is also applicable in this case.
2. The Intersection-over-Union (IoU) loss or Jaccard index was extensively used in semantic segmentation tasks. It is focused on determining the similarity between finite sample sets (Jaccard, 1912). For images, the IoU measures the agreement between any predicted region and its corresponding ground-truth region by measuring the intersection between the prediction and the ground-truth normalized by their union. The IoU loss can take into account the frequency of the classes, and thus it is considered robust to the class imbalance problem (Leivas Oliveira, 2019). For multi-class classification tasks like the one we pursue here, the mean IoU (mIoU) loss function is

often used, which initially computes the Jaccard index for each class and then computes the average over all classes.

For comparative purposes, we computed the standard evaluation metrics for semantic segmentation. We computed the overall accuracy, measured as the ratio between the correctly predicted pixels and the total number of pixels, and the mean pixel accuracy per class measured as the average of the correctly predicted pixels per class over the total ground-truth pixels per class. Likewise, we compare the test performance with performance parameters during the execution of the training. In this sense, we monitor the average per epoch of the loss value given its loss function (average loss) and the average overall accuracy per epoch (accuracy).

3 RESULTS

We implemented the model and the training using the open-source *PyTorch* (Paszke et al., 2019) framework[3]. All our training cases have been performed with an NVIDIA GeForce 2080 Ti GPU and an NVIDIA Tesla P100 GPU. We generated 27,000 sub-maps in the training dataset and 3,000 sub-maps in the validation dataset from the augmentation procedure previously described. We used the Adam stochastic gradient algorithm (Kingma and Ba, 2014) for optimization. In this case, the gradient is estimated from subsets of the training dataset or *batches*. We used batches of 32 samples, generating around 840 subsets containing all the training dataset, which are used to update the hyperparameters and to computed the metric in each cycle or *epoch*. We considered 100 or 200 epochs depending on the loss function that we used, thus we executed around 84,000 or 168,000 iterations in total respectively. The learning rate was annealed from the initial value of 0.001 with a dynamic adjustment, lowering the value by a factor 0.9 when the minimization did not decrease in 5 consecutive epochs. We performed several training cases by turning off some of the transformations of the data augmentation process, changing the weight values for the stratified random sampling maps, and changing the loss functions and its parameters. Using this setup, each epoch took ~ 210 s.

We present two training cases that reached the highest accuracies from all testing that we ran. Test 1 uses the *U-net* architecture with a α-balanced Focal Loss. For this experiment, we consider a proportion of 1:100 for the background classes with respect to the underrepresented classes to build the weighted pixel maps for the stratified sampling selection. On the other hand, Test 2 uses the *U-net* architecture using the mean IoU as the loss function. Similarly, we consider a proportion of 1:100 for the background classes with respect to the underrepresented classes to build the weighted pixel maps. In both experiments, we only used random rotations and perspective transformations as augmentation. **Figures 3A,B** present the monitoring plots of these two test cases during the training execution. These graphs show the evolution of the performance parameters of the selected metrics per epoch, i.e., average loss and accuracy. For both cases, the performance parameters behave as expected for the training dataset (see blue curves in **Figures 3A,B**), however, the parameters associated with the validation dataset show differences between each other.

For test 1, the overall pixel accuracy, the accuracy per class and the Jaccard index for the validation dataset increase slowly over every iteration reaching 0.84, 0.60, and 0.47 respectively, while its corresponding loss increases heavily. We interpret that the noise and rising trend in the loss profile are due to the accumulation of misclassified examples, such as pixels at the edge of granules or clusters of pixels that have features of multiple classes, which the model slowly corrects thus improving the accuracy. The model reaches high levels of saturation, with hints a slight overfitting close to the end of the training (see **Figure 3A**).

This effect can be observed in the full map prediction and in the predicted probability distribution per class shown in **Figures 4A,B**. In the whole map, the model reaches 0.74 of overall pixel accuracy, a mean accuracy per class of 0.52 and a Jaccard index of 0.40. These values are slightly lower than those achieved during training (evaluated in sub-maps of size 128×128) but still compatible, since they come from the same map. **Figure 4A** displays the direct comparison between the ground-truth map with the predicted map. As a first conclusion, we highlight the efficiency of the model to segregate granules and intergranular lanes, which contributes mostly to the overall pixel accuracy.

Regarding the correct identification of underrepresented morphologies, the model behaviour is different per class. Based on the probability maps in **Figure 4B**, we identify that the model develops different reliability levels depending on the class. For clean morphologies such as uniformly shaped granules, the model is slightly more confident as compared with more structured and complex classes. In that sense, granules with multiple and similar features give rise to a prediction with a high degree of uncertainty. This is manifested in classes such as granules with a dot, granules with a lane and complex-shaped granules.

On the other hand, test 2 shows a completely different behaviour. As shown in **Figure 3B**, the performance parameters related to the validation dataset reaches a threshold at an early stage of training without appreciable changes along the epochs. From the first cycle, a value of 0.87 for overall pixel accuracy and a Jaccard index of 0.52 are achieved. However, in terms of the mean accuracy per class, the average threshold value during the first 20 iterations is 0.64, which then decreases and stabilizes around 0.60 correspondingly. Using this training setup, we suggest that the model is able to learn the basic morphological patterns from few batches, but it is unable to extract more specific information from the full dataset provided during the training. Signatures of over-fitting are also observed, but the invariance of the loss for the training and validation datasets indicates an upper limit in the learning process in the defined training time.

Figure 5A shows the full map prediction and the predicted probability distribution maps of the model with the lowest loss

[3]All codes are placed in a free access repository (https://gitlab.leibniz-kis.de/smdiazcas/SegGranules_Unet_model.git)

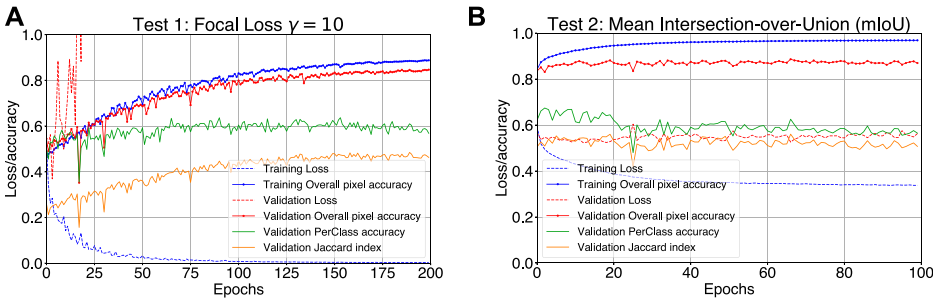

FIGURE 3 | (A) Tracking plot for the training for test 1: *U-net* model using α-balanced Focal loss with $\gamma = 10$ as training loss function. **(B)** Tracking training plot for test 2: *U-net* model using Mean IoU as training loss function.

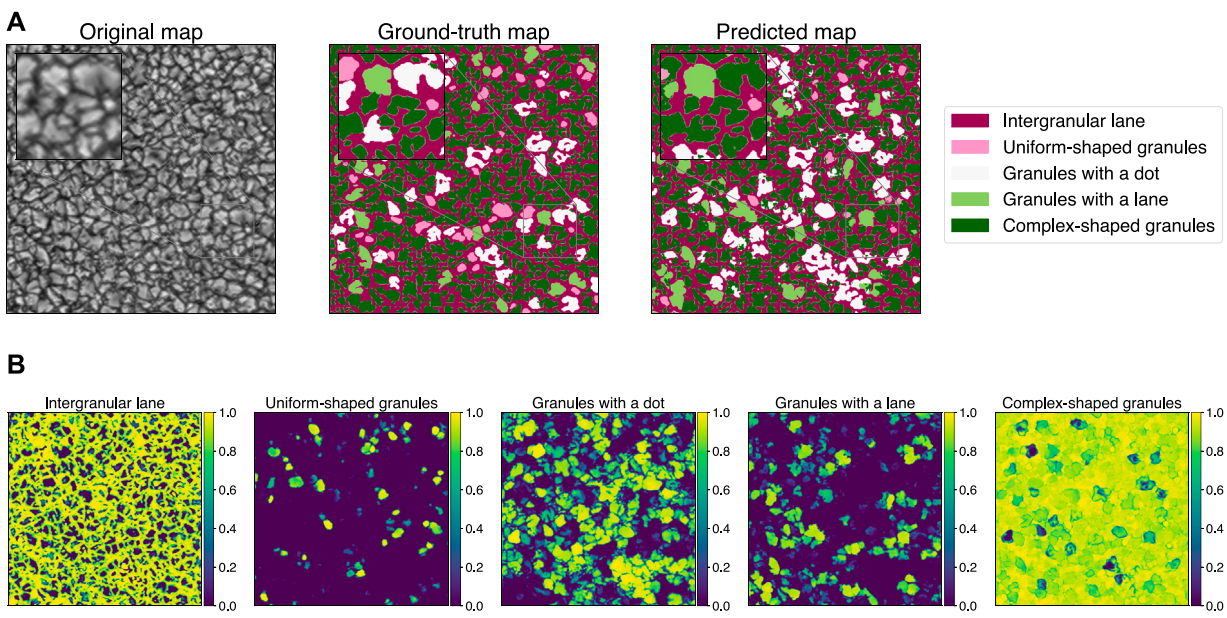

FIGURE 4 | (A) Comparative maps using the validation full size map–Test 1: *U-net* model using α-balanced Focal loss with $\gamma = 10$ as training loss function. A zoomed region is highlighted showing in detail the diverse constrains of the segmentation. **(B)** Probability maps per class using the validation full size map–Test 1: *U-net* model using α-balanced Focal loss with $\gamma = 10$ as training loss function.

value obtained during the training (epoch 12). In the whole map, the model reaches 0.71 of overall pixel accuracy, a mean accuracy per class of 0.58 and a Jaccard index of 0.40. Again, these values are slightly lower than those achieved during training (evaluated in sub-maps of size 128 × 128) but still compatible, since they come from the same map. A good efficiency to segregate granules and intergranular lanes is also obtained in this test, contributing mostly to the overall pixel accuracy as well (see the **Figure 5A**).

Based on the probability maps generated (see **Figure 5B**), we notice that the model reproduces high levels of confidence in all classes, managing to identify detailed morphological patterns associated with the classes, i.e., dots, lanes or combinations even within individual cells, which promotes the over-labelling of structures, i.e., single granules contain pixels of different classes as shown in the predicted map of **Figure 5A**.

So far, we have been referring to class-average quantities of the performance parameters, which are biased by the well-know imbalance between granulation structures. In **Table 1**, we present a summary of the overall pixel accuracy (OPA) and the Jaccard index per class. As we mentioned, the identification of the intergranular lane provides the major contribution to the effectiveness of the models, reaching accuracy values around 0.90 in both metrics. On the other hand, the identification of specific morphologies has a significantly lower performance, especially for underrepresented classes. We identify that the contribution of precision and recall on the Jaccard index are unequal, i.e. the models are acceptably sensitive (higher recall) in detecting the simple shapes characteristics of each class, but those are inexact during the ground-truth comparison.

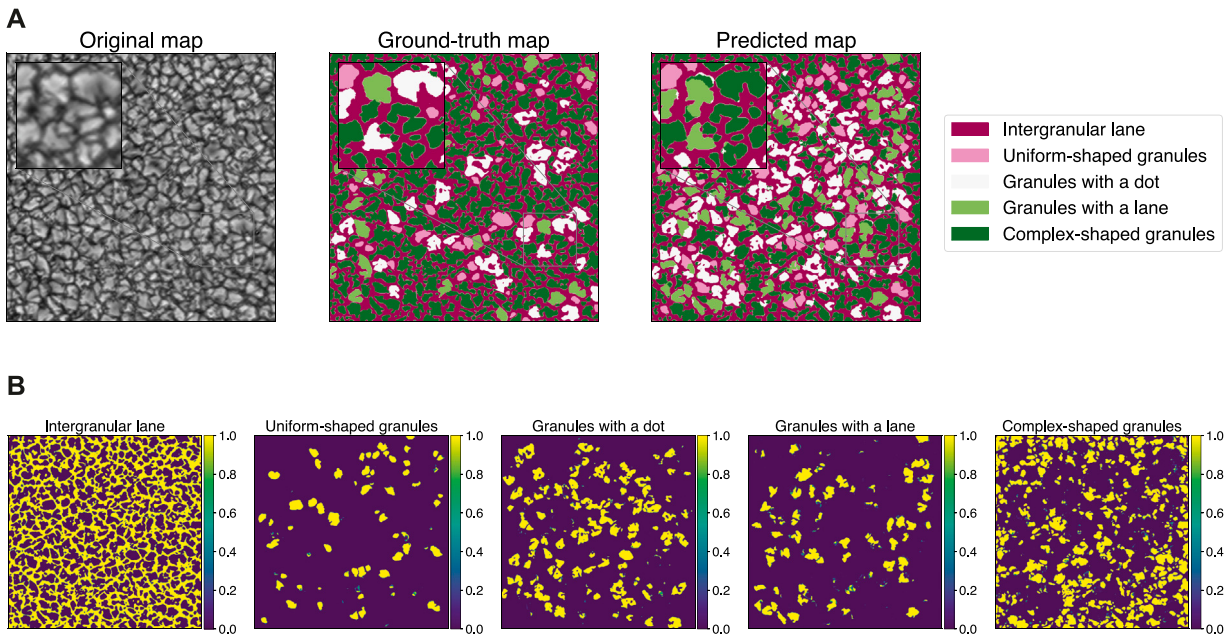

FIGURE 5 | (A) Comparative maps using the validation full size map–Test 2: *U-net* model using Mean IoU as training loss function. A zoomed region is highlighted showing in detail the diverse constrains of the segmentation. **(B)** Probability maps per class using the validation full size map–Test 2: *U-net* model using Mean IoU as training loss function.

TABLE 1 | Summary of the performance parameters per class calculated for the full size verification map prediction.

	IGL[a]		UG[b]		DG[c]		LG[d]		CG[e]	
	OPA[f]	Jaccard	OPA	Jaccard	OPA	Jaccard	OPA	Jaccard	OPA	Jaccard
Test 1: Focal Loss	0.90	0.85	0.12	0.10	0.58	0.29	0.31	0.15	0.71	0.58
Test 2: mIoU	0.95	0.90	0.44	0.20	0.67	0.31	0.33	0.11	0.54	0.49

[a]*Intergranular lane.*
[b]*Uniform-shaped granules.*
[c]*Granules with a dot.*
[d]*Granules with a lane.*
[e]*Complex-shaped granules.*
[f]*Overall Pixel Accuracy.*

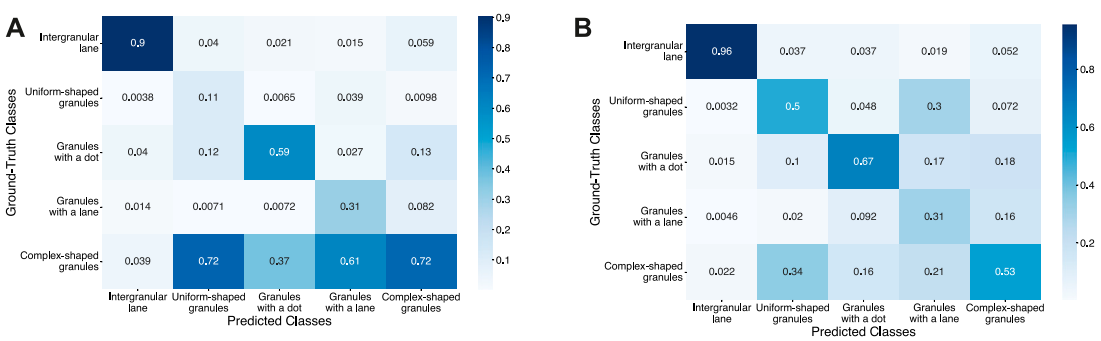

FIGURE 6 | (A) Probability distribution at pixel level of the category prediction given its original category (confusion matrix) for test 1: *U-net* model using α-balanced Focal loss with $\gamma = 10$ as training loss function. **(B)** Probability distribution at pixel level of the category prediction given its original category (confusion matrix) for test 2: *U-net* model using Mean IoU as training loss function.

Based on the multi-class confusion matrix at the pixel level for each model shown in the **Figures 6A,B**, we notice specific behavior per model. For test 1, the high uncertainty levels generate a strong effect on the maximum probability to match a category given the original category. In this case, granules belonging to under-represented classes have a high tendency to be classified as complex-shape granules due to the homogeneous probabilities between classes in granules where similar morphologies are shared. On the other hand, for test 2, the over-labeling effect and the high levels of reliability, tend to homogenize the maximum probabilities by class, negatively affecting classes such as uniform-shaped granules and granules with a lane.

4 DISCUSSION

In summary, we present the first attempts to classify and identify structures in the solar granulation based on the semantic segmentation paradigm using a deep learning method. As our main objective, we found an interesting potential of the *U-net* architecture to identify and classify cellular patterns in solar granulation images, but modifications to the current model should be implemented to ensure its optimal performance. With the proposed training procedure, the model achieves high levels of accuracy in the identification of the intergranular network which allows the effective separation of granular morphologies. We have also established that the network architecture is sensitive in identifying characteristic patterns in granules, such as granules with a dot (overall accuracy greater than 0.5 in both tests), but it looses efficacy when it comes to discerning between structures with combined morphologies, i.e., granules with multiple features and complex-shaped structures. This outcome drives high uncertainty levels (test 1) or an over-labeling effect in single granules (test 2).

During our experiments, we have identified recurrent hints of overfitting in all performed tests, meeting the highest accuracy for the tests presented here. We implemented some functional strategies such as the Dropout regularization and hyperparameters scaling but without obtaining any improvement. Going further, we identified that the preparation of the ground-truth dataset played a crucial role in the model generalization ability. The semi-manual and manual labeling process introduced unwanted constraints, e.g., over-merging, poor contours separation and small incorrect areas. Moreover, we noted the difficulty in defining closed classification criteria, which would allow us to represent the samples of each category unambiguously. Labeling structures of this specific phenomenon is a complex task, even for human classifications. The phenomena in the photosphere are so diverse that it is effectively easy to under-classify or over-classify morphologies. Thus, it is fundamental to improve the initial labeling for future supervised testing including the use of ground-truth segmentation methods that involve the least amount of user intervention in order to reduce ambiguity.

Another source of over-fitting may be related to the augmentation process, which is highly affected by the wrong-labeled data. In this case, the geometric transformations applied in the limited available labeled samples, especially for the underrepresented classes, can induce over-fitting (Shorten and Khoshgoftaar, 2019). Granules, as individual elements, are unique at a very detailed level, i.e., in super-high-resolution images. However, as we have already mentioned, they share similar phenomenologies that makes it possible to classify them into groups with comparable patterns at basic levels of similarity. Therefore, the use of extensive and random geometric transformations can produce non-deterministic effects, negatively affecting the training performance. Other strategies exist in the literature to prevent overfitting in skewed data, i.e. transfer learning (Weiss et al., 2016), pre-training (Singh Punn and Agarwal, 2021) or one-shot and zero-shot learning (Xian et al., 2017), which we plan to study in future works.

We extensively highlight these initial experiments as a starting point for further investigation. As this research is still under development, we seek to improve the levels of sensitivity and precision as much as possible to unequivocally detect the existing phenomenologies in solar granulation. We anticipate that extending our approach to include time-series, i.e., video segmentation, and other physical observables such as polarization and Doppler maps can be fruitful. This additional information would reveal other characteristics associated with the considered phenomena, allowing the definition of precise selection criteria, e.g., granular lane cases have been unambiguously detected based on the host granule evolution. Besides, the exploration of self-supervised or unsupervised methods is in our sights for further studies.

5 RESOURCE IDENTIFICATION INITIATIVE

All code of the model was constructed based on Python Programming Language, RRID:SCR_008394 version 3.9.7, and *PyTorch* libraries, RRID: SCR_018536 version 1.10.0.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://star.mps.mpg.de/sunrise/ and https://gitlab.leibniz-kis.de/smdiazcas/SegGranules_Unet_model.git.

AUTHOR CONTRIBUTIONS

SD, CF, and AA conceived of the presented idea. SD and CF prepared the ground-truth dataset. SD and AA constructed the algorithm and the training strategy. SD executed the network training and analysed the results with help of AA All authors contributed to the final version of the manuscript. SB supervised the project.

FUNDING

SMDC and CEF are funded by the Leibniz Association grant for the SAW-2018-KIS-2-QUEST project. The German contribution to Sunrise is funded by the Bundesministerium für Wirtschaft und Technologie through Deutsches Zentrum für Luft und Raumfahrt e.V. (DLR), Grant No.50 OU 0401, and by the Innovationsfond of the President of the Max Planck Society (MPG). The Spanish contribution has been funded by the Spanish MICINN under projects ESP2006-13030-C06 and AYA2009-14105-C06 (including European FEDER funds). HAO/NCAR is sponsored by the National Science Foundation, and the HAO Contribution to Sunrise was partly funded through NASA grant number NNX08AH38G. The National Solar Observatory (NSO) is operated by the Association of Universities for Research in Astronomy, Inc. (AURA), under cooperative agreement with the National Science Foundation.

ACKNOWLEDGMENTS

We acknowledge the Instituto de Astrofísica de Canarias (IAC) and the Leibniz-Institut für Sonnenphysik (KIS) for providing us with compute time on their GPUs: NVIDIA GeForce 2080 Ti GPU and an NVIDIA Tesla P100 GPU respectively, resources extensively used in this research. This research has made use of NASA's Astrophysics Data System Bibliographic Services. We acknowledge the community effort devoted to the development of the following open-source packages that were used in this work: numpy (numpy.org), matplotlib (matplotlib.org) and Pytorch (pytorch.org).

REFERENCES

Aggarwal, C. C. (2018). *An Introduction to Neural Networks*. Cham: Springer International Publishing, 1–52. doi:10.1007/978-3-319-94463-0_1

Aloysius, N., and Geetha, M. (2017). "A Review on Deep Convolutional Neural Networks," in 2017 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, April 6–8, 2017, 0588–0592. doi:10.1109/ICCSP.2017.8286426

Armstrong, J. A., and Fletcher, L. (2019). Fast Solar Image Classification Using Deep Learning and its Importance for Automation in Solar Physics. *Sol. Phys.* 294, 80. doi:10.1007/s11207-019-1473-z

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Analysis Mach. Intell.* 39, 2481–2495. doi:10.1109/TPAMI.2016.2644615

Baek, J. H., Kim, S., Choi, S., Park, J., Kim, J., Jo, W., et al. (2021). Solar Event Detection Using Deep-Learning-Based Object Detection Methods. *Sol. Phys.* 296, 160. doi:10.1007/s11207-021-01902-5

Barthol, P., Gandorfer, A., Solanki, S. K., Schussler, M., Chares, B., Curdt, W., et al. (2011). The Sunrise Mission. *Solar Phys.* 268, 1–34. doi:10.1007/s11207-010-9662-9

Bellot Rubio, L., and Orozco Suárez, D. (2019). Quiet Sun Magnetic Fields: an Observational View. *Living Rev. Sol. Phys.* 16, 1. doi:10.1007/s41116-018-0017-1

Bovelet, B., and Wiehr, E. (2001). A New Algorithm for Pattern Recognition and its Application to Granulation and Limb Faculae. *Sol. Phys.* 201, 13–26. doi:10.1023/A:1010344827952

Bovelet, B., and Wiehr, E. (2007). Multiple-Scale Pattern Recognition Applied to Faint Intergranular G-Band Structures. *Sol. Phys.* 243, 121–129. doi:10.1007/s11207-007-9010-x

Carlier, A., Chauveau, F., Hugon, M., and Rösch, J. (1968). Cinématographie à Haute Résolution Spatiale de la Granulation Photosphérique. *Acad. Des. Sci. Paris Comptes Rendus Ser. B Sci. Phys.* 266, 199–201.

Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. (2018). Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected Crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 834–848. doi:10.1109/tpami.2017.2699184

Chola, C., and Benifa, J. V. B. (2022). Detection and Classification of Sunspots via Deep Convolutional Neural Network. *Glob. Transitions Proc.* doi:10.1016/j.gltp.2022.03.006

De Pontieu, B. (2002). High-Resolution Observations of Small-Scale Emerging Flux in the Photosphere. *Astrophys. J.* 569, 474–486. doi:10.1086/339231

Domínguez Cerdeña, I. (2003). Evidence of Mesogranulation from Magnetograms of the Sun. *Astron. Astrophys.* 412, L65–L68. doi:10.1051/0004-6361:20034617

Dumoulin, V., and Visin, F. (2016). *A Guide to Convolution Arithmetic for Deep Learning*. arXiv:1603.07285 [Online].

Ellwarth, M., Fischer, C. E., Vitas, N., Schmiz, S., and Schmidt, W. (2021). Newly Formed Downflow Lanes in Exploding Granules in the Solar Photosphere. *Astron. Astrophys.* 653, A96. doi:10.1051/0004-6361/202038252

Fernández, A., García, S., Galar, M., Prati, R., Krawczyk, B., and Herrera, F. (2018). "Foundations on Imbalanced Classification," in *Learning from Imbalanced Data Sets* (Cham: Springer). doi:10.1007/978-3-319-98074-4_2

Fischer, C. E., Borrero, J. M., Bello González, N., and Kaithakkal, A. J. (2019). Observations of Solar Small-Scale Magnetic Flux-Sheet Emergence. *Astron. Astrophys.* 622, L12. doi:10.1051/0004-6361/201834628

Fischer, C. E., Vigeesh, G., Lindner, P., Borrero, J. M., Calvo, F., and Steiner, O. (2020). Interaction of Magnetic Fields with a Vortex Tube at Solar Subgranular Scale. *Astrophys. J. Lett.* 903, L10. doi:10.3847/2041-8213/abbada

Ghosh, A., Sufian, A., Sultana, F., Chakrabarti, A., and De, D. (2020). "Fundamental Concepts of Convolutional Neural Network," in *Recent Trends and Advances in Artificial Intelligence and Internet of Things, Intelligent Systems Reference Library*. Editors V. Balas, R. Kumar, and R. Srivastava (Springer), Vol. 172. doi:10.1007/978-3-030-32644-9_36

Girshick, R. B. (2015). "Fast R-Cnn," in 2015 IEEE International Conference on Computer Vision (ICCV) Santiago, Chile, December 7–13, 2015, 1440–1448. doi:10.1109/iccv.2015.169

Guglielmino, S. L., Martínez Pillet, V., Ruiz Cobo, B., Bellot Rubio, L. R., del Toro Iniesta, J. C., Solanki, S. K., et al. (2020). On the Magnetic Nature of an Exploding Granule as Revealed by Sunrise/IMaX. *Astrophys. J.* 896, 62. doi:10.3847/1538-4357/ab917b

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., et al. (2017). Brain Tumor Segmentation with Deep Neural Networks. *Med. Image Anal.* 35, 18–31. doi:10.1016/j.media.2016.05.004

He, H., and Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* 21, 1263–1284. doi:10.1109/TKDE.2008.239

He, H., and Ma, Y. (2013). *Imbalanced Learning: Foundations, Algorithms, and Applications*. 1st edn. Hoboken, New Jersey: Wiley-IEEE Press.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). "Adasyn: Adaptive Synthetic Sampling Approach for Imbalanced Learning," in 2008 IEEE International Joint Conference on Neural Networks Hong Kong, China, June 1–8, 2008 (IEEE World Congress on Computational Intelligence), 1322–1328. doi:10.1109/IJCNN.2008.4633969

Hirzberger, J., Bonet, J. A., Vázquez, M., and Hanslmeier, A. (1999). Time Series of Solar Granulation Images. III. Dynamics of Exploding Granules and Related Phenomena. *Astrophys. J.* 527, 405–414. doi:10.1086/308065

Huang, C., Li, Y., Loy, C. C., and Tang, X. (2016). "Learning Deep Representation for Imbalanced Classification," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, June 27–30, 2016. doi:10.1109/cvpr.2016.580

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). "Densely Connected Convolutional Networks," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, July 21–26, 2017, 2261–2269. doi:10.1109/CVPR.2017.243

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., and Brox, T. (2017). "Flownet 2.0: Evolution of Optical Flow Estimation with Deep Networks," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, July 21–26, 2017. doi:10.1109/cvpr.2017.179

Jaccard, P. (1912). The Distribution of the Flora in the Alpine zone.1. *New Phytol.* 11, 37–50. doi:10.1111/j.1469-8137.1912.tb05611.x

Javaherian, M., Safari, H., Amiri, A., and Ziaei, S. (2014). Automatic Method for Identifying Photospheric Bright Points and Granules Observed by Sunrise. *Sol. Phys.* 289, 3969–3983. doi:10.1007/s11207-014-0555-1

Kaithakkal, A. J., and Solanki, S. K. (2019). Cancelation of Small-Scale Magnetic Features. *Astron. Astrophys.* 622, A200. doi:10.1051/0004-6361/201833770

Khan, S. H., Hayat, M., Bennamoun, M., Sohel, F., and Togneri, R. (2017). "Cost Sensitive Learning of Deep Feature Representations from Imbalanced Data," in *IEEE Transactions on Neural Networks and Learning Systems* vol. 29, 3573–3587. doi:10.1109/TNNLS.2017.2732482

Kingma, D. P., and Ba, J. (2014). "Adam: A Method for Stochastic Optimization," in *3rd International Conference on Learning Representations, ICLR 2015* San Diego, CA, May 7–9, 2015.

Kitai, R., and Kawaguchi, I. (1979). Morphological Study of the Solar Granulation. *Sol. Phys.* 64, 3–12. doi:10.1007/BF00151111

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet Classification with Deep Convolutional Neural Networks," in Proceedings of the 25th International Conference on Neural Information Processing Systems (Red Hook,NY, USA: Curran Associates Inc.), 1097–1105. NIPS'12.Vol. 1.

Le Cun, Y., Bottou, L., and Bengio, Y. (1997). "Reading Checks with Multilayer Graph Transformer Networks," in 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, Munich, Germany, April 21–24, 1997, 151–154. doi:10.1109/ICASSP.1997.599580

Leivas Oliveira, G. (2019). *Encoder-decoder Methods for Semantic Segmentation: Efficiency and Robustness Aspects*. Freiburg im Breisgau, Germany: Technische Fakultät, Albert-Ludwigs-Universität Freiburg. Ph.D. thesis. doi:10.6094/UNIFR/150065

Lin, T. Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). "Focal Loss for Dense Object Detection," in 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, October 22–29, 2017, 2999–3007. doi:10.1109/ICCV.2017.324

Love, T., Neukirch, T., and Parnell, C. E. (2020). Analyzing Aia Flare Observations Using Convolutional Neural Networks. *Front. Astron. Space Sci.* 7. doi:10.3389/fspas.2020.00034

Malherbe, J.-M., Roudier, T., Stein, R., and Frank, Z. (2018). Dynamics of Trees of Fragmenting Granules in the Quiet Sun: Hinode/SOT Observations Compared to Numerical Simulation. *Sol. Phys.* 293, 4. doi:10.1007/s11207-017-1225-x

Martínez Pillet, V., Del Toro Iniesta, J. C., Álvarez-Herrero, A., Domingo, V., Bonet, J. A., González Fernández, L., et al. (2011). The Imaging Magnetograph eXperiment (IMaX) for the Sunrise Balloon-Borne Solar Observatory. *Sol. Phys.* 268, 57–102. doi:10.1007/s11207-010-9644-y

Muller, R., and Roudier, T. (1984). Variability of the Quiet Photospheric Network. *Sol. Phys.* 94, 33–47. doi:10.1007/BF00154805

Namba, O. (1986). Evolution of 'exploding Granules. *Astron. Astrophys.* 161, 31–38.

Nordlund, Å., Stein, R. F., and Asplund, M. (2009). Solar Surface Convection. *Living Rev. Sol. Phys.* 6, 2. doi:10.12942/lrsp-2009-2

Oksuz, K., Cam, B. C., Kalkan, S., and Akbas, E. (2020). "Imbalance Problems in Object Detection: A Review," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 43, 3388–3415. doi:10.1109/TPAMI.2020.2981890

Olã Bressan, P., Marcato Junior, J., Correa Martins, J. A., Nunes Gonçalves, D., Matte Freitas, D., Prado Osco, L., et al. (2022). Semantic Segmentation With Labeling Uncertainty and Class Imbalance Applied to Vegetation Mapping. *Internat. J. Appl. Earth Obs. Geoinf.* 108, 102690. doi:10.1016/j.jag.2022.102690

Palacios, J., Blanco Rodríguez, J., Vargas Domínguez, S., Domingo, V., Martínez Pillet, V., Bonet, J. A., et al. (2012). Magnetic Field Emergence in Mesogranular-Sized Exploding Granules Observed with sunrise/IMaX Data. *Astron. Astrophys.* 537, A21. doi:10.1051/0004-6361/201117936

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "Pytorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems*. Editors H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc.), 32, 8024–8035.

Rempel, M. (2018). Small-scale Dynamo Simulations: Magnetic Field Amplification in Exploding Granules and the Role of Deep and Shallow Recirculation. *Astrophys. J.* 859, 161. doi:10.3847/1538-4357/aabba0

Riethmüller, T. L., Solanki, S. K., Zakharov, V., and Gandorfer, A. (2008). Brightness, Distribution, and Evolution of Sunspot Umbral Dots. *Astron. Astrophys.* 492, 233–243. doi:10.1051/0004-6361:200810701

Rimmele, T. R., Warner, M., Keil, S. L., Goode, P. R., Knölker, M., Kuhn, J. R., et al. (2020). The Daniel K. Inouye Solar Telescope - Observatory Overview. *Sol. Phys.* 295, 172. doi:10.1007/s11207-020-01736-7

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015 Lecture Notes in Computer Science. Editors N. Navab, J. Hornegger, W. Wells, and A. Frangi Cham: Springer. doi:10.1007/978-3-319-24574-4_28

Roudier, T., and Muller, R. (2004). Relation between Families of Granules, Mesogranules and Photospheric Network. *Astron. Astrophys.* 419, 757–762. doi:10.1051/0004-6361:20035739

Roudier, T., Lignières, F., Rieutord, M., Brandt, P. N., and Malherbe, J. M. (2003). Families of Fragmenting Granules and Their Relation to Meso- and Supergranular Flow Fields. *Astron. Astrophys.* 409, 299–308. doi:10.1051/0004-6361:20030988

Roudier, T., Malherbe, J. M., Rieutord, M., and Frank, Z. (2016). Relation between Trees of Fragmenting Granules and Supergranulation Evolution. *Astron. Astrophys.* 590, A121. doi:10.1051/0004-6361/201628111

Schmidhuber, J. (2014). Deep Learning in Neural Networks: An Overview. *Neural Networks* 61, 85–117. doi:10.1016/j.neunet.2014.09.003

Sebe, N., Cohen, I., Garg, A., and Huang, T. (2005). *Machine Learning in Computer Vision*. New York: Springer.

Shelhamer, E., Long, J., and Darrell, T. (2017). Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 640–651. doi:10.1109/tpami.2016.2572683

Shorten, C., and Khoshgoftaar, T. M. (2019). A Survey on Image Data Augmentation for Deep Learning. *J. Big Data* 6, 60. doi:10.1186/s40537-019-0197-0

Simonyan, K., and Zisserman, A. (2015). "Very Deep Convolutional Networks for Large-Scale Image Recognition," in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, May 7–9, 2015 .

Singh Punn, N., and Agarwal, S.(2021). *BT-Unet: A Self-Supervised Learning Framework for Biomedical Image Segmentation Using Barlow Twins with U-Net Models*. arXiv e-prints. arXiv:2112.03916.

Solanki, S. K., Barthol, P., Danilovic, S., Feller, A., Gandorfer, A., Hirzberger, J., et al. (2010). Sunrise: Instrument, Mission, Data, and First Results. *Geophysical Monograph Series* 723, L127–L133. doi:10.1088/2041-8205/723/2/l127

Solanki, S. K., Riethmüller, T. L., Barthol, P., Danilovic, S., Deutsch, W., Doerr, H. P., et al. (2017). The Second Flight of the Sunrise Balloon-Borne Solar Observatory: Overview of Instrument Updates, the Flight, the Data, and First Results. *Astrophys. J. Suppl. Ser.* 229, 2. doi:10.3847/1538-4365/229/1/2

Steiner, O., Franz, M., Bello González, N., Nutto, C., Rezaei, R., Martínez Pillet, V., et al. (2010). Detection of Vortex Tubes in Solar Granulation from Observations with SUNRISE. *Astrophys. J. Lett.* 723, L180–L184. doi:10.1088/2041-8205/723/2/L180

Stix, M. (2002). *The Sun: An Introduction*. Germany: Springer.

Szeliski, R. (2011). *Computer Vision: Algorithms and Applications*. New York: Springer.

Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A Survey of Transfer Learning. *J. Big Data* 3, 9. doi:10.1186/s40537-016-0043-6

Xian, Y., Lampert, C. H., Schiele, B., and Akata, Z. (2017). "Zero-Shot Learning – A Comprehensive Evaluation of the Good, the Bad and the Ugly," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41, 2251–2265. doi:10.1109/TPAMI.2018.2857768

Yanli, S., and Pengpeng, S. (2021). J-net: Asymmetric Encoder-Decoder for Medical Semantic Segmentation. *Secur. Commun. Netw.* 2021, 2139024. doi:10.1155/2021/2139024

Zou, Y., Weinacker, H., and Koch, B. (2021). Towards Urban Scene Semantic Segmentation with Deep Learning from Lidar Point Clouds: A Case Study in Baden-Württemberg, germany. *Remote Sens.* 13. doi:10.3390/rs13163220

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Methods
PUBLISHED 12 August 2022
DOI 10.3389/fspas.2022.897301

OPEN ACCESS

EDITED BY
Bala Poduval,
University of New Hampshire,
United States

REVIEWED BY
Juan Carlos Martínez Oliveros,
University of California, Berkeley,
United States
Naoto Nishizuka,
National Institute of Information and
Communications Technology, Japan

*CORRESPONDENCE
Chetraj Pandey,
cpandey1@gsu.edu

SPECIALTY SECTION
This article was submitted to Space Physics,
a section of the journal
Frontiers in Astronomy and Space Sciences

RECEIVED 16 March 2022
ACCEPTED 30 June 2022
PUBLISHED 12 August 2022

CITATION
Pandey C, Ji A, Angryk RA, Georgoulis MK and Aydin B (2022), Towards coupling full-disk and active region-based flare prediction for operational space weather forecasting. *Front. Astron. Space Sci.* 9:897301. doi: 10.3389/fspas.2022.897301

Towards coupling full-disk and active region-based flare prediction for operational space weather forecasting

Chetraj Pandey[1]*, Anli Ji[1], Rafal A. Angryk[1], Manolis K. Georgoulis[2] and Berkay Aydin[1]

[1]Department of Computer Science, Georgia State University, Atlanta, GA, United States, [2]Research Center for Astronomy and Applied Mathematics, Academy of Athens, Athens, Greece

Solar flare prediction is a central problem in space weather forecasting and has captivated the attention of a wide spectrum of researchers due to recent advances in both remote sensing as well as machine learning and deep learning approaches. The experimental findings based on both machine and deep learning models reveal significant performance improvements for task specific datasets. Along with building models, the practice of deploying such models to production environments under operational settings is a more complex and often time-consuming process which is often not addressed directly in research settings. We present a set of new heuristic approaches to train and deploy an operational solar flare prediction system for ≥M1.0-class flares with two prediction modes: full-disk and active region-based. In full-disk mode, predictions are performed on full-disk line-of-sight magnetograms using deep learning models whereas in active region-based models, predictions are issued for each active region individually using multivariate time series data instances. The outputs from individual active region forecasts and full-disk predictors are combined to a final full-disk prediction result with a meta-model. We utilized an equal weighted average ensemble of two base learners' flare probabilities as our baseline meta learner and improved the capabilities of our two base learners by training a logistic regression model. The major findings of this study are: 1) We successfully coupled two heterogeneous flare prediction models trained with different datasets and model architecture to predict a full-disk flare probability for next 24 h, 2) Our proposed ensembling model, i.e., logistic regression, improves on the predictive performance of two base learners and the baseline meta learner measured in terms of two widely used metrics True Skill Statistic (TSS) and Heidke Skill Score (HSS), and 3) Our result analysis suggests that the logistic regression-based ensemble (Meta-FP) improves on the full-disk model (base learner) by ~9% in terms TSS and ~10% in terms of HSS. Similarly, it improves on the AR-based model (base learner) by ~17% and ~20% in terms of TSS and HSS respectively. Finally, when compared to the baseline meta model, it improves on TSS by ~10% and HSS by ~15%.

KEYWORDS

solar flares, solar magnetograms, ensemble, machine learning, deep learning

1 Introduction

A solar flare is an intense burst of electromagnetic radiation through magnetic reconnection and plasma instability coming from the release of magnetic energy associated with active regions (AR) and they transpire as a sudden brightening of light on the Sun's corona Toriumi and Wang (2019). Coronal mass ejections (CMEs), which are often associated with solar flares, have comparable energies, and can release large amounts of mass resulting into major geomagnetic storms which creates intense currents in the Earth's magnetosphere, changes in the radiation belts, and in the ionosphere Feng et al. (2020). When particles emitted by the Sun are accelerated during a flare or by a CME event and reach the Earth along interplanetary magnetic field lines, Solar energetic particle (SEP) events are produced Núñez and Paul-Pena (2020). Primarily, solar flares are considered to be the central phenomena in space weather forecasting, and this paper discusses on the predictive models for solar flares. Solar flares can induce intense variation in Earth's magnetic field, causing potential disruptions to many stakeholders such as the electricity supply chain, airlines industry, astronauts in space, and communication systems including satellites and radio. Forecasting solar flares has been a major challenge in heliophysics owing to the yet unsolved fundamental cause of this phenomenon which makes it difficult to predict the exact occurrence of a flare, especially for relatively large ones. However, recent advancements in machine learning and deep learning methods have demonstrated great experimental success and catalyzed the efforts in prediction of solar flares, which captivated the interest of many interdisciplinary researchers Li et al. (2020); Nishizuka et al. (2018); Huang et al. (2018). Developing predictive models for flare prediction is limited to the nature, quantity, and quality of flaring instances as well as the inductive bias of learning algorithms when predicting such flare events. As a consequence of the intrinsic limitations pre-incorporated by the predictive models during problem formulation or model selection or utilizing different data products, an individual flare prediction model is limited in performance. Although all the models built so far for flare forecasting have limitations, different comprehensions and insights on data distribution are still valuable for making the final decision in an operational flare forecasting system. Therefore, it is intuitive to use as many pieces of information that can be gathered from different sets of models such as machine learning or deep learning models obtained from different data modalities in terms of active region magnetogram patches, full-disk magnetograms or magnetogram's metadata (magnetic field parameters) to issue a reduced risk prediction.

In active region-based models, predictions are issued for certain areas on the Sun with greatly enhanced magnetic flux, known as active regions. Active regions have lifetimes of days to month, feature strong and entangled magnetic fields and are the exclusive locations of strong flares and major eruptions, including fast coronal mass ejections (CMEs). This said, only a slim minority (10% or less) of active regions appearing in a given solar cycle provide flares of GOES class ≥M1.0 and fast CMEs [e.g., Georgoulis et al. (2019); Toriumi and Wang (2019)]. These regions can host solar eruptions. To employ active region-based models in an operational setting, individual active region forecasts are aggregated by computing the probability of flare from at least one active region assuming conditional independence and then these flare probabilities are used to compute a full-disk flare occurrence probability. However, for an operational system, working with near-real time data and issuing near-real time predictions, active region-based models relying on magnetic field observations possess a limited forecasting ability as they restrict the training datasets within central regions (±70°) due to severe projection effects Hoeksema et al. (2014). Besides the unreliable measurements, foreshortening closer to the solar limbs greatly impacts the operational use of magnetic field data. This leads to reduction in significant information required to make reliable flare predictions in active regions. Moreover, predictions from active region-based models often rely on sampled subset of statistical features that were used to train the model and therefore when examining forecasts from different subsets of features, it is common to observe that for the similar condition of the photospheric magnetic field, they can give varying values for prediction probabilities of a particular flare to happen.

To account for the limitations of active region-based flare predictors, full-disk prediction models provide a complementary approach for operational flare forecasting systems Pandey et al. (2021). The full-disk model utilize the compressed line-of-sight magnetograms and these magnetograms are used for shape based parameters (such as size, directionality, borders of sunspots) and do not possess the magnetic field properties as in the magnetogram rasters which is advantageous over the active region-based models where individual active region magnetic field parameters used near the limb are more prone to projection effects. The significant part of an operational flare forecasting model is to issue a reliable forecast for which we use a heterogeneous ensemble that combines two different base learners. In addition, to address the operational aspect of our system, we consider two essential system-level criteria: 1) near-real-time availability of input data is ensured given that both of our base learners are trained with line-of-sight magnetograms and physical parameters obtained from a line-of-sight magnetograms and vector magnetograms available at a cadence of 12 min, and 2) our proposed system is scalable in a sense that it allows the flexibility of adding a new base learner (if needed in the future) in the system as it will be one step away from retraining the ensemble and deploying it back to our forecasting system.

In this work, to issue more reliable forecasts in an operational settings, we propose a heuristic ensemble approach which consolidates the predictive results of the two aforementioned

prediction modalities into one combined solar flare forecast. The major contributions of this paper are following: we present a methodology on how to train and validate an ensemble flare prediction model in regard to its operations-ready characteristics. The ensemble combines the predictions from two base learners: 1) a deep learning-based full-disk flare predictor using SDO/HMI images and 2) a set of probabilistic predictions from a time series classifier utilizing active region patches' magnetic field metadata in the form of multivariate time series. For both base learners, we use the similar time-segmented tri-monthly data partitioning strategy Pandey et al. (2021) to perform 3-fold cross-validation experiments. Finally, we use the probability scores of these two base learners obtained from the validation and test partitions to train and validate our proposed meta-learner which converges to a more robust full-disk flare predictor.

The remainder of this paper is organized as follows. In Section 2, we present the related work on ensemble solar flare forecasting models. In Section 3, we provide a detailed workflow of our methodology. In Section 4, we present our detailed experimental evaluation with settings and results. In Section 5 we present a discussion on the ensembles created and, lastly, in Section 6, we present our conclusions and discuss future work.

2 Related works

The idea of automatically extracting forecast patterns from the large volume of intrinsic magnetic field data on the photosphere of the sun using machine learning methods has begun from the early 1990's Aso et al. (1994). Since then, with the rapid development in machine learning and deep learning approaches, a number of research groups Nishizuka et al. (2018); Huang et al. (2018); Li et al. (2020), Nishizuka et al. (2021), and references therein present their efforts in applying such methods to build flare forecasting models.

In recent years, Li et al. (2020); Huang et al. (2018) used a deep learning model based on CNN with different data products for flare forecasting. Although they show an impressive performance on flare classification, they limit the scope of the prediction to smaller areas by using active region-based data within ± 30°–45° of the central meridian of the Sun which may counter their performance for true operational forecasting. In addition, Florios et al. (2018) calculated physical features of flaring and non-flaring ARs obtained from the SDO/HMI's near-real-time vector magnetogram data and trained SVMs, multilayer perceptrons (MLPs), and decision tree algorithms to predict occurrences of ≥M1.0-class and ≥C1.0-class flares with a forecast horizon of 24 h. In Benvenuto et al. (2018), a combination of supervised lasso regression for identifying the significant features and then an unsupervised fuzzy clustering is used for the classification of ≥M1.0-class and ≥C1.0-class flares. Furthermore, Park et al. (2018); Pandey et al. (2021) uses full-disk magnetograms data as a point in time observation with CNN based deep learning models, which have limitations in capturing the evolution of solar flares and they do not account for flares that are on the eastern-limb of the Sun. Overall, some methods are appropriate for constructing prediction models for the temporal data variation, whereas others are beneficial for spatial data variation, which demands a need for a coupled hybrid model that can exploit the gains of multiple models.

Jonas et al. (2018) designed a time series data set using photospheric and coronal images from HMI/SDO and AIA/SDO instruments to forecast ≥M1.0-class flares within the next 24 h. They utilize random partitioning of datasets into 80 and 20% for training and testing the linear classifier. Apart from devising flare forecasting as a binary classification task, Abduallah et al. (2021) formulates it as a multiclass classification problem to classify B-, C-, M- and X-class flares by utilizing the physical parameters within ± 70° provided by the SHARP series of HMI/SDO. Finally, the author uses majority voting as an ensemble to issue a final flare forecast from three different models trained on the same data. The training procedure in their work uses random 10-fold cross-validation.

Instead of using a single prediction model, ensembles use a set of predictions and combine these results to improve on a single-model prediction. In addition, an ensemble can be created with a single model itself by perturbing its initial conditions or parameter settings to produce multiple results and then combine those results into one called homogeneous ensembles Breiman (1996); Freund and Schapire (1996). Flare forecasting problems also make use of decision tree-based homogeneous ensembles. Liu C. et al. (2017) apply random forest (RF) Breiman (2001)—a meta-algorithm that fits a number of decision tree classifiers on different sub-samples of a dataset and utilizes averaging to improve the model's performance. Similarly, Nishizuka et al. (2017) employed an extremely randomized tree (ERT) classifier Geurts et al. (2006) by fitting several decision-tree classifiers on a random subset of features with a randomly defined threshold to prevent overfitting. While RF and ERT are meta-algorithms based on the bagging technique, XGBoost Chen and Guestrin (2016) follows boosting approach to ensemble construction and focuses on incorrect predictions. It varies from Random Forest such that XGBoost always prioritizes functional space while reducing the cost of a model, whereas Random Forest tries to prioritize hyperparameters when optimizing the model. McGuire et al. (2019) uses XGBoost for window-based feature extraction from time series of physical parameters to classify solar flares. However the aforementioned ensembles can optimize on one set of data modality.

Besides decision trees, different models trained with different algorithms but with same data modalities can also be used in an ensemble as in Liu J.-F. et al. (2017). However, they only included magnetograms with ARs within ± 30° of the central meridian of the Sun for ≥C1.0-class flares and then designed a multimodel integrated learner (MIM) by fitting several distinct base learners, such as neural networks, naive classifiers, and SVMs. Finally, the

outputs of base learners were combined by a genetic algorithm. Similar efforts for ≥C1.0-class flares forecasting can be seen in Campi et al. (2019) where ARs extracted from SDO/HMI images from 2012 September 14 and 2016 April 30 are used and two-third of the instances are randomly selected for training and one-third for testing their models. Furthermore, in Domijan et al. (2019) they study the predictive capabilities of magnetic-feature properties located within ± 45° from the solar central meridian and detected using Solar Monitor Active Region Tracker Higgins et al. (2011) in Michelson Doppler Imager (MDI) magnetograms and analyze the features to predict ≥C1.0-class flares within the 24 h following the observation. In this data-driven era of predictive models, complex models can bring on higher accuracy, but also ensembles allow many weak models to be combined to produce a meta model that can compete with the state-of-the-art research efforts Murray (2018).

In recent years, the usage of ensembles have become a more popular research topic in space weather forecasting. Guerra et al. (2015) created a multi-model ensemble from four base learners for ≥M1.0-class flare prediction, finding an improved forecast output compared to any one single model. Similarly, Schunk et al. (2016) built an ionosphere-thermosphere-electrodynamics multimodel ensemble prediction system based on seven physics-based data assimilation models. Furthermore, in Guerra et al. (2020), full-disk probabilistic forecasts from six operational forecasting methods are converted to an ensemble for ≥M1.0-class flares by a linear classifier and create a total of 28 ensembles to show the improvement of such a technique over individual model forecasts. Although, ensemble methods are increasingly being used by space weather researchers, much of this research has yet to be implemented into operations, where transitioning comes with issues of model compatibility.

It is worth noting that using a flare forecasting model in operational settings, generally it is preferred to use more simplistic robust methods. Diving into meteorology's scenario, The NASA Community Coordinated Modeling Center's (CCMC) CME Scoreboard [1]) and solar flare Scoreboard [2]) provide an weighted and equi-weight average of multiple forecast scores. Using an equal weighted average of multiple forecasts can be used as a reliable first guess over a more complex model runs or deciding on one specific forecast out of several in operations Murray (2018), however, an ensemble derived from a linear combination of multiple models can add to the decision making capabilities on one final forecast leveraging the advantage of simplicity and hence making it more reliable to trust its decision while in operation.

To evaluate a flare forecasting system in an operational scenario, Cinto et al. (2020) provides a set of criteria that are worth considering and can be used to distinguish a non-operationally evaluated system: 1) model evaluation without truly unseen data, 2) using active region (AR) magnetograms only near the center of the solar disk, 3) only using AR magnetograms linked to ≥C1.0-class flares, and 4) using insufficient data instances. The author argues that the non-operationally evaluated system are evaluated under certain bias and that does not make them wrong, however, evaluating under such specific conditions might impair their predictive capabilities in real operational settings. In addition to these guidelines, it is essential to note that, most of the studies, create a cross-validation dataset by randomizing the process of data splitting. While such data splitting leads to higher experimental accuracy scores, it often fails to deliver similarly real-time performance as discussed in Ahmadzadeh et al. (2021). We build our models that meet the standard of the aforementioned criteria as they can address the near-limb events with the full-disk base learner, they are trained and tested with a time-segmented partitioning of data from solar cycle 24, and we evaluate our models using data instances that were not presented to the models during training to address the operational settings of flare forecasting.

In this work, we combine the prediction probabilities of two types of base learners by the means of a linear classifier based on logistic regression. Our first base learner, which is a deep learning based model which focuses on spatial variation of a full-disk magnetogram. Similarly, our second base learner is a heuristic-based aggregation model which outputs full disk probability using the results from active region-based multivariate time series classifiers. We train and validate an operations-ready ensemble flare prediction model which optimizes the predictive performance of both our base learners and provides a better confidence while issuing a flare forecast.

3 Methodology

Ensemble approaches integrate multiple forecasts into a single prediction by combining the predictions from multiple base learners. A simplistic way of integrating the forecasts is to use an equal weighting for each forecast and combine to improve on a single-model prediction which we use as our baseline meta-model. As mentioned earlier, we attempt to combine the predictions of two base learners: 1) a deep learning-based full-disk flare predictor using Helioseismic and Magnetic Imager (HMI) instrument onboard Solar Dynamics Observatory (SDO) images and 2) a multivariate time series classifier utilizing magnetic field metadata to issue one combined full-disk flare forecast.

3.1 Base learners

3.1.1 Time-series forest

Our active region-based prediction model is a multivariate Time Series Forest (TSF), trained with Space Weather Analytics

1 https://kauai.ccmc.gsfc.nasa.gov/CMEscoreboard/.

2 https://ccmc.gsfc.nasa.gov/challenges/flare.php.

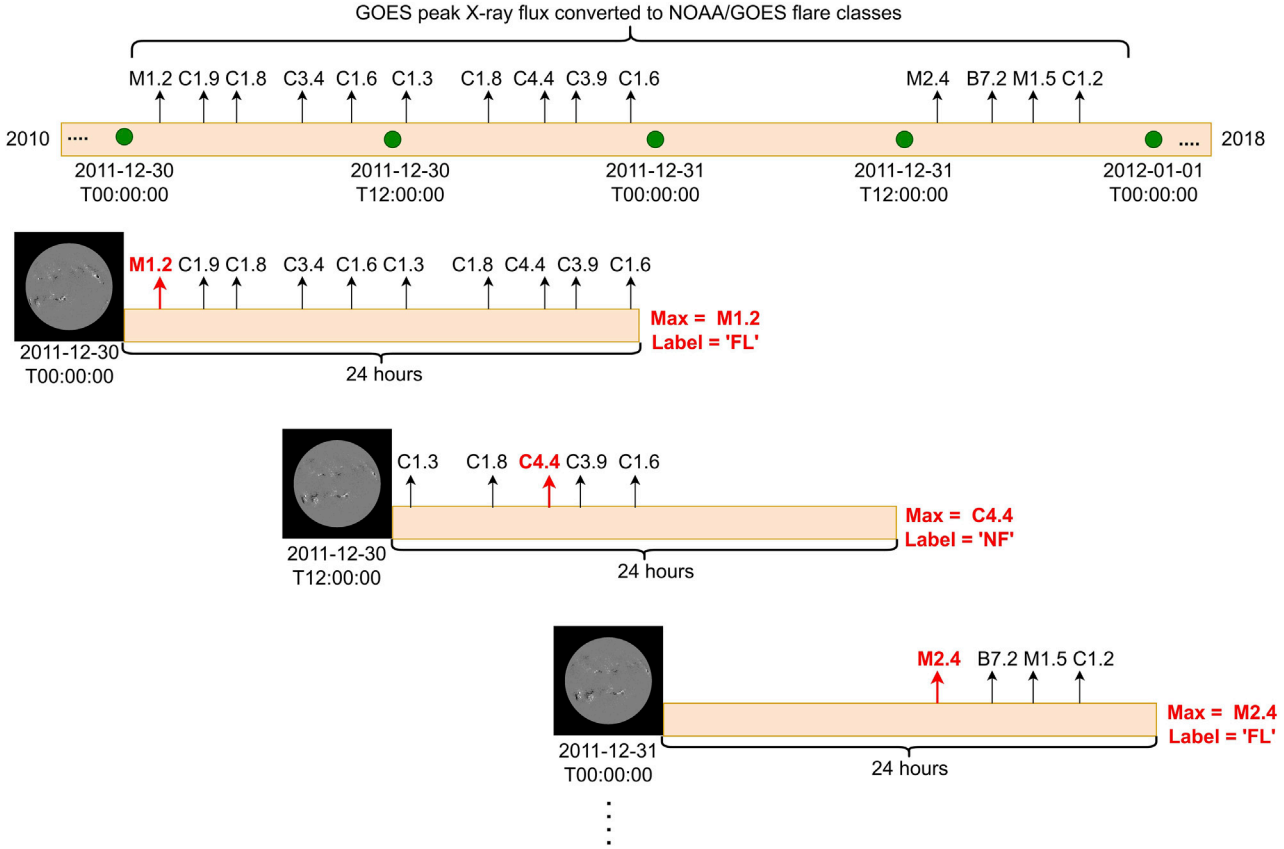

FIGURE 1
A timeline diagram to present the problem formulation of our deep learning-based full-disk flare prediction model using bi-daily observations of full-disk line-of-sight magnetograms and prediction window of 24 h considered to label the magnetogram instances.

benchmark dataset for solar flare prediction (SWAN-SF) Angryk et al. (2020a,b) to predict the occurrence of ≥M1.0-class flares within the next 24 h by using an observation window of 12 h. The SWAN-SF is an open source multivariate time series (MVTS) dataset that provides time series instances for a collection of space weather related physical parameters within ± 70° primarily calculated for each active regions from solar photospheric magnetograms. The TSF model is trained by utilizing six magnetic-field parameters: 1) TOTUSJH (Total unsigned current helicity), 2) TOTPOT (Total photospheric magnetic free energy density), 3) TOTUSJZ (Total unsigned vertical current), 4) ABSNJZH (Absolute value of the net current helicity), 5) SAVNCPP (Sum of the modulus of the net current per polarity), and 6) USFLUX (Total unsigned flux) from the suggested list of 13 parameters in Bobra and Couvidat (2015) as these are available in near-real time, which is a necessity for an operational system. The model outputs the flaring probability for an individual active region and the implementation of this model is based on Ji et al. (2020).

3.1.2 Deep learning model

We trained an AlexNet-based Krizhevsky et al. (2012) Convolutional Neural Network to perform full-disk binary flare prediction for ≥M1.0-class flares. Similar to the active region-based counterparts, the full-disk model assumes a 24 h prediction window, but uses a single image (point-in-time observation) to perform the predictions. For this task, we collected compressed 8-bit images created from full-disk line-of-sight magnetograms provided by HMI/SDO. We collected two compressed magnetogram images per day (bi-daily image samples) at 00:00 UT and 12:00 UT from December 2010 to December 2018 using Helioviewer API Muller et al. (2009) and labeled them based on maximum of GOES peak X-ray flux converted to NOAA/GOES flare classes observed in next 24 h as shown in Figure 1. Unlike the TSF model, this deep learning model outputs flaring probability for the entire full-disk and its implementation is based on Pandey et al. (2021).

We used trimonthly partitioning for training our models, which is non-chronological time-segmented partitioning

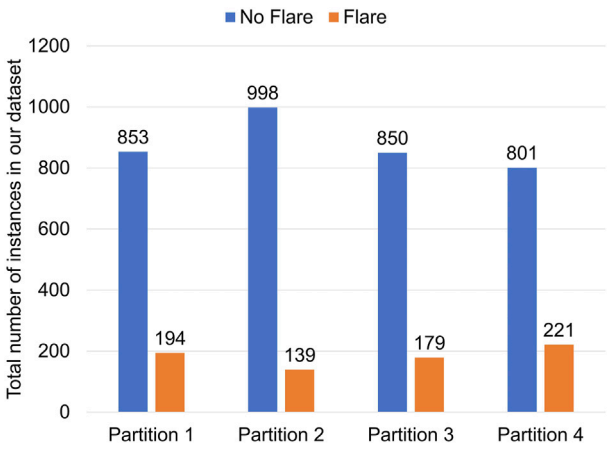

FIGURE 2
Total number of Flare and No Flare instances across 4 trimonthly partitions used in this work[3].

strategy, where Partition-1 contains data from January to March, Partition-2 from April to June, Partition-3 from July to September, and Partition-4 from October to December in a timeline from 2010 to 2018. The AR-based model also uses the same partitioning for aligning our training partitions and avoiding the penetration of training partitions into testing data in different prediction modalities to ensure the fair comparisons and avoid partial memorization through temporal coherence Ahmadzadeh et al. (2021).

3.2 Flare prediction ensemble

Our active region-based model outputs probabilities of flare (P_{FL}) for each active region which we then aggregate to obtain a restricted full-disk flaring probability (i.e., from active regions in central locations). We use the following heuristic function in Eq. 1 to determine aggregated active region flaring probability[3].

$$P_{aggregated} = 1 - \prod_i [1 - P_{FL}(AR_i)] \quad (1)$$

where P_{FL} (AR_i) is the flaring probability of an active region, and the aggregated result calculates the probability of having at least one flaring active region, assuming the flaring events from active regions are conditionally independent. The product term calculates the probability of having no flaring active regions. These aggregated results from the active-region based model are then concatenated with full-disk model's output. The aggregation procedure searches for most-recent valid active-region predictions up to 6 h prior to the designated forecast issue time. These gathered predictions from full-disk and aggregated full-disk probabilities are then combined to issue a final flare forecast using an ensemble. In this work, while preparing our final dataset for the full-disk model, we do not include magnetogram images where the observation time of the available image and requested image timestamp is more than 6 hours. Therefore due to data unavailability through helioviewer, we have used a total of 4,235 data instances, where 3,502 are No Flare (NF) instances and 733 are Flare (FL) instances. The detailed distribution of the dataset for each tri-monthly partition is shown in Figure 2 and the class imbalance ratios across the partitions are generally consistent from ~ 12–22% (~3.6:1 to ~7.2:1).

In our baseline meta-model approach, we use equal weighted averaging of flare probabilities from aggregated active-regions and full-disk flaring probabilities for issuing a final forecast. In other words, given two flaring probabilities from two approaches, the baseline approach is to compute the arithmetic average of the probabilities, assuming equal importance. This simplistic combination of flare probabilities will serve as our baseline, although it is a naive approach that does not consider the intrinsic characteristics of long-term diagnostic results from the models.

Our alternative approach to the baseline meta-model is logistic regression-based classifier that is trained with flaring probabilities from the base learners. As we already use two powerful algorithms to train our base learner to extract the complex dynamics of the datasets, we chose a linear model, logistic regression, because of its simplicity and computational efficiency for the final prediction result. The infrastructure of our complete flare prediction system design is presented in Figure 3 which shows our overall methodology for creating an ensemble using two heterogeneous base learners that outputs a full-disk flare forecast.

Given the flare probability scores of two base learners which we utilize as two input features—P_{FL} (FD) and P_{FL} (Aggregated), and one binary (0/1) target feature (y) where 0 is used for No flare (NF) and 1 is used for Flare (FL). Logistic Regression aim to optimize the weights (w_1, w_2, and b), such that:

$$Z = w_1 \times P_{FL}(FD) + w_2 \times P_{FL}(Aggregated) + b \quad (2)$$

$$\hat{y} = \sigma(Z) \quad (3)$$

where, Z in Eq. (2) is the linear combination of two base learners' output, σ is the sigmoid activation function, and $\hat{y}$ is the predicted output as shown in Eq. (3). The above problem of finding the optimized weights w_1, w_2 for two base learners is formulated as an optimization problem where the loss is minimized to get the better values of weights using a logistic loss function as shown in Eq. (4).

$$loss(L) = -\frac{1}{N}\sum_{i=1}^{N}[(y_i \cdot log(\hat{y}_i)) + ((1 - y_i) \cdot log(1 - \hat{y}_i))] \quad (4)$$

3 We note that, while aggregating active regions based outputs to full-disk probabilities, there were instances that were not available even when we search for most-recent valid active-region predictions up to 6 h prior to the designated forecast. Therefore, such instances are also removed from full-disk models for consistency.

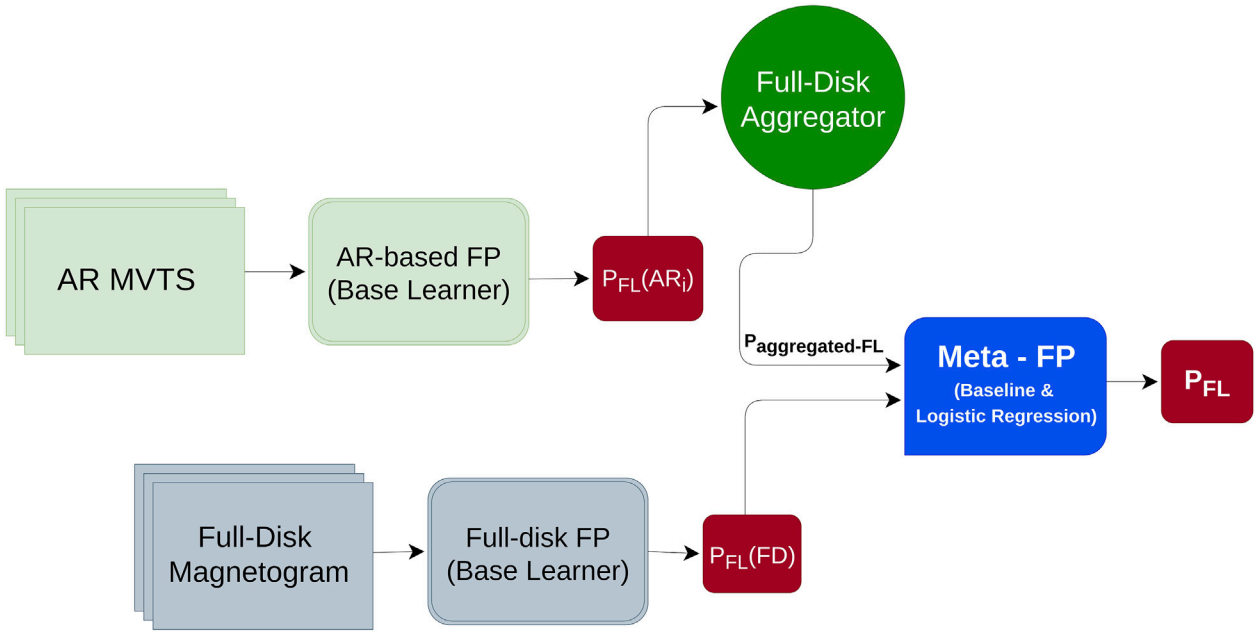

FIGURE 3
An illustration of our ensemble flare prediction pipeline showing two base learners (AR-based FP) and (Full-disk FP) and the ensemble (Meta-FP) followed by full-disk aggregation of AR-based FP's flare probabilities.

We use stochastic gradient descent (SGD) as our solver for the optimization with hyperparameter tuning. The hyperparameters we considered are learning rate and different regularization parameters which includes L1 loss Tibshirani (1996), L2 loss Hoerl and Kennard (1970), and linear mixings of L1 and L2 loss Zou and Hastie (2005). And As we will describe later on Section 4, we employ 2-fold cross-validation for our meta-model where we use one of the test partition scores of the base learners to train and another for testing our meta model, referred to as Meta-FP, interchangeably. We note that we aim to provide full-disk forecasts by computing the aggregated flare probability scores from active regions to make it compatible with the full-disk model using the probabilistic heuristic shown in Eq. (1).

4 Experimental evaluation

4.1 Experimental settings

In this work, we trained two base learners for flare prediction (≥M1.0-class flares) with two different dataset and model configurations and architectures. Although our two base learners utilize two different data modalities (i.e., point-in-time image and multivariate time series), we used time-segmented tri-monthly partitioning when training both of these models. We divided our datasets into four partitions to ready our 3-fold holdout cross-validation dataset. The data in Partition-1 contains images from the months of January to March, Partition-2 from April to June, Partition-3 from July to September, and Partition-4 from October to December. Here, this partitioning of the dataset is created by dividing the data timeline from Dec 2010 to Dec 2018 into four partitions on the basis of months rather than chronological partitioning, to incorporate approximately equal distribution of flaring instances in every fold for training, validating, and testing the model. Furthermore, such a partitioning strategy diversify the data instances in both the training and testing phase of our models as it considers instances during solar maxima and minima of solar cycle 24 used in this work.

We create three sets of base learner models from 3-fold cross-validation experiments as our base learners where we use Partition-3 as our hold-out test set (i.e., never used in training and validation). Then,

- In Fold-1, we trained both of our base learners with Partition-1 and Partition-2 and validated on Partition-4
- In Fold-2, we trained both of our base learners with Partition-1 and Partition-4 and validated on Partition-2
- In Fold-3, we trained both of our base learners with Partition-2 and Partition-4 and validated on Partition-1.

All of these three base learners are tested on Partition-3. Partition-3 as a test differs from the validation sets in each fold such that, we used the validation set in every epoch to track the performance of our model whereas the test set, Partition-3, is used only once to confirm the performance of the trained models and meta-models at the end.

To train and validate our Meta-FP, we create our dataset based on the probability scores of our three base learner sets obtained from 3-

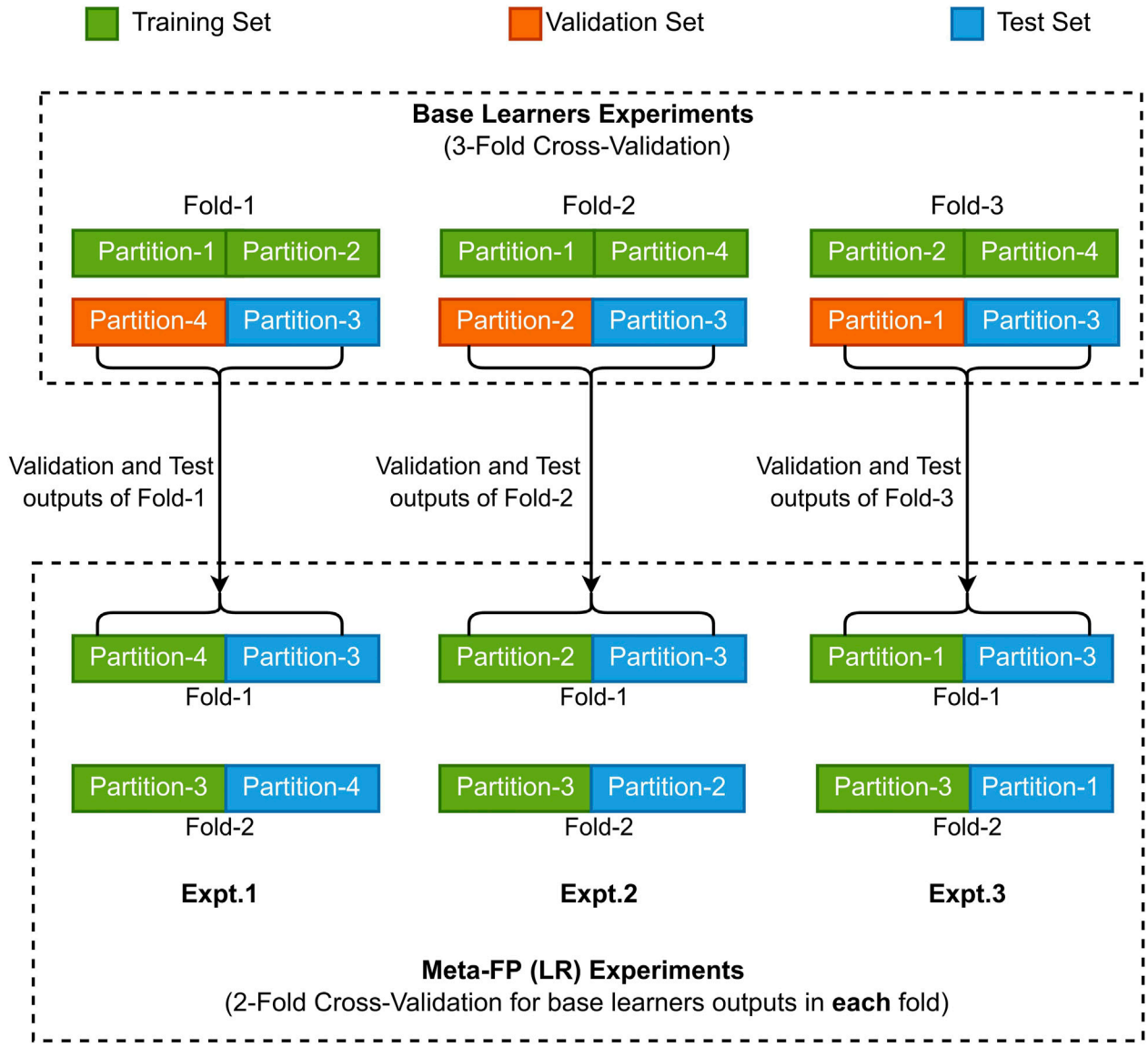

FIGURE 4
An experimental design diagram to depict the flow of our experiments for this work. Meta-FP experiments for logistic regression (LR) are cross-validated using each fold results of base learners. This results into 2-fold cross-validation in each experiments of Meta-FP.

Fold cross validation experiments. The details of our experimental design is shown in Figure 4. We used the flare probability scores from the validation set and test set used in respective base learners interchangeably to train and validate our Meta-FP model which is a general linear model i.e., logistic regression (LR). The experiments for Meta-FP are performed in such way that:

- In Expt. 1, we performed 2-fold cross validation with Partition-4 and Partition-3.
- In Expt. 2, we performed 2-fold cross validation with Partition-2 and Partition-3.
- In Expt. 3, we performed 2-fold cross validation with Partition-1 and Partition-3.

In doing so, we trained six Meta-FP models based on logistic regression and compared our results with a baseline Meta-FP which is an equal weighted average of two base learners.

To evaluate the performance of our models, we create a contingency matrix, which includes information on True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) to evaluate the performance of our base learners and Meta-FP. Note that, in the context of our flare

prediction task, Flare (FL) is considered as the positive outcome while No Flare (NF) is the negative. Using these four outcomes we use two widely used performance metrics in space weather forecasting, True Skill Statistics [TSS, shown in Eq. (5)] and Heidke Skill Score (HSS, shown in Eq. (6)) to evaluate our model.

$$TSS = \frac{TP}{TP + FN} - \frac{FP}{FP + TN} \tag{5}$$

$$HSS = 2 \times \frac{TP \times TN - FN \times FP}{((P \times (FN + TN) + (TP + FP) \times N))} \tag{6}$$

The values of TSS range from -1 to 1, where 1 indicates all correct predictions, -1 represents all incorrect predictions, and 0 represents no-skill, often transpiring as the random or one-sided (all positive/all negative) predictions. It is defined as the difference between True Positive Rate (TPR) and False Positive Rate (FPR) and does not account for class-imbalance, i.e., treats false positives (FP) and false negatives (FN) equally. Similarly, HSS measures the forecast skill of the models over an imbalance-aware random prediction. It ranges from -∞ to 1, where 1 represents the perfect skill and 0 represents no skill gain over a random prediction. It is common practice to use HSS for the solar flare prediction models (similar to weather predictions where forecast skill has more value than accuracy or single-class precision), due to the high class-imbalance ratio present in the datasets.

4.2 Evaluation

Although AR-based classifiers are better for pinpointing the source active regions for flares and giving more accurate estimations for forecasting flaring phenomena, the aggregated results drop significantly in contrast to our expectation. The results from AR-based models shows TSS = 0.82±0.02 and HSS = 0.20±0.04 when these methods are evaluated solely on active region based confusion matrices. However, when we aggregate them, these models fail to reach the acceptable levels of skill scores as they drop to TSS = 0.32±0.04 and HSS = 0.15±0.02. The reason for these issues may stem from three reasons: 1) limb events are not considered (beyond ± 70°) as there are no reliable magnetic field readings, 2) these models are not optimized for full-disk flare prediction, and/or 3) an independent, equally weighted aggregation scenario in our heuristic approach. Furthermore, the drop in aggregated skill scores can be attributed to the number of high false positives, which is common in rare-event forecasting problems and particularly in flare forecasting. The reason we empirically observed throughout the years for these false positives are often the models' inability to distinguish [C4+ to C9.9] flares from ≥M-class flares as discussed in Pandey et al. (2022). All in all, our first observation is that for full-disk flare prediction, our designated deep learning models are more effective when compared to the AR aggregations as it considers the near-limb events by using a compressed full-disk magnetogram which are suitable to capture the shape parameters in the active regions within and beyond ± 70° of the central meridian of the Sun.

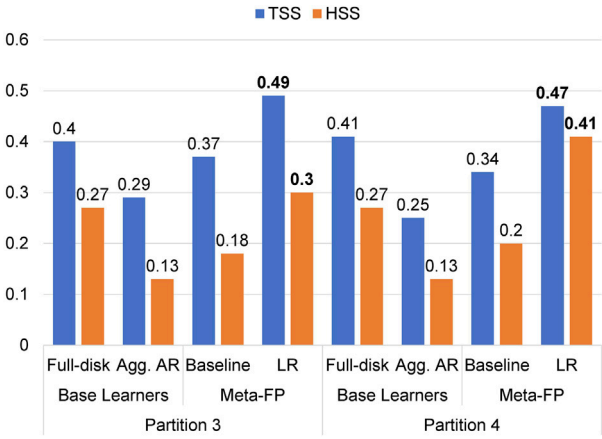

FIGURE 5
Validation Scores of base learners in Fold-1 (base learners trained with Partitions 1 and 2) and the corresponding validation scores of Meta-FP (meta models trained in Expt. 1).

Analyzing our results, we observed that our logistic regression-based Meta-FP improves on both TSS and HSS compared to two base learners and equal weighting baseline meta learner on respective test partitions as shown in Figures 5–7. In our first experiment, we trained two Meta-FP models that utilizes the flare probability scores of two base learners that are trained with Partition-1 and Partition-2 of the respective datasets. We train and validate our Meta-FP with respect to the unused two partitions that are Partition-3 and Partition-4 for the first experiment as shown in Figure 5. Our other two experiments are also consistent with making sure to only use two such partitions that have not been used while training the base learners as shown in Figures 6, 7. While the improvement in terms of TSS and HSS on both the base learner and baseline Meta-FP can be seen across all six logistic regression-based Meta-FP model, the maximum improvement of logistic regression over base learners and baseline can be seen with base learners in Fold-1 (trained with Partition-1 and Partition-2) where the Meta-FP is trained with Partition 3 and tested on Partition-4 (right side of the Figure 5). In this experiment, the logistic regression model improves on full-disk (base learner) in terms of TSS by ~6% and HSS by ~14%. Similarly, it improves on aggregated AR-based models in terms of TSS by ~22% and HSS by ~28%. While we used the equal weighted averaging as a baseline model, it does not improve on the results from the full-disk base learner. However, compared to the baseline for the same experiment (Fold-1) as explained above, the logistic regression model improves by ~13% and ~21% in terms of TSS and HSS respectively.

On an average, we observe that our full-disk model (base learner) has TSS = 0.40±0.07 and HSS = 0.25±0.07 and the AR-based model (base learner) has TSS = 0.32±0.04 and HSS = 0.15±0.02 computed over both test and validation results from all three folds. When we

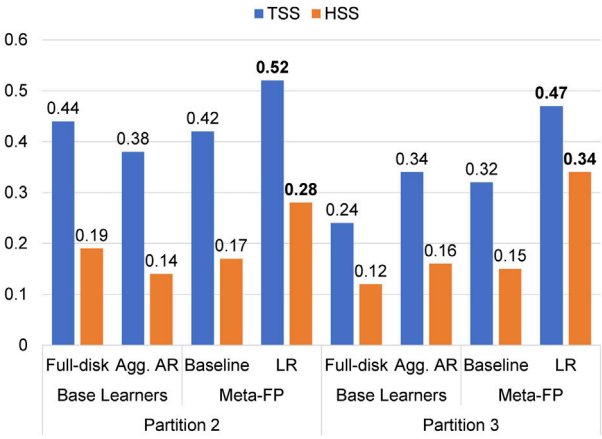

FIGURE 6
Validation Scores of base learners in Fold-2 (base learners trained with Partitions 1 and 4) and the corresponding validation scores of Meta-FP (meta models trained in Expt. 2).

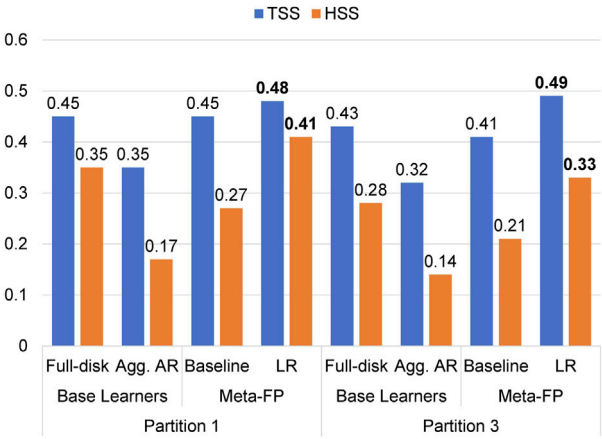

FIGURE 7
Validation Scores of base learners in Fold-3 (base learners trained with Partitions 2 and 4) and the corresponding validation scores of Meta-FP (meta models trained in Expt. 3).

employed the baseline meta learner (equal-weighted average), the average TSS = 0.39±0.05 and HSS = 0.20±0.04 is observed. Given that, equal weighted average is used as a common way to ensemble two or more models, it can be problematic as it could not even surpass the scores of a base learner (full-disk model). With the logistic regression-based meta learner (Meta-FP), the average TSS and HSS observed is 0.49±0.02 and 0.35±0.05 respectively. Therefore, we see that on an average, the Meta-FP improves on the full-disk model by ~9% in terms of TSS and ~10% in terms of HSS. Similarly, it improves on the AR-based model by ~17% and ~20% in terms of TSS and HSS respectively. Finally, when compared to the baseline meta model, it improves on TSS by ~10% and HSS by ~15%.

5 Discussion

Ensemble methods combines multiple models to obtain better predictive performance than could be obtained from any of the constituent model alone. By using an ensemble method, we learn how the single model output can be improved based on 1) maximum voting, 2) equal weighted averaging, and 3) weighted voting. Learning the weights in weighted voting, in the scope of this paper, is structured as a logistic regression problem. One usual way to create an ensemble is to simply average the forecast probabilities of multiple models and provide a final forecast decision, however, it is naive to assume that all base-learners are equally good. Therefore, the main objective of training an ensemble here is to learn and assign better weights for two base-learner predictions by quantifying the level of impact of individual models predictions on the final forecast. The prediction distribution for Partition-3 and Partition-4 used in Experiment-1 for training and testing the ensemble alternatingly and the learned decision-boundary by Meta-FP LR is shown in Figure 8 as an example to show how an ensemble improves over the base-learner by coupling using a linear classifier. The predicted probability distribution and learned decision boundary in Experiment-2 and 3 is presented in Supplementary Figures S1, S2 respectively. Furthermore, the confusion matrices for base-learners predictions and for the consequent ensembles created in all three experiments are presented in Supplementary Tables S1–S6.

Ensemble methods defy the idea of making one model and relying on this model as the best/most accurate predictor we can make. It rather take a multitude of models into account, and combine those models to produce one final model that issues a final forecast. At this point, we do have access to very complex machine learning paradigms that have proven to be very effective in several areas, such as computer vision and image classification. However, relying on the forecast of a single model for rare events like major solar flares might be critical for a system in operation. The model thus obtained might be biased on the dataset used to train the model and can be just as good as the curated dataset used to create the model. Therefore, it is essential to have a reliable flare forecasting model obtained by assembling multiple models with different data modalities to leverage the most with coupling.

6 Conclusion and future work

In this work, we trained a logistic regression-based meta learner for flare prediction that combines the probabilities of two flare prediction models trained with different datasets and machine learning paradigms. While we have two models (base learners) with their own advantages in prediction capabilities, we observed that for base learners, full disk models have better performance for full disk flare forecasting compared to AR-aggregation. Therefore, with a motive of further improving the

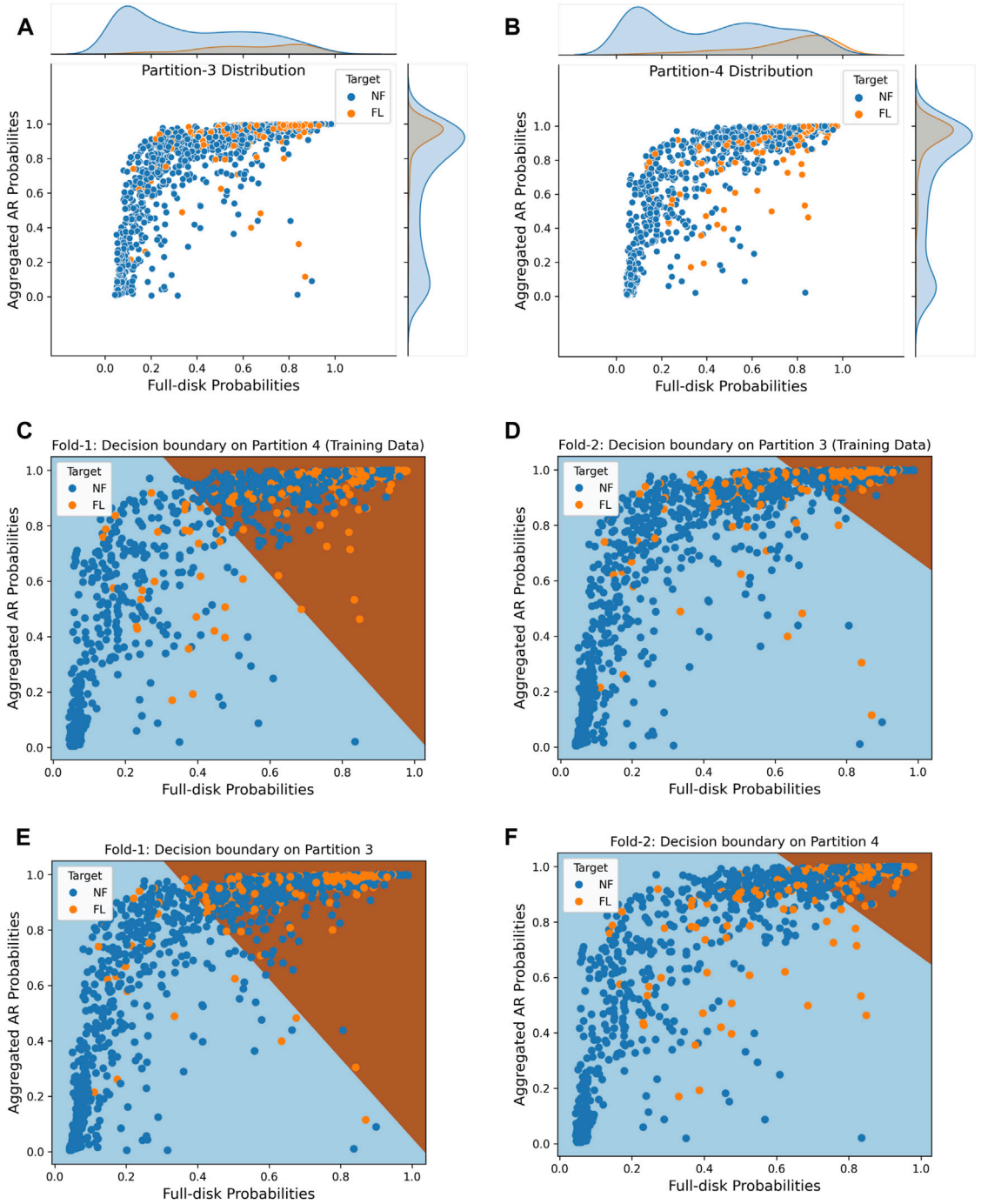

FIGURE 8
The figure shows the distribution of predicted probability scores for all the data instances used in Experiment-1 and the decision boundary learned by the trained logistic regression classifier for both the partitions: **(A)** Distribution of probabilities scores in partition-3 of two base learner mapping to actual ouput. **(B)** Distribution of probabilities scores in partition-4 of two base learner mapping to actual ouput. **(C)** The learned-decision boundary by Meta-FP LR while training on partition-4. **(D)** The learned-decision boundary by Meta-FP LR while training on partition-3. **(E)** The learned-decision boundary by Meta-FP LR validated on partition-3. **(F)** The learned-decision boundary by Meta-FP LR validated on partition-4.

performance of base learners, we explored a simplest way to combine them by training an ensemble flare predictor which automates the task of assigning weights to the outputs of our base learners, thus improving the overall performance of our models and adding robustness to the prediction task compared to equal weighted ensembling.

Furthermore, considering that we only used bi-daily observations, the shape parameters considered in compressed magnetograms proves to be actually powerful. AR-based models on the other hand, using magnetic field data, either as images or derived products, as they are now, will have limited capability although they have higher sensitivity per active region. Therefore, a complementary approach is necessary that does not only rely directly on magnetic field rasters and this work introduces a technique which considers both the magnetic-field parameters and shape-based parameters to obtain flare forecasting models with their own essence and abilities. Finally, we combine these two heterogeneous models into one coupled model using a linear ensemble to improve overall performance. Although we see significant improvements in skill scores after ensembling, our coupled models are not without limitations that are also inherited from our full-disk based model trained with point-in-time bi-daily observations, which overlooks the temporal evolution of magnetic-field parameters of the active regions which can limit the predictive capabilities of full-disk flare predictors. Therefore, our next goal is to formulate the flare prediction task as a video classification problem using full-cadence image sequences that will account for the temporal evolution of active regions. Furthermore, there are several other directions that can be explored such as using a basis function on the aggregated active region prediction probabilities, finding other better aggregation strategies that could boost the performance of AR-based models while computing a full-disk probability and elaborate the ensemble using more sophisticated classifiers, aiming to further improve the predictive capabilities of our models.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

CP built deep learning-based full-disk models, ensemble models, contributed in design of experiments and writing the manuscript. AJ built AR-based classifiers and contributed on reviewing the manuscript. RA was involved in planning and participated in writing the manuscript. MG provided the domain expertise, contributed in model coupling concepts, and reviewing the manuscript. BA conceived of the original idea, and contributed in design of experiments, model coupling concepts, planning, and writing the manuscript.

Funding

This project is supported in part under two NSF awards #2104004 and #1931555 jointly by the Office of Advanced Cyberinfrastructure within the Directorate for Computer and Information Science and Engineering, the Division of Astronomical Sciences within the Directorate for Mathematical and Physical Sciences, and the Solar Terrestrial Physics Program and the Division of Integrative and Collaborative Education and Research within the Directorate for Geosciences. This work is also partially supported by National Aeronautics and Space Administration (NASA) grant award No. 80NSSC22K0272.

Acknowledgments

The data used in this project is a courtesy of NASA/SDO and the AIA, EVE, and HMI science teams [4]. We also want to thank the developers of Helioviewer Project [5] for providing an API for input images.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fspas.2022.897301/full#supplementary-material

4 https://sdo.gsfc.nasa.gov/data/.

5 https://api.helioviewer.org/docs/v2/.

References

Abduallah, Y., Wang, J. T. L., Nie, Y., Liu, C., and Wang, H. (2021). DeepSun: Machine-learning-as-a-service for solar flare prediction. *Res. Astron. Astrophys.* 21, 160. doi:10.1088/1674-4527/21/7/160

Ahmadzadeh, A., Aydin, B., Georgoulis, M. K., Kempton, D. J., Mahajan, S. S., Angryk, R. A., et al. (2021). How to train your flare prediction model: Revisiting robust sampling of rare events. *Astrophys. J. Suppl. Ser.* 254, 23. doi:10.3847/1538-4365/abec88

Angryk, R. A., Martens, P. C., Aydin, B., Kempton, D., Mahajan, S. S., Basodi, S., et al. (2020b). Multivariate time series dataset for space weather data analytics. *Sci. Data* 7, 227. doi:10.1038/s41597-020-0548-x

[Dataset] Angryk, R., Martens, P., Aydin, B., Kempton, D., Mahajan, S., Basodi, S., et al. (2020a). *SWAN-SF*. doi:10.7910/DVN/EBCFKM

Aso, T., Ogawa, T., and Abe, M. (1994). Application of back-propagation neural computing for the short-term prediction of solar flares. *J. Geomagn. Geoelec.* 46, 663–668. doi:10.5636/jgg.46.663

Benvenuto, F., Piana, M., Campi, C., and Massone, A. M. (2018). A hybrid supervised/unsupervised machine learning approach to solar flare prediction. *Astrophys. J.* 853, 90. doi:10.3847/1538-4357/aaa23c

Bobra, M. G., and Couvidat, S. (2015). Solar flare prediction using sdo/hmi vector magnetic field data with a machine-learning algorithm. *Astrophys. J.* 798, 135. doi:10.1088/0004-637x/798/2/135

Breiman, L. (2001). *Mach. Learn.* 45, 5–32. doi:10.1023/a:1010933404324

Breiman, L. (1996). Bagging predictors. *Mach. Learn.* 24, 123–140. doi:10.1023/a:1018054314350

Campi, C., Benvenuto, F., Massone, A. M., Bloomfield, D. S., Georgoulis, M. K., Piana, M., et al. (2019). Feature ranking of active region source properties in solar flare forecasting and the uncompromised stochasticity of flare occurrence. *Astrophys. J.* 883, 150. doi:10.3847/1538-4357/ab3c26

Chen, T., and Guestrin, C. (2016). "XGBoost," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM). doi:10.1145/2939672.2939785

Cinto, T., Gradvohl, A. L. S., Coelho, G. P., and da Silva, A. E. A. (2020). A framework for designing and evaluating solar flare forecasting systems. *Mon. Not. R. Astron. Soc.* 495, 3332–3349. doi:10.1093/mnras/staa1257

Domijan, K., Bloomfield, D. S., and Pitié, F. (2019). Solar flare forecasting from magnetic feature properties generated by the solar monitor active region tracker. *Sol. Phys.* 294, 6. doi:10.1007/s11207-018-1392-4

Feng, L., Gan, W., Liu, S., Wang, H., Li, H., Xu, L., et al. (2020). Space weather related to solar eruptions with the aso-s mission. *Front. Phys.* 8. doi:10.3389/fphy.2020.00045

Florios, K., Kontogiannis, I., Park, S.-H., Guerra, J. A., Benvenuto, F., Bloomfield, D. S., et al. (2018). Forecasting solar flares using magnetogram-based predictors and machine learning. *Sol. Phys.* 293, 28. doi:10.1007/s11207-018-1250-4

Freund, Y., and Schapire, R. E. (1996). "Experiments with a new boosting algorithm," in Machine Learning: Proceedings of the Thirteenth International Conference (Burlington, MA, USA: Morgan Kaufmann), 148–156.

Georgoulis, M. K., Nindos, A., and Zhang, H. (2019). The source and engine of coronal mass ejections. *Phil. Trans. R. Soc. A* 377, 20180094. doi:10.1098/rsta.2018.0094

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Mach. Learn.* 63, 3–42. doi:10.1007/s10994-006-6226-1

Guerra, J. A., Murray, S. A., Bloomfield, D. S., and Gallagher, P. T. (2020). Ensemble forecasting of major solar flares: Methods for combining models. *J. Space Weather Space Clim.* 10, 38. doi:10.1051/swsc/2020042

Guerra, J. A., Pulkkinen, A., and Uritsky, V. M. (2015). Ensemble forecasting of major solar flares: First results. *Space weather..* 13, 626–642. doi:10.1002/2015sw001195

Higgins, P., Gallagher, P., McAteer, R., and Bloomfield, D. (2011). Solar magnetic feature detection and tracking for space weather monitoring. *Adv. Space Res.* 47, 2105–2117. doi:10.1016/j.asr.2010.06.024

Hoeksema, J. T., Liu, Y., Hayashi, K., Sun, X., Schou, J., Couvidat, S., et al. (2014). The helioseismic and magnetic imager (HMI) vector magnetic field pipeline: Overview and performance. *Sol. Phys.* 289, 3483–3530. doi:10.1007/s11207-014-0516-8

Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67. doi:10.1080/00401706.1970.10488634

Huang, X., Wang, H., Xu, L., Liu, J., Li, R., Dai, X., et al. (2018). Deep learning based solar flare forecasting model. i. results for line-of-sight magnetograms. *Astrophys. J.* 856, 7. doi:10.3847/1538-4357/aaae00

Ji, A., Aydin, B., Georgoulis, M. K., and Angryk, R. (2020). "All-clear flare prediction using interval-based time series classifiers," in 2020 IEEE International Conference on Big Data (Big Data) (Atlanta, GA, USA: IEEE). 4218–4225. doi:10.1109/bigdata50022.2020.9377906

Jonas, E., Bobra, M., Shankar, V., Hoeksema, J. T., and Recht, B. (2018). Flare prediction using photospheric and coronal image data. *Sol. Phys.* 293, 48. doi:10.1007/s11207-018-1258-9

Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. *Commun. ACM* 25, 84–90. doi:10.1145/3065386

Li, X., Zheng, Y., Wang, X., and Wang, L. (2020). Predicting solar flares using a novel deep convolutional neural network. *Astrophys. J.* 891, 10. doi:10.3847/1538-4357/ab6d04

Liu, C., Deng, N., Wang, J. T. L., and Wang, H. (2017a). Predicting solar flares Using*SDO*/HMI vector magnetic data products and the random forest algorithm. *Astrophys. J.* 843, 104. doi:10.3847/1538-4357/aa789b

Liu, J.-F., Li, F., Wan, J., and Yu, D.-R. (2017b). Short-term solar flare prediction using multi-model integration method. *Res. Astron. Astrophys.* 17, 034. doi:10.1088/1674-4527/17/4/34

McGuire, D., Sauteraud, R., and Midya, V. (2019). "Window-based feature extraction method using XGBoost for time series classification of solar flares," in 2019 IEEE International Conference on Big Data (Big Data) (Los Angeles, CA, USA: IEEE). doi:10.1109/bigdata47090.2019.9006212

Muller, D., Fleck, B., Dimitoglou, G., Caplins, B., Amadigwe, D., Ortiz, J., et al. (2009). JHelioviewer: Visualizing large sets of solar images using JPEG 2000. *Comput. Sci. Eng.* 11, 38–47. doi:10.1109/mcse.2009.142

Murray, S. A. (2018). The importance of ensemble techniques for operational space weather forecasting. *Space weather.* 16, 777–783. doi:10.1029/2018sw001861

Nishizuka, N., Kubo, Y., Sugiura, K., Den, M., and Ishii, M. (2021). Operational solar flare prediction model using deep flare net. *Earth Planets Space* 73, 64. doi:10.1186/s40623-021-01381-9

Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., and Ishii, M. (2018). Deep flare net (DeFN) model for solar flare prediction. *Astrophys. J.* 858, 113. doi:10.3847/1538-4357/aab9a7

Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., Watari, S., Ishii, M., et al. (2017). Solar flare prediction model with three machine-learning algorithms using ultraviolet brightening and vector magnetograms. *Astrophys. J.* 835, 156. doi:10.3847/1538-4357/835/2/156

Núñez, M., and Paul-Pena, D. (2020). Predicting >10 MeV SEP events from solar flare and radio burst data. *Universe* 6, 161. doi:10.3390/universe6100161

Pandey, C., Angryk, R. A., and Aydin, B. (2022). "Deep neural networks based solar flare prediction using compressed full-disk line-of-sight magnetograms," in *Information management and big data* (Berlin, Germany: Springer International Publishing), 380–396. doi:10.1007/978-3-031-04447-2_26

Pandey, C., Angryk, R. A., and Aydin, B. (2021). "Solar flare forecasting with deep neural networks using compressed full-disk HMI magnetograms," in 2021 IEEE International Conference on Big Data (Big Data) (Orlando, FL, USA: IEEE), 1725–1730. doi:10.1109/bigdata52589.2021.9671322

Park, E., Moon, Y.-J., Shin, S., Yi, K., Lim, D., Lee, H., et al. (2018). Application of the deep convolutional neural network to the forecast of solar flare occurrence using full-disk solar magnetograms. *Astrophys. J.* 869, 91. doi:10.3847/1538-4357/aaed40

Schunk, R. W., Scherliess, L., Eccles, V., Gardner, L. C., Sojka, J. J., Zhu, L., et al. (2016). Space weather forecasting with a multimodel ensemble prediction system (MEPS). *Radio Sci.* 51, 1157–1165. doi:10.1002/2015rs005888

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* 58, 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

Toriumi, S., and Wang, H. (2019). Flare-productive active regions. *Living Rev. Sol. Phys.* 16, 3. doi:10.1007/s41116-019-0019-7

Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* 67, 301–320. doi:10.1111/j.1467-9868.2005.00503.x

frontiers | Frontiers in Astronomy and Space Sciences

TYPE Review
PUBLISHED 15 November 2022
DOI 10.3389/fspas.2022.896877

OPEN ACCESS

EDITED BY
Bala Poduval,
University of New Hampshire,
United States

REVIEWED BY
Marie Farrell,
Maynooth University, Ireland
Verena Heidrich-Meisner,
University of Kiel, Germany

*CORRESPONDENCE
Haroun El Mir,
H.El-Mir@cranfield.ac.uk

SPECIALTY SECTION
This article was submitted to Space Physics,
a section of the journal
Frontiers in Astronomy and Space Sciences

RECEIVED 15 March 2022
ACCEPTED 02 November 2022
PUBLISHED 15 November 2022

CITATION
El Mir H and Perinpanayagam S (2022), Certification of machine learning algorithms for safe-life assessment of landing gear.
Front. Astron. Space Sci. 9:896877.
doi: 10.3389/fspas.2022.896877

Certification of machine learning algorithms for safe-life assessment of landing gear

Haroun El Mir* and Suresh Perinpanayagam

Integrated Vehicle Health Management Centre, Cranfield University, Cranfield, United Kingdom

This paper provides information on current certification of landing gear available for use in the aerospace industry. Moving forward, machine learning is part of structural health monitoring, which is being used by the aircraft industry. The non-deterministic nature of deep learning algorithms is regarded as a hurdle for certification and verification for use in the highly-regulated aerospace industry. This paper brings forth its regulation requirements and the emergence of standardisation efforts. To be able to validate machine learning for safety critical applications such as landing gear, the safe-life fatigue assessment needs to be certified such that the remaining useful life may be accurately predicted and trusted. A coverage of future certification for the usage of machine learning in safety-critical aerospace systems is provided, taking into consideration both the risk management and explainability for different end user categories involved in the certification process. Additionally, provisional use case scenarios are demonstrated, in which risk assessments and uncertainties are incorporated for the implementation of a proposed certification approach targeting offline machine learning models and their explainable usage for predicting the remaining useful life of landing gear systems based on the safe-life method.

KEYWORDS

explainable AI, landing gear systems, certification, risk management, safe-life design

1 Introduction

The aircraft maintenance, repair and operations (MRO) industry is seeing a rise in demand for new aircraft, as well as an increased need for seamless integration and cost-effective maintenance digitisation. Digital, or avionics systems, are rooted as a progressively-important part of the predictive maintenance processes used in aircraft. Examples of such systems are a division of structural health monitoring (SHM), named damage monitoring systems. It consists of load monitoring, also known as operational loads monitoring (OLM), and fatigue monitoring (Staszewski and Boller, 2004). With the advancement in processing power, computing capabilities of onboard systems are rendered able to effortlessly accommodate improved and more demanding loads monitoring sensors and software. This paper explores the improvement of fatigue monitoring systems for landing gear (LG). LG are certified for usage on aircraft using

the safe-life fatigue approach. This approach attributes each component of the LG with a predefined and unchanging service life, after which the component is either:

1) Used as a replacement to a similar component onto the LG assembly of another aircraft, wherein it is certified for a longer life span due to the less impactful load profile on the aircraft in which it will be used.
2) Scrapped and deemed unworthy of service.

The safe-life calculation consists of a load spectrum assigned to the aircraft LG, which consists of an assumption that forms a safety factor. This load spectrum estimation can accommodate improvements, due to its high safety factors. The loads applied in-service are highly probable to be less impactful on the life of the part than what is proposed by the safe-life estimation. The assigned service life may therefore be extended if the loads are monitored with OLM equipment. The disparity in stress-life (S-N) curves also contributes to the value of the safety factor applied when setting the safe-life of the component (Irving et al., 1999).

The safe-life method assumes a set of load profiles to result with a number of trips that the LG will be safely able to travel. Instead, basing the replacement of the LG on the amount and severity of loads encountered can be accomplished by collecting data with the use of sensors, thereby allowing for the quantification and classification of the factors causing imminent fatigue failure. Currently, such an ideology is approached by a form of OLM systems, which consists of strain gauges placed on military aircraft (Hunt and Hebden, 2001; Dziendzikowski et al., 2021) wherein the strain output is transformed into digital signals that are thereby converted into stress histories, resulting in a loading sequence. Nevertheless, this method of fatigue assessment is inadequate for structural damage detection, by virtue of leaving out "a factor of two to three in fatigue life to be gained if damage could be monitored more adequately" (Staszewski and Boller, 2004). Furthermore, placing additional devices for measuring such parameters invites more reliability issues and an increase in maintenance costs (Cross et al., 2012).

This has, in turn, given birth to the use of Artificial Intelligence (AI)-handled solutions, with an expected growth due to commercial demands, closing the gap where safety-critical applications and the novelty of machine learning (ML) algorithms are deemed to ultimately collide and remould the way that the MRO industry has been assessing aircraft structural health. Successively, the emergency of placing a basis for the certification and risk management of such approaches arises, ranging from the ML explainability levels to the uncertainties in data exchange and collection in-service, due to the non-deterministic qualities of ML when compared to currently-used avionics software and equipment.

Aerospace industry regulators have put forward their interest in the use of ML, for its data-driven benefits, in digital systems related to all levels of the aircraft development cycle, from design to manufacturing, maintenance and operation to communication, by assigning committees and publishing recommendations. EUROCAE created working group WG-114, and SAE started committee G-34, both working in conjunction with the aim of certifying AI for the safe operation of aerospace vehicles and systems, including Unmanned Air Systems. Their published work so far has been the "*SAE AIR6988 & EUROCAE ER-022 Artificial Intelligence in Aeronautical Systems: Statement of Concerns*" (SAE International, 2021a; EUROCAE, 2021). It critically assesses current aeronautical systems encompassing the whole lifecycle of airborne vehicles and equipment and how they fall short of covering AI and, more specifically, ML challenges.

A coverage of upcoming certification requirements for the usage and collection of data from aircraft sensors to predict LG remaining useful life (RUL) is employed in this paper. It is based on:

1) The WG-114/G-34 SAE AIR6988 document (SAE International, 2021a).
2) EASA AI Roadmap (EASA, 2020).
3) EASA CoDANN & CoDANN II reports (EASA and Daedalean AG, 2020, 2021).
4) EASA Concept Paper: First Usable Guidance for L1 ML Applications (EASA, 2021).

These documents have been chosen due to their relevance to the subject of this paper. Nevertheless, the documents also incorporate previous standardisation requirements (such as ARP4754A and DO-178C, DO-254) and guidance by means of addressing their limitations in light of AI requirements for avionics applications. For a survey and taxonomy of the recently-published proposals and guidance papers on practical ML application for use in aviation, the article by the subgroups of the SAE G-34/EUROCAE WG-114 standardisation working group on ML lifecycle development (Kaakai et al., 2022) is recommended to the reader. It sets out the ML development lifecycle guidelines for certification in aeronautics, that are to be the core of the forthcoming publication by SAE: the "*AS6983 Process Standard for Development and Certification/ Approval of Aeronautical Safety-Related Products Implementing AI*".

2 Paper contribution

This paper encapsulates the certification approaches and requirements currently available for landing gear (LG) and AI applications in the aerospace industry, to cover all issues related to machine learning (ML) and safe-life, which will eventually lead

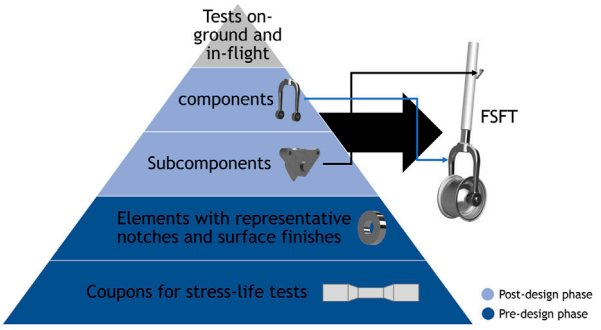

FIGURE 1
The life cycle and process of testing for aircraft components begins, generally, with the pre-design (non-specific) phase, consisting of material coupon and element testing. The process may be compared to stages, or levels, where the testing of each stage, with its corresponding attribute, is required to be completed successfully in order for the next, upper, and more structurally complicated stage to begin. At the bottom of the pyramid, a coupon is stress-tested physically; a specimen of material tested for the nature of its properties. Then, to be more representative of the part used, elements with a similar surface finish and treatment, in addition to including notches characteristic of the part in question, are tested. At the post-design phase, sub-component and component physical testing, which considers the impact of environmental conditions and key features of the LG e.g., piston-cylinder assembly performance, leads to FSFT tests, where the whole structure, or its vital system component sets, such as the shock absorber system, steering system, and wheel and braking system are flight-simulated. Finally, the structure is tested on-ground and in-flight. The process shown in the figure is also known as the "building block" test approach.

to a philosophy for the certification of ML for LG, and whether it may be employed using AI in the next decade. Major issues related to AI that affect the LG environment will have been identified by the reader. It is important to note that the goal of this paper is to illuminate and ease the process of the development of a certification methodology, where a ML algorithm/set of algorithms are to be used for the purpose of LG remaining useful life (RUL) prediction. The paper does so by assisting with confusions a newcomer to this field may have, as the area of certification is quite tough to manoeuvre. The reader may then form a method with which to begin and is guided along the way with the allocation of their requirements through the elimination of current standards and allocation of assurance case tools available, as well as building, block by block, a clearer image of where they stand in the process of complying with those standards, in order to develop adequate use case scenarios.

3 Current Safe-Life Assessment

Safe-life fatigue analysis of aircraft structures is a principle of design in which an estimation is placed prior to the first operation of a component in-service. This estimation is based on the evaluation of the structure's ability to sustain its original crack-free status while being exposed to cyclic loads in-service, such as landing, take-off, and taxiing, which all contribute to impacting the fatigue life of the LG components (Ladda and Struck, 1991). The safe-life analysis places a value of operational hours for the part in question in which it would be replaced afterwards, regardless of whether visible fatigue cracks form in the structure. This approach therefore deems the part inoperable and unsafe for use on the aircraft after those specified hours or cycles. Looking towards how this approach begins, the component's lifecycle and its workarounds, as well as how they fit into the whole aircraft's production plan, come into question.

3.1 Aircraft testing lifecycle

The life cycle and process of testing for aircraft components begins, generally, with the pre-design (non-specific) phase, consisting of material coupon and element testing. The process may be compared to stages, or levels, where the testing of each stage, with its corresponding attribute, is required to be completed successfully in order for the next, upper, and more structurally complicated stage to begin. At the bottom of the pyramid in Figure 1, a coupon is stress-tested physically; a specimen of material tested for the nature of its properties. Then, to be more representative of the part used, elements with a similar surface finish and treatment, in addition to including notches characteristic of the part in question, are tested. At the post-design phase, sub-component and component physical testing, which considers the impact of environmental conditions and key features of the LG e.g. piston-cylinder assembly performance, leads to FSFT tests, where the whole structure, or its vital system component sets, such as the shock absorber system, steering system, and wheel and braking system are flight-simulated. Finally, the structure is tested on-ground and in-flight (Ball et al., 2006). The process shown in the figure is also known as the "building block" test approach (Wanhill, 2018). The post-design phase is directly connected to airworthiness certification due to the phase containing components and parts of the aircraft ready for use and in their final stage of design (Ball et al., 2006). Compliance with airworthiness standards demands the identification of loads encountered and the load cycles in order to schedule corresponding component visual check-ups (Wong et al., 2018).

3.2 Safe-life requirements

Currently, the only components in the aircraft to which this safe-life fatigue estimation may be applied are the LG. The LG's incapability of accommodating crack initiation and expansion is due to its components consisting of high-strength alloys that

motivate rapid crack propagation. Two fatigue detection approaches may be used for the safe-life fatigue analysis of a metallic aircraft component: the stress-life approach and the strain-life approach (Wanhill, 2018). The ways in which a safe-life is specified to obtain certification allowing the use of the component on large aircraft requires:

1) Full-scale fatigue tests (FSFT) encompassing the whole structure physically being tested with methods, such as strain gauges mounted to localise and quantify strain (Dziendzikowski et al., 2021).
2) The testing of specific components of that structure in question–in the case of LG, that would be its individual components each tested separately for fatigue resistance.
3) The use of hypotheses and the stress-life approach *via* Miner's rule for damage accumulation, whereby damage fixated by each repetition of stress due to load applications is assumed equal (Federal Aviation Administration, 2005). The Miner's rule, also referred to as the Palmgren-Miner linear accumulation hypothesis, states that the damage due to fatigue is equal to a singular value of "one" as long as cyclic application of this load has reached an amount validating its appearance on the fatigue curve (Schmidt, 2021).

The LG encounters multiple loads in succession, contributing to high cycle fatigue (HCF). Low cycle fatigue, which is correlated with strain life curves, is characterized by plastic strain. Stress-life curves, on the other hand, are used in high cycle fatigue, where fatigue is mostly in the elastic region and plasticity can be neglected. Landing gear stresses do not reach the plastic deformation region of the material in each of its components, which is why the stress-life fatigue approach is used. There is an abundance of available data for the stress-life approach, and it is applicable specifically to HCF. In addition to the pre-design nature of the landing gear structural CS-25 airworthiness certification requirements for large airplanes, the safe-life fatigue analysis that is currently used in the LG certification process utilises Miner's rule for damage accumulation, using S-N curves. These curves conform to a certain material coupon, where the material must be the same as that used in the component in question. As Pascual and Meeker (1999) discuss, an S-N curve for a certain material is a representation of the fatigue data of a coupon of that material, in the form of a log-log plot containing cyclic stress 'S' values *versus* 'N', the median fatigue life articulated in cycles to failure. It is key to note that S-N curves are derived from a specific stress-ratio. They also contain scatter, which is an uncertainty associated with failure in fatigue. Additionally, two factors parametrise S-N curves: probability of survival and probability of failure. Both introduce uncertainty factors to be applied for the final prediction of a component life. Fatigue is non-deterministic, as opposed to static loads, e.g. Component A tested for fatigue using the identical test parameters as Component B will result with a fatigue life significantly different than that of its proponent. This introduces scatter in S-N curves used for fatigue prediction.

Required in addition to these curves is the fatigue spectrum: data on the applied loads, how frequently they manifest, and how their occurrence fits in the grand scheme of load sets applied, in terms of their timing and repetitions. Flight profiles are a set of load variances, representative of a certain flight block. These profiles add up to form a spectrum for fatigue prediction (Schmidt, 2021). The spectrum may also consist of flight hours in addition to flight cycles if the nature of the mission of the aircraft is mixed in terms of range duration. Established design lives can be divided into three categories with their corresponding cycle ranges:

1) 50,000 cycles for short-haul flight aircraft, e.g. A320.
2) 25,000 cycles for long-haul aircraft, e.g. A350.
3) 10,000 cycles for tactical aircraft.

The steps for safe-life fatigue analysis of LG are as follows, summarised in Figure 2:

Step 1. S-N curves are generated by performing uniaxial cyclic stress amplitude loads on numerous material samples until failure. This material data may also be extracted from readily-available scatter data and must comply with the 99/95 standard, with an applied scatter factor of 3 at a minimum (Fatemi and Vangt, 1998) The curves are also altered according to the in-service factors of the landing gear environment, which are not experienced by the coupons tested in monitored conditions. The resulting curves are referred to as "working curves" (Wanhill, 2018).

Step 2. Meanwhile, a stress-time history plot is derived from a load-time history plot for the LG component by referring to the geometry of the component.

Step 3. Methods such as Bathtub/Rainflow counting are performed on the load-time history plot which is a stress-time history plot after Step 2, to result in stress cycles and a mean stress value for each cycle count (Le-The, 2016).

Step 4. The cycles are converted with their mean values to fully-reversed stress cycles in order to extract equivalent data when referring to the S-N curves for the material used in the component of the LG (for data compatibility purposes). This is done *via* mean stress correction techniques, such as the Goodman mean stress correction Eq. 1. (σ_0), as discussed by Hoole (2020), is the value of the fully-reversed stress cycles. (σ_a) is the stress amplitude value of those stress cycles, (σ_m) is their mean stress level, and (σ_{UTS}) is the material's defined ultimate tensile strength.

Step 5. Fatigue damage (d) accrued by each applied cyclic stress amplitude (σ_0) is formulated using Miner's rule. As per Equation 2 (n) is the frequency at which (σ_0) is applied, and (N_f) is the number of cycles to failure. (D_T) is total damage

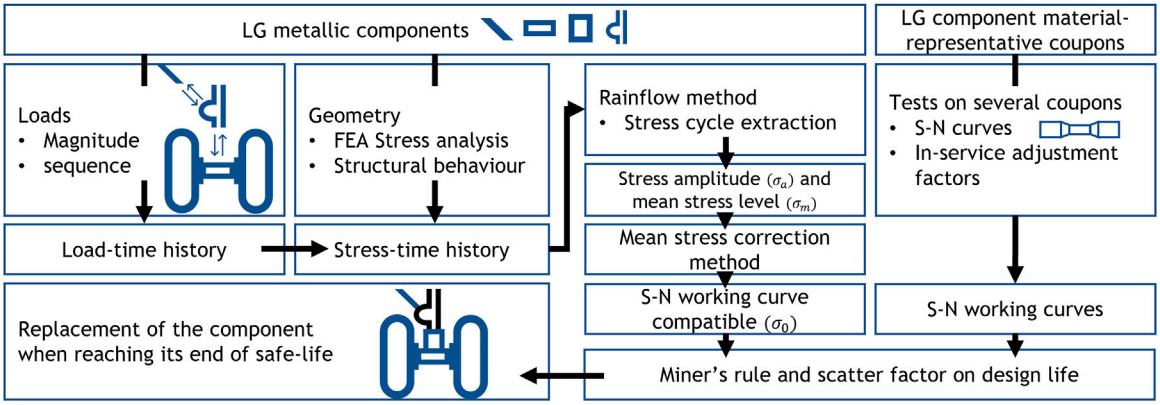

FIGURE 2
Steps for safe-life fatigue analysis of landing gear.

accumulated from the stress formed by the cycles. As (Hoole, 2020) mentions, a value of 1 for (D_T) signifies failure of the part in question, meaning it has reached the end of its fatigue life, and representing failure Eq. 3.

Equation 1 Goodman mean stress correction

$$\sigma_0 = \frac{\sigma_a}{1 - \frac{\sigma_m}{\sigma_{UTS}}} \tag{1}$$

Equation 2 Fatigue damage, Miner's rule

$$d = \frac{n}{N_f} \tag{2}$$

Equation 3 Total damage, Miner's rule

$$D_T = \sum d \tag{3}$$

3.3 Machine learning for safe-life prediction

Studies performed by Holmes et al. (2016) attempt to form a correlation between flight parameters and loads applied to a LG structure attached to a drop test rig, *via* the use of two types of nonlinear regression models as part of their ML approach: multi-layer perceptron (MLP), and Bayesian MLP. The data accumulated consists of inputs, such as wheel speed, accelerations in the LG, and similar flight variables, consisting of kinematic approaches; related with changes in velocity and displacement, in order to result in load induced on the LG. Since the MLP is Bayesian, it requires a specification of a prior. A gaussian prior distributions was used. The functional efficiency of the used neural network (NN) is calculated by acquiring the mean-square error between the predictions formed by the model and the measured targets. Optimising the NN is done using gradient descent. As for the weight uncertainty of the NN, it is reduced by assigning each weight a probability distribution. Additionally, the input datasets were filtered due to noise in acceleration measurements being higher than actual load values recorded through strain. The physical test of the LG rig included assumptions made to simulate a landing environment *via* spinning the wheels before impact, changing the angle of impact of the LG, and dropping the structure from variable heights. These impacts were then measured using strain gauges placed on the LG rig components and load cells placed on the platform on which the LG drops. Another method used for data collection and prediction included the use of Greedy algorithms and Gaussian process (GP) regression; a class of Bayesian non-parametric models. With the use of flight test data parameters to predict landing gear vertical load. GP was used as it trains faster than MLP, and the computations necessary for GP regression are simplified by the fact that a distribution directly over candidate functions can be defined, rather than over the parameters of a predefined function (as would be necessary for a Bayesian neural network for example). They are likewise compact. Cross et al. (2013) found correlations with the general trend of data prediction. Later studies put forth the requirement of physics-informed data to predict landing gear loads to a usable level. These ML approaches result in models that are able to predict loads, where a model requires that it be aircraft-specific. Nonetheless, different surfaces on which the physics-informed ML model (using both LG drop test data and flight test data) was used on still produced acceptable outcomes.

4 Machine learning techniques

As a branch of AI, ML is a computing field that operates with the use of computational methods related to statistics, probability, and computing theory. ML is used by systems to

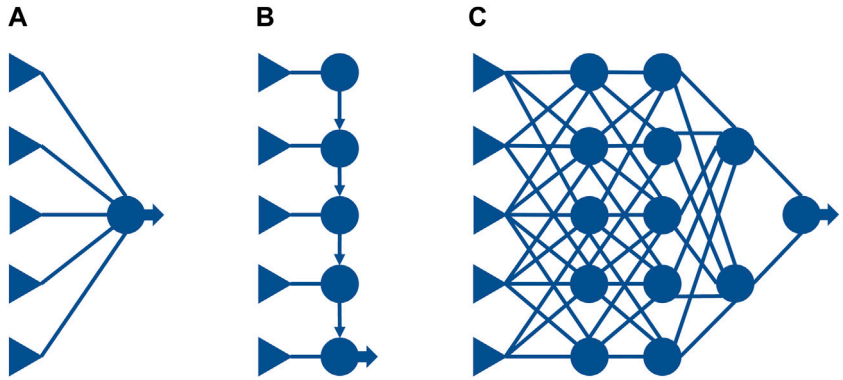

FIGURE 3
Neural networks in comparison to linear regression and decision lists. A linear model **(A)** such as linear or logistic regression is able to compute and take in a high number of variables for input. Nevertheless, the path from input to output is relatively short due to all variables being multiplied by a single weight, in addition to the principle that these input variables are not capable of communicating within themselves. This renders them able to only act for linear functions and boundaries related to the input space. Decision lists **(B)** allow for these long paths of computation to occur, but depends on the input variables being of a similar size to the output variables. Neural networks **(C)** merge these two methods together, allowing for the input variable interactions to be complex and incorporate long computation paths. The benefit of this model is the ability to represent applications, such as speech, photo and text recognition.

learn patterns or monitor data input and apply statistical algorithms to infer the required output depending on the type of algorithm being used. The method by which models of ML operate may be described as follows: "a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, measured by P, improves with experience E" (Mitchell, 1997). An example is the use of statistical methods, where algorithms classify or foresee similarities in data being extracted to suggest a best-case scenario. The input data used to build the ML model, through the stages of its creation, are categorised into three common datasets, forming the ML algorithm: training, validation and testing. Furthermore, ML may be categorised into four types in terms of the method with which it learns: supervised, unsupervised, semi-supervised and reinforcement learning. These reflect the types of feedback-input relationships. Supervised learning occurs when input and output pairs of labelled data are monitored and a function is learned as a result, mapping input and output accordingly. In unsupervised learning, the unlabelled data input is studied without any feedback and patterns are found within that input. Semi-supervised learning trains on labelled and unlabelled data, improving model accuracy when compared to a supervised learning algorithm. As for reinforcement learning, the algorithm is given a response at the end of each set of decisions made, as part of each step in its decision process. Its aim is twofold: the initial improvement of performance due to learning from previous action-result combinations, and the eventual output of the most optimal long-term reward that it may be assigned, e.g. lengthening the duration of a game in order to win eventually instead of winning over an opponent earlier on only to ultimately lose in a game of checkers (Russell and Norvig, 2022).

4.1 Artificial neural networks

When ML involves layers of computing segments that are adaptable and unembellished, that is the term known as deep learning. Deep neural networks (DNN), a subset of ML, are the most common form of deep learning. They are based on one or more layers adapted for large data input sizes. When containing less than 3 layers, the term neural networks is used. Figure 3 is a demonstration of how a DNN may relate to shallower ML models. A linear model (a) such as linear or logistic regression is able to compute and take in a high number of variables for input. Nevertheless, the path from input to output is relatively short due to all of the variables being multiplied by a single weight, in addition to the principle that these input variables are not capable of communicating within themselves. This renders them able to only act for linear functions and boundaries related to the input space. Decision lists (b) allow for these long paths of computation to occur, but depends on the input variables being of a similar size to the output variables. Neural networks (c) merge these two methods together, allowing for the input variable interactions to be complex and incorporate long computation paths. The benefit of this model is the ability to represent applications, such as speech, photo and text recognition (Russell and Norvig, 2022). DNN, which are characterized by multiple layers instead of one (usually three

layers or more), tend to be more accurate and effective in task purveyance. A term commonly found when dealing with the inexplicability of DNN, the black-box is scientifically associated with a system of known and observable inputs and outputs, and no knowledge or observation to be made on how the inner mechanisms of that system may be. In the case of a NN, although the code may be observed, it is functionally referred to as a black-box due to the nature of constant reorganization of the computational NN layers. Modelled based on the workings of the brain; firing neurons with correlated weights to result with decisions, the black-box model and nature of DNN has recently been subjected to theories attempting to explain its method of operation, some attempting to generalize to all types of DNN modes of operations (Alain and Bengio, 2016; Zhang et al., 2016; Schwartz-Ziv and Tishby, 2017; Poggio et al., 2020), and others focusing on certain NN methods and the available interpretation approaches (Guidotti et al., 2018; Montavon et al., 2018; Azodi et al., 2020).

4.2 Machine learning challenges

In health monitoring of aerospace structures, an advisory system provides recommendations that are backed up with evidence, which are in the form of:

1) Sensor output from damage monitoring systems, which consists of direct measurements from the aircraft component/s in question.
2) Flight parameter and environmental conditions derived outputs, that are indirect measurements. These materialize in the form of operational monitoring systems (OMS).

The OMS is a sub-component of SHM, and a system similar to damage monitoring, with the difference being that its measurements are of a derived nature (SAE International, 2021b). The former is the system most useful for the purpose of sensor replacement purposes. Nevertheless, ML may be used as a part of both damage monitoring and OMS. Requirement-wise, the software that provides an envelope around the ML tool needs to be developed to a defined quality process, according to a distinct software control method. That occurs when embedding the software. In addition, it must be demonstrated that the ML black-box may be used in a reliable and robust manner. Questions important for the setting of requirements in the ML uncertainties capture are:

1) 'Is it using recognized libraries?'; code pre-written for repeated usage. The reader is referred to (Nguyen et al., 2019) for a description and comparison of current ML libraries and frameworks.
2) 'What was the quality process used in creating that software?' A framework by Murphy, Kaiser and Arias (2006) proposes a ranking for supervised ML algorithms, consisting of "tools to compare the output models and rankings, several trace options inserted into the ML implementations, and utilities to help analyse the traces to aid in debugging".
3) 'What is the validation process of the model itself (the data-driven part of the training)?'

Just as important, the training, testing and validation processes must be robust and contain a level of assurance that provides accurate predictions when implemented live, in order to be moved from an advisory status to a fully-trusted status. What data is used, its source, reliability, coverage provided by the data (e.g., whether it covers all types of landing for the aircraft type in question), and all operational cases (e.g., heavy landing, light landing, crosswind conditions, icy conditions on runway) are questions to be asked when formulating a data-based rigorous selection process. Moreover, whether the validation data is based on physics data from finite element (FE) models, or testing rig scenarios, plays a significant role in the assurance process.

Deep learning encounters challenges pertaining to its data in which features are represented, specifically with the initial step of obtaining that data, wherein labelling is required (Khan and Yairi, 2018). Furthermore, challenges introduce themselves, according to Khan and Yairi (2018) in the following aspects and identifiers of the deep learning bubble:

1) Specific deep learning architectures and their categorisations into the most suitable pertaining applications have not been yet solidified due to researchers' inadequate justifications of why they used those specific methods and as to why a certain number of layers was most suitable for their applications.
2) Comparison of the architectures has not been standardised, whether it be in terms of time consumption, resource management, computational requirements, or data loss.
3) With regard to structural health management, deep learning applications will have to recognise the failures or faults according to their corresponding environments and be able to diagnose issues, such as no fault found (Khan et al., 2014a; 2014b).

Of the problems faced, imbalanced data issues arise. As discussed by Liu et al. (2009), the cause of imbalanced data results while learning is due to classification and clustering situations, as a result of the classes being learned having considerably more data when compared to their counterparts. Furthermore, cases which are uneven occur due to the intrinsic nature of those events, as well as the additional expense that may result from obtaining these examples for learning in the algorithm. These imbalanced data classification issues may be overcome with the following approaches: pre-processing, cost-sensitive learning, algorithm-centred, and hybrid methods (Kaur et al., 2019).

The data used for training an algorithm may be improved with pre-processing methods, when the algorithm faces a class of data containing an abundant number of examples while the other class contains a lower amount. Due to the accuracy of classification being negatively affected if not for sampling methods, they represent an important step towards avoiding bias (Barandela et al., 2004). The aim of these methods is to balance the classes of data and result with less bias *via* either over-sampling or under-sampling. These two methods operate by manipulating the training data space.

Over-sampling: Of the classes available in pre-processing data, the minority class that happens to bias the data is duplicated in sample packets and the data is therefore balanced in terms of the final dataset. Under-sampling, on the other hand, performs the opposite by randomly extracting samples from the major class (leading to the probable negative aspect of deleting important data) in order to result in equal amounts of the minor and major class. Over- and under-sampling may be combined to form the hybrid sampling method, where they are both used to result with balanced data for pre-processing (Xu et al., 2020).

Bias and variance are concluded to be the key issues in ML applications. They are to be addressed, according to (EASA and Daedalean AG, 2020), based on the following two methodologies:

1) Datasets with bias and variance need to be distinguished from opposing datasets and effort shall be put into reducing such bias and variance within the data itself.
2) The bias and variance need to be evaluated based on the level of risk they impose upon the ML model.

Feature selection and extraction are another means of selecting features more suitable for the classification at hand at the pre-processing stage (Kursa and Rudnicki, 2011). The classes of feature selection would be the filter method, wrapper method, and embedded methods (Guyon and Elisseeff, 2003).

4.3 ML risk management

A ML workspace is a framework in which the algorithm's training takes place. The workspace allows for the specification of the coding language package to be used, its training preferences, and the workspace variables. According to SAE AIR6988, as part of the advised requirements to forming certification standards for the data selection and validation of ML systems, the workspace should be covered with a certain level of protection to prevent "data poisoning or tampering", whether it be intentional or not, by the workspace user or intruder. The effects of such an intrusion would include false outputs and algorithm decisions, e.g., importing additional data into the training dataset which cause the algorithm to develop a deceptive result while assuming that the training process is untampered with. Moreover, any non-complying data must be detected and removed from the dataset after the validation step. Additionally, the "probabilistic nature of ML applications" must be taken critically when assessing and forming the safety process analysis.

For the certification of the method in which data is selected and validated, validation for ML would partition a block of data, representing the entire operational profile of a landing system, into 3 types:

1) Training, in which the model in this cycle is trained and compared with the results from an independent dataset which would be the validation set, and a decision is formulated: is this model good enough or does it require further refinement? This decision set is part of the training cycle, clarifying the need for the validation set to be independent of the training set.
2) Testing, where each of these datasets needs to conform with IID (Independent Identity Distributed) and be of good coverage. For example, in a scenario where hard landings are part of the data input, a similar number of hard landings in each of those three datasets must be clearly present in order to avoid the inevitability of bias.
3) The validation process of the model itself, in which the safe-life approach for LG RUL assessment would be the benchmark for this paper's purposes. The model's performance in this step is evaluated by means of using the validation dataset (set aside and unused, as part of the data partitioning procedure done beforehand) and observing the output to decide whether it is acceptable, signalling the readiness of the ML algorithm for use in a real-life scenario, if so.

Risks in ML are categorised, in terms of robustness, into two kinds (EASA and Daedalean AG, 2021):

1) Algorithm robustness, where the algorithm used for learning is tested for robustness as the training dataset is changed.
2) Model robustness, in which perturbations in the input to the algorithm are used for the identification and quantification of the robustness of the training model.

As pointed out in AIRC6988, the traditional form of safety assessment has always been to realise the orders of system failure by means of its own component-level intercommunication with other systems. This could be improved for the case of AI applications due to their complicated ecosystem interactions. The interaction of the system with "external factors" is one improvement to be noticeably important, due to its probability of forming failure conditions in the case of AI applications. Such a safety approach already exists as part of the SOTIF_ISO 21448 document for certification based on the automotive industry's "advanced algorithms" system inclusions. This approach assesses the following:

TABLE 1 RUL ML black-box issues and proposed mitigations.

RUL ML black-box issue	Corresponding mitigation
ML models need to cater to the varying nature of fatigue life scatter in data points in order to appropriately "characterise the probabilistic property of fatigue lives given a specific condition"	Certification requirements must capture fatigue life scatter data and be able to predict fatigue life probabilistically
ML models learn correlations between data input and output via the means of data extraction, leading to the possibility of contradicting physics principles	Certification requirements must adapt to models trained with different datasets
Using a model trained on one data range may result inaccurately when the same model is implemented on a different framework due to the possibility of data overfitting	Certification requirements need to incorporate vital landing gear operational uncertainties, such as hard landings, as well as temperature and environmental variations

1) Both the system and sub-system levels of AI are tested for functionality and performance.
2) The probable sources of failures mitigated by the functional aspect of the system must be pointed out and their causes reassessed.
3) These probable failures must be avoided by the means of "functional modifications".

Putting these advisories into effect, in the case of issues that will arise due to the usage of black-box ML models in order to model fatigue life, an advisory example is shown in Table 1. A high-level mitigation, or requirement, is set up for each ML data issue.

Certain ML infrastructures, such as continual learning pipelines, allow for the ability to add continuous data points in a well-formulated algorithm, allowing for the data output to be optimised in terms of the assessment of structural integrity and maintenance scheduling. This is deemed an improvement for data collection purposes, but increases the risks for uncertainties specifically when considering external data collection factors, where the potential sources of data in the case of LG fatigue detection include:

1) Fatigue tests implemented physically on the parts themselves in a controlled environment.
2) Flight data of the same aircraft and landing gear from other operators.
3) Maintenance observations.
4) IVHM data, including output from strain gauges on-board the aircraft and LG assembly.

4.4 Explainability

Certification for ML applications in LG may be applied *via* explainability, by the means of connecting data point values from features; values and properties of a monitored process (Bishop, 2006). Among the important requirements for the acceptance of a ML algorithm for use in an industry that is to accept AI solutions over the coming years, trust reappears as a main question at hand, which is where explainability comes into play. Applications and methods for instilling trust into a certain AI approach are reflected in the currently-adopted Intelligence Community Directive (ICD 203) and the SAE AIR6988 documents. These both serve the purpose of proposing the standards required for the application of AI in the aerospace industry, as well as emphasising the need for explainability (Blasch et al., 2019). Additionally, explainability is a part of the four building blocks of the framework in EASA's guidance for ML applications paper (EASA, 2021), in addition to the DEEL white paper (DEEL Certification Workgroup, 2021) that concentrates on the properties an ML system should have, and specifies those to be "auditability, data quality, explainability, maintainability, resilience, robustness, specifiability, and verifiability" (Kaakai et al., 2022).

Explainability is a method by which the transparency of a ML black-box may be improved, where the ML model being explained gets its model prediction uncertainties specified by the user, as well as the clarification of the method with which the feedback of the model is interpreted takes place. Such explainable methods have already been achieved by the means of the research done by Smith-Renner et al. (2020):

1) Ensuring fairness in the model with which the end users may interpret the meaning of the results in a language that conforms with their own specific knowledge and terminologies, while assessing bias in the meantime (Dodge et al., 2019).
2) Adjusting the expectancies of end users to comply with the end results of the explainable AI method being used in which uncertainties in the ML model itself are incorporated for the user to be prepared in terms of the model perception (Kocielnik et al., 2019).
3) Enclose trust of an explainable AI agent in order for users to return to such an ML algorithm repeatedly for similar use case scenarios encompassing the model's features of its system, its agents' reliability, and the intentions with which trust is to be instilled (Pu and Chen, 2006).
4) Improve the recommendation rigor of the explainability of the black-box ML model by means of clarifying to the user

which parts of the model are the most important for the use case scenario at hand while referring to the conceptual model in the user's mindset (Herlocker et al., 2000).

Furthermore, explainability may be organised in regard to its approach to the ML algorithm, in which feature selection and feature extraction are two distinguished methods. Feature extraction creates non-detectable features from those that have already been found in the algorithm (Guyon et al., 2006),whereas feature selection evaluates each and every feature in the model after which these features are deemed either adequate or inconsistent for use in the model (Guyon and Elisseeff, 2003).

Another term important to the explainability approach is whether it is local or global in reach. If local, when provided with a conditional distribution, the input clusters of small regions of that distribution lead to how the ML model's predictions are interpreted by the explainable method. As for the case of it being global, average values are the lead source taken for interpretation of distributions fully encompassing the model's condition (Hall et al., 2017).

The method of adopting a ML model's features depends on both the ML model being used, as well as the fatigue failure model being implemented, resulting in the dependence on the sensor data taken ultimately during flight, take-off, and manoeuvres on the landing strip. The usage of features has been resorted to due to the nature of the way in which an ML model operates; by operating on 'single values per case' (Ten Zeldam, 2018). As the ML model formulated to operate on failure diagnosis trains on maintenance data and usage data, while simultaneously filtering outliers, and labelling each feature for the readiness of the model, these labelled features will then need to be categorised based on their relative importance to the fatigue failure of the LG components being studied. These values are compared to predefined value ranges that dictate whether a component's stress reactions qualify it as leading to fatigue failure due to the likely repeatability of this value and its cycling resulting in a HCF failure. The values shall include tyre wear, side-stay loads, impact loads, shock absorber travel distance, as well as distance travelled by the wheel, in addition to forces applied on the axle of the LG. The features are then transferred to classes, or diagnoses (Ten Zeldam, 2018). This methodology does result in relative feature importance, informing the end user of how critical a feature is by relating its likelihood of occurrence to the results of a simulated model.

The need for explainability in certification-required applications is bringing forth work such as that by Viaña et al. (2022), where an algorithm is formed of explainable layers; using clustering for parameter initialisation, overcoming state-of the-art algorithms when it comes to fuzzy system-based combinations.

5 Certification and its challenges

Commercial avionics systems and equipment are composed of software and hardware components, developed to comply with their corresponding design standards. These standards are covered by the two leading documents that the FAA and EASA certification authorities refer to for the approval of the systems in question: DO-178C/ED-12C for the compliance of avionics software development with airworthiness requirements (RTCA, 2012), and its complementary document, DO-254/ED-80 for the design assurance of avionics equipment, consisting of both hardware and software (RTCA, 2000). These documents introduce an iteration of design assurance levels (DAL) that are also used in other avionics certification requirement documents, such as ARP4754A. DAL are measures assigned to each function in the avionics system of an aircraft, be it software-based in the case of DO-178C or hardware based for DO-254. The values of these functional measures range from A to E in alphabetical order. They correspond to cases of catastrophic effect to those of no safety effect on the operation of the aircraft, any form of overload on the crew, and therefore the safety of both (Fulton and Vandermolen, 2017). ARP4754A separates DAL into two: FDAL, function development phase, for aircraft functions and systems, and IDAL, item development phase, for electronic hardware and software items. The FDAL process assigns assurance cases ranging from A to E severity levels for functions which are allocated to items in a system. IDAL then assigns assurance levels for each item that is a part of the function in question, as part of electronic hardware or software. Detailed analyses examples may be read in ARP4754A (SAE International, 2010).

Also emerging are assurance cases toolsets, such as AdvoCATE, developed by Denney et al. (2012), offering an alternative to the manual labour of creating safety cases, and their linkage graphically with similar case scenarios, thus reducing time by providing available risk and hazard options, along with the assigning of requirements whether they be high or low level, in a seamless manner. Furthermore, fragments of the sources of documents for the assigned assurance cases can be linked to each correlated node, creating an easily exportable diagram, the software also works in coordination with AUTOCERT, a tool that evaluates modelling-and-design-stage flight and simulation code for safety violations, *via* clarifying it in a form of wording for the purpose of certification (Denney and Trac, 2008). "*Guidance on the Assurance of Machine Learning for Use in Autonomous Systems*" (AMLAS) is provided with a tool offered by the Institute for Safe Autonomy at the University of York. It focuses on the development of assurance cases for the use of ML in autonomous systems. The tool enables the addition of objects for each ML component, and its corresponding safety cases, while referring to AMLAS detailed means of compliance (Hawkins et al., 2021).

The limitations of ML algorithms require a scope to be identified within, and since they can handle non-deterministic

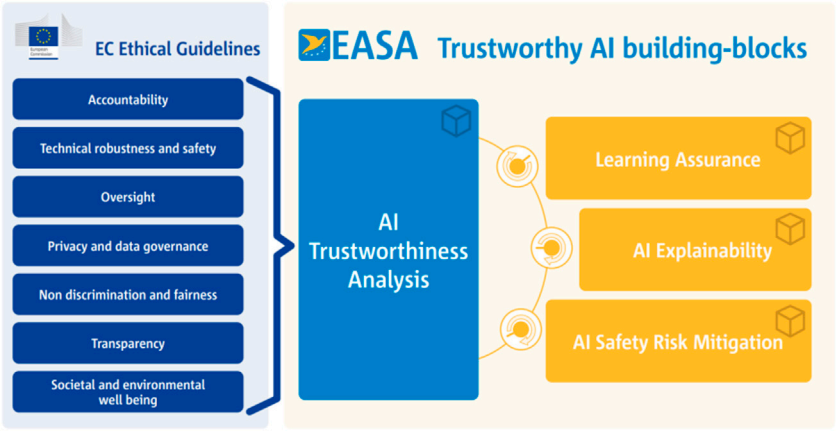

FIGURE 4
EASA Ethical Guidelines and AI building-blocks for trustworthiness. As per EASA's AI Roadmap, which has been formed with the goal of placing standards for ML applications in the EASA-related aerospace sector, seven ethical guidelines were placed for the operation of AI deemed trustworthy. They are subsequently managed by the four blocks in the figure, wherein: AI Trustworthiness Analysis supports the methodology on how to approach the seven guidelines in the use case of civil aviation. Learning Assurance develops the ideology of making sure that the ML algorithm in use is appropriate for the case at hand. AI Explainability focuses on the reason behind why the algorithm decides and its importance with respect to the end user in terms of delivering the desired output. AI Safety Risk Mitigation highlights the nature of how an AI black-box may require supervision due to its understandability and openness being limited in terms of decisions made.

behavioural scenarios, SOTIF (International Organization for Standardization, 2022), which was developed to address the new safety challenges that autonomous (and semi-autonomous) vehicle software developers are facing, may be used as part of the basis for certification application. Another challenge for certification is the constitution of a dataset, and whether it be sufficient for the required application and when compared to the function in operation. In the case of explainability, the lack of such a measure affects confidence in the model's learning capability. While ML is being implemented, the deployment of such a program would not be successful when supplied with a low-level set of tools for the inference. New practices in the aeronautics domain for certification encompass an initiative known as overarching properties. Here, assurance cases, which have been previously used in aeronautics and NN, may define themselves as the bridge between the need to comply with the overarching properties (which are intent, correctness and innocuity) and the quality possession of the product being considered by placing a strong argument. Artificial Intelligence in Aviation workgroups (such as SAE G-34/EUROCAE WG-114) are experimenting with the aforementioned new practices in order to produce guidance material for the standards being developed for ML in the aeronautical domain.

As per EASA's AI Roadmap, which has been formed with the goal of placing standards for ML applications in the EASA-related aerospace sector, seven ethical guidelines were placed for the operation of AI deemed trustworthy, as can be seen in Figure 4. They are subsequently managed by the four blocks in the figure, wherein:

1) AI Trustworthiness Analysis supports the methodology on how to approach the seven guidelines in the use case of civil aviation.
2) Learning Assurance develops the ideology of making sure that the ML algorithm in use is appropriate for the case at hand.
3) AI Explainabilityes focus on the reason behind why the algorithm decides and its importance with respect to the end user in terms of delivering the desired output.
4) AI Safety Risk Mitigation highlights the nature of how an AI black-box may require supervision due to its understandability and openness being limited in terms of decisions made.

5.1 Load profile uncertainties and risk management

Risk management for aircraft commences with following the standards placed by regulatory bodies, such as EASA for the European market, and the FAA in the US market. The next step in uncertainty management would be the categorisation of failure events and their probabilities, wherein there exists an inverse relation between the failure condition of an aircraft and its probability, and the resulting consequence on the aircraft and/or its occupants. Classifications by EASA are defined as Minor, Major, Hazardous, and Catastrophic, where they differ in their

definitions on levels of workload and crew impairment as well as passenger fatality probabilities. In addition, failure types must be stated. These include (Au et al., 2022):

1) Particular Risk: Failures impacting the system from the outside that could affect the system unfavourably.
2) Common Mode: Failure of a component as part of the system that contains a component identical to it dictates that the other component shall fail similarly.
3) Other Isolated Failures: The use of "undetected failures" on systems ensures that a failure not specified explicitly is encompassed in the placed standards and classifications, confirming the robustness of a system, must it pass said introduced diagnosis evaluation without any failure.

The LG operating environment consists of abrupt changes and the electrical sensors are susceptible to such changes and exterior elements. DO-160G covers avionics requirements in terms of environmental test conditions and procedures (Sweeney, 2015). For LG, these include waterproofness, shocks and vibrations, brake temperature, atmospheric conditions, lightning, electromagnetic emissions and susceptibility, and contaminants, such as dust and sand (Au et al., 2022).

A LG's components must be all tested against a "qualification test plan" to prove its usability in the harshest of environmental conditions (Au et al., 2022). This does not, however, include the component's entire life's combinations, resulting with the need to add experience from the industry and a "system development process" to add to the system's decisions in terms of verification for its use-case on-site.

Uncertainties resulting from the fatigue design process may be realised in:

1) Material properties of the components.
2) Geometry of the components.
3) Loads applied in-service onto the components.

The process in which components are manufactured, e.g. machining results with variations in the dimensions of the components, thereby directly affecting stress values of the components while in loading (Hoole, 2020). These variations may add up and amount to a failure as was the case with an aircraft nose landing gear strut examined by Barter et al. (1993), failing due to the formation of a fatigue-induced crack, as a result of an initial defect during manufacturing that grew in-service until the part was overloaded. As for material S-N curve datasets used for the stress-life approach, they naturally contain variability for each stress amplitude when compared to the cycles to failure. Furthermore, during the aircraft manoeuvres, the changes in magnitudes of the loads being applied, as well as when these loads occur, and the order of these occurrences, are factors to be considered for uncertainties. These are overcome *via* the use of safety factors within the stress-life analysis.

Loads imposed on the landing gear as part of the aircraft's life cycle can be divided into two types:

1) High and unexpected landing loads that occur during the aircraft's manoeuvres on the ground e.g. touchdown (Tao et al., 2009).
2) Loads that are repeated during the designated aircraft's trip and while on the ground, e.g. turning, braking, and taxiing, and being towed.

When extracting data, the order in which landing gear loads are applied may be inferred from load-time histories using open-source data, e.g. Flightradar24 such as in the case of Hoole (2020). He further categorises this variability in-service into the following: magnitude of the load, number of manoeuvres on-ground, and the order in which these manoeuvres occur. The latter two depend on factors related to the airport's structure and design, as well as the weather conditions on the day of service, in addition to the aircraft traffic at that point, changing the manoeuvres for an aircraft, also based on each airport's taxi operations locally, as well as gate locations.

For fatigue analysis, and with referral to EASA CS-25, (Hoole, 2020) mentions six methods that are commonly used for RUL conservatism:

1) Safety factors placed on components directly impacting their safe-life in order to indicate that they should be used ahead of assumed failure.
2) A safety factor to adjust the Miner's rule as part of the stress-life life approach discussed previously.
3) A safety factor placed on the application of stress on the components in order to assume that they are larger than their actual values.
4) An S-N curve reduction derived statistically.
5) A downwards shift on the S-N curve, causing the assumed stress required in order to reach failure for a certain number of cycles to be decreased.
6) A shift acting to the left on the S-N curve, indicating the assumption of a lower number of cycles needed in order for a part to fail under the specified load.

6 Proposed scenarios

As is the case with applications that would be deemed safety-critical, the following have requirements imposed upon them by learning assurance standards:

1) Datasets that are important for the development of the system.
2) The method and order in which this development takes place.
3) The behaviour of the system while both the development and operational stages take place.

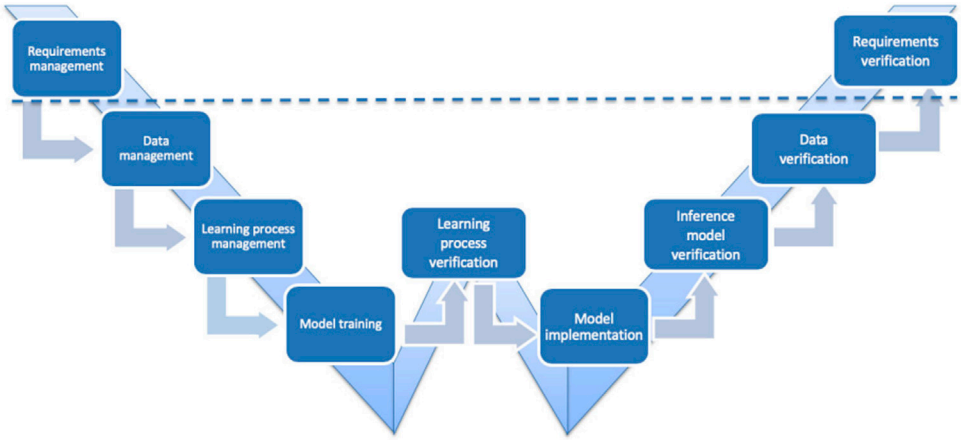

FIGURE 5
EASA Development life-cycle in the case of ML implementation. The process begins and ends with requirements management and verification while taking into reference ED-79A and ARP4754A documentation. In the midst of this life-cycle is data management where training, validation, and test datasets are collected and labelled as well as validated in comparison with the system requirements while sustaining a reliable amount of bias and variance within the data. Learning process management then prepares the model for training by selecting the appropriate algorithm for training, as well as the corresponding functions required for performance maintenance, while risk-checking the frameworks being used in the training environment. Model training merges the data management and process management steps to run the algorithm after which the data is validated using the validation dataset to evaluate the model's bias, variance, and quality of execution. Learning process verification uses the test dataset only. It evaluates the model's quality of execution, data bias, and data variance. It is not related in any way to validation, which is the last step of the model training stage. Model implementation moves the training model into one that may be run on the hardware targeted for the use case intended and any optimisations necessary are made in this stage in terms of computing requirements and necessities accommodated for. Inference model verification is the process in which the performance of the final inference model is evaluated through comparisons with the trained model. Additionally, compliance measures about software verifications are implemented according to ED-12C and DO-178C documentation.

TABLE 2 A use case advised by AIR6988. Shown is a predictive maintenance-involved system, where the ML-based system's functionality is summarized in the *Example* column, the *ID* column is "a unique identifier useful for reference in future work of the joint EUROCAE SAE G-34/WG-114 committee", the *Goal* details the ML-based system's functional operation is, *Inputs* counts the system's sensors and type of data, *Outputs* returns the message displayed as a result of the interaction between the ML-based system and the systems beneath, *Details* demonstrates the problems the use case targets, and *Integration* narrows down the system to be used with this AI application, whereas *Safety Concerns* raise severity level of the issues to be avoided for the completion of this use-case scenario.

Example	ID	Goal	Inputs	Outputs
Off-Board Predictive Maintenance	UC-SC322	Predict with high-specificity and high-accuracy an on-board failure with enough lead time to plan an optimized reaction	Low-level time-series sensor data collected and sent through a digital acquisition unit or data gateway	Failure message (can be EICAS/ECAMS message) + anticipated failure time + confidence of failure prediction
	Details		**Integration**	
	Combination of existing data cleansing/ETL + ML and other statistical methods to do big-data predictive maintenance		Aircraft owner, maintenance operation	
			Safety Concerns	
			Minimal, assuming existing procedures + instructions for parts handling are followed, and that scheduled maintenance is performed, as required	
Example	**ID**	**Goal**	**Inputs**	**Outputs**
On-Board Predictive Maintenance	UC-SC23	Predict with high-specificity and high-accuracy an on-board failure without having to send data to an off-board data center for analysis	Low-level time-series sensor data managed through high-bandwidth digital acquisition unit	EICAS/ECAMS message with predictive notation + anticipated failure time + confidence of failure prediction
	Details		**Integration**	
	Embedded NNs + other existing statistical methods (embedded) + on-board hardware for complex analytical processing		Aircraft owner, maintenance operation	
			Safety Concerns	
			Minimal, assuming existing procedures + instructions for parts handling are followed, and that scheduled maintenance is performed, as required	

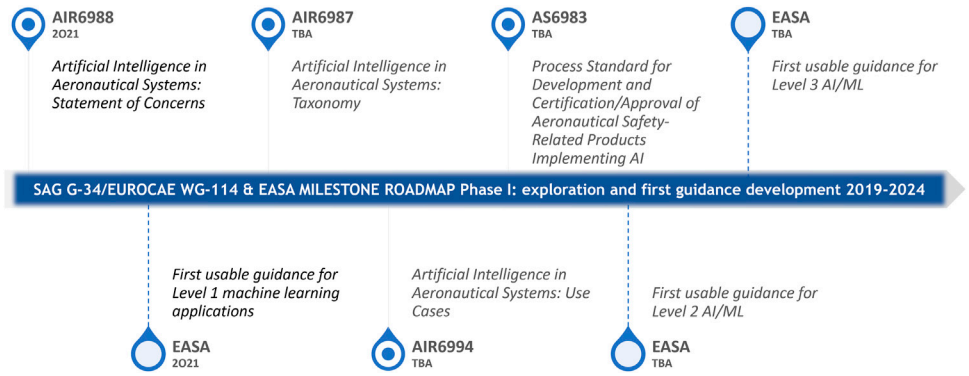

FIGURE 6
The standards for ML usage in aerospace are still being introduced, with a planned completion of frameworks by 2028 set by EASA. The first guidance development phase is to end by 2024, with the planned publication of ML use cases by the SAE workgroup, and the level 3 ML guidance paper by EASA. The SAE AIR6988 and EASA Level 1 ML Application guidance papers shown to the left of the timeline have been discussed in this paper.

(EASA and Daedalean AG, 2020) placed a layout for such a development life-cycle in the case of ML implementation, shown in Figure 5.

The process begins and ends with requirements management and verification while taking into reference ED-79A and ARP4754A documentation. In the midst of this life-cycle is data management where training, validation, and test datasets are collected and labelled as well as validated in comparison with the system requirements while sustaining a reliable amount of bias and variance within the data. Learning process management then prepares the model for training *via* selecting the appropriate algorithm for training as well as the corresponding functions required for performance maintenance, while risk-checking the frameworks being used in the training environment. Model training merges the data management and process management steps to run the algorithm after which the data is validated using the validation dataset in order to evaluate the model's bias, variance, and quality of execution. Learning process verification uses the test dataset only. It evaluates the model's quality of execution, data bias, and data variance. It is not related in any way to validation, which is the last step of the model training stage. Model implementation moves the training model into one that may be run on the hardware targeted for the use case intended and any optimisations necessary are made in this stage in terms of computing requirements and necessities accommodated for. Inference model verification is the process in which the performance of the final inference model is evaluated through comparisons with the trained model. Additionally, compliance measures with regard to software verifications are implemented according to ED-12C and DO-178C documentation (EASA and Daedalean AG, 2020).

The methodologies of certification discussed earlier may lead to a suggested use case advised by AIR6988 (SAE International, 2021a). The use case in Table 2 is an example of a predictive maintenance-involved system, where the ML-based system's functionality is summarized in the Example column, the ID column is "a unique identifier useful for reference in future work of the joint EUROCAE SAE G-34/WG-114 committee", the Goal details the ML-based system's functional operation is, Inputs counts the system's sensors and type of data, Outputs returns the message displayed as a result of the interaction between the ML-based system and the systems beneath, Details demonstrates the problems the use case targets, and Integration narrows down the system to be used with this AI application, whereas Safety Concerns raise severity level of the issues to be avoided for the completion of this use-case scenario.

7 A Roadmap and further research

Additional methods of data extraction for the use of ML, such as transfer learning, are currently being developed and seem promising for the benefit of this paper's direction. Transfer Learning is based on the development of a model's information for the use in another model performing similar tasks, while maintaining a low consumption of computationally-hungry processes and large amounts of data-requiring techniques. The aim is to keep the output and the task constant while changing the probability distributions required for the operation that leads to these tasks and outputs (EASA and Daedalean AG, 2020). Risks that may arise in correlation with resorting to such an approach include the necessity to verify the results of an empirical method-styled process, since transfer learning does include this approach. Another risk appears due to the requirement of transfer learning for a

"representative test set for the target function" (EASA and Daedalean AG, 2020), as a result of the source and target domain not being adequately related, causing an extra step and risk mitigation, trying to prevent what is known as a "negative transfer". Additional risk is due to uncertainty from using public dataset trained models as it may be more difficult to confirm that they comply with the learning assurance requirements. (Gardner et al., 2020) is an example of work in progress in this field, where the focus is on structures that have no data on their damage state obtained yet. The group uses data procured from an analogous structure to inference the damage on the former structure mentioned, using ML and non-destructive evaluation. The standards for ML usage in aerospace are still being introduced, with a planned completion of frameworks by 2028 set by EASA. Shown in Figure 6, the first guidance development phase is to end by 2024, with the planned publication of use cases by the SAE workgroup, and the level 3 ML guidance paper by EASA.

Author contributions

SP conceived the original idea and it was discussed with HM, whereby the main focus and paper's ideas were agreed upon with guidance taken from SP. The text of the paper is written by HM with sources for explainability and AI taken from colleagues, including those under the supervision of SP.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alain, G., and Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. Available at: http://arxiv.org/abs/1610.01644 (Accessed January 6, 2022).

Au, J., Reid, D., and Bill, A. (2022). "Challenges and opportunities of computer vision applications in aircraft landing gear," in 2022 IEEE Aerospace Conference (Big Sky, MT), March 5–12, 2022. doi:10.1109/aero53065.2022.9843684

Azodi, C. B., Tang, J., and Shiu, S. H. (2020). Opening the black box: Interpretable machine learning for geneticists. *Trends Genet.* 36, 442–455. doi:10.1016/j.tig.2020.03.005

Ball, D. L., Norwood, D. S., and TerMaath, S. C. (2006). "Joint strike fighter airframe durability and damage tolerance certification," in 47th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference, 01 May 2006–04 May 2006 (Newport, Rhode Island. doi:10.2514/6.2006-1867

Barandela, R., Valdovinos, R. M., Salvador Sánchez, J., and Ferri, F. J. (2004). "The imbalanced training sample problem: Under or over sampling?," in. *Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition* (Lisbon, Portugal: SSPR), 806. doi:10.1007/978-3-540-27868-9_88

Barter, S. A., Athiniotis, N., and Clark, G. (1992). "Cracking in an aircraft nose landing gear strut," in *Handbook of case histories in failure analysis*. Editor K. A. Esaklul (Ohio: ASM International), 11.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Blasch, E., Sung, J., Nguyen, T., Daniel, C. P., and Mason, A. P. (2019). "Artificial intelligence strategies for national security and safety standards," in *Aaai FSS-19: Artificial intelligence in government and public sector* (Arlington, Virginia. arXiv). doi:10.48550/arXiv.1911.05727

Cross, E., Sartor, P., Worden, K., and Southern, P. (2013). "Prediction of landing gear loads from flight test data using Gaussian process regression," in *Structural health monitoring 2013: A Roadmap to intelligent structures*. Editor F.-K. Chang (Lancaster, Pennsylvania: DEStech Publications), 1452

Cross, E., Sartor, P., Worden, K., and Southern, P. (2012). "Prediction of landing gear loads using machine learning techniques," in 6th European Workshop on Structural Health Monitoring (Dresden, Germany), July 3–6, 2012. Available at: http://www.ndt.net/?id=14124 (Accessed April 24, 2021).

EASA and Daedalean (2021). Concepts of design assurance for neural networks (CoDANN) II public extract. Available at: https://www.easa.europa.eu/en/document-library/general-publications/concepts-design-assurance-neural-networks-codann-ii (Accessed June 16, 2021).

EASA and Daedalean (2020). Concepts of design assurance for neural networks (CoDANN) public extract. Available at: https://www.easa.europa.eu/en/document-library/general-publications/concepts-design-assurance-neural-networks-codann (Accessed May 13, 2021).

DEEL Certification Workgroup (2021). White paper: Machine learning in certified systems (S079l03t00-005). Available at: https://hal.archives-ouvertes.fr/hal-03176080 (Accessed July 8, 2021).

Denney, E., Pai, G., and Pohl, J. (2012). "AdvoCATE: An assurance case automation toolset," in *Safecomp 2012: Computer safety, reliability, and security* (Magdeburg, Germany). 8–21. doi:10.1007/978-3-642-33675-1_2

Denney, E., and Trac, S. (2008). "A software safety certification tool for automatically generated guidance, navigation and control code," in 2008 IEEE Aerospace Conference (Big Sky, MT), 1–8 March 2008, Big Sky, Montana. doi:10.1109/AERO.2008.4526576

Dodge, J., Vera Liao, Q., Zhang, Y., Bellamy, R. K. E., and Dugan, C. (2019). "Explaining models: An empirical study of how explanations impact fairness judgment," in IUI '19: 24th International Conference on Intelligent User Interfaces, March 16–20, 2019, Marina del Ray, California (California), 275–285. doi:10.1145/3301275.3302310

Dziendzikowski, M., Kurnyta, A., Reymer, P., Kurdelski, M., Klysz, S., Leski, A., et al. (2021). Application of operational load monitoring system for fatigue estimation of main landing gear attachment frame of an aircraft. *Materials* 14, 6564. doi:10.3390/ma14216564

EASA (2020). Artificial intelligence Roadmap: A human-centric approach to AI in aviation. Available at: https://www.easa.europa.eu/en/newsroom-and-events/news/easa-artificial-intelligence-roadmap-10-published (Accessed May 13, 2021).

EASA (2021). EASA Concept paper: First usable guidance for level 1 machine learning applications: A deliverable of the EASA AI Roadmap. Available at: https://www.easa.europa.eu/en/newsroom-and-events/news/easa-releases-its-concept-paper-first-usable-guidance-level-1-machine-0 (Accessed February 22, 2022).

EUROCAE (2021). ER-022: Artificial intelligence in aeronautical systems: Statement of Concerns. Available at: https://eshop.eurocae.net/eurocae-documents-and-reports/er-022/# (Accessed December 8, 2021).

Fatemi, A., and Vangt, L. (1998). Cumulative fatigue damage and life prediction theories: A survey of the state of the art for homogeneous materials. *Int. J. Fatigue* 20, 9–34. doi:10.1016/S0142-1123(97)00081-9

Federal Aviation Administration (2005). AC 23-13a: Fatigue, fail-safe, and damage tolerance evaluation of metallic structure for normal, utility, acrobatic, and commuter category Airplanes. Available at: https://www.faa.gov/regulations_policies/advisory_circulars/index.cfm/go/document.information/documentid/22090 (Accessed October 4, 2020).

Fulton, R., and Vandermolen, R. (2017). *Airborne electronic hardware design assurance*. Boca Raton, FL: CRC Press.

Gardner, P., Fuentes, R., Dervilis, N., Mineo, C., Pierce, S. G., Cross, E. J., et al. (2020). Machine learning at the interface of structural health monitoring and non-destructive evaluation. *Phil. Trans. R. Soc. A* 378, 20190581. doi:10.1098/rsta.2019.0581

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Comput. Surv.* 51, 1–42. doi:10.1145/3236009

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3, 1157

Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. A. (2006). *Feature extraction: Foundations and applications*. Heidelberg: Springer-Verlag.

Hall, P., Ambati, S., and Phan, W. (2017). *Ideas on interpreting machine learning*. O'Reilly Media. Available at: https://www.oreilly.com/radar/ideas-on-interpreting-machine-learning/ (Accessed September 25, 2021).

Hawkins, R., Paterson, C., Picardi, C., Jia, Y., Calinescu, R., and Habli, I. (2021). Guidance on the assurance of machine learning in autonomous systems (AMLAS). Available at: https://www.york.ac.uk/assuring-autonomy/guidance/amlas/ (Accessed May 13, 2021).

Herlocker, J. L., Konstan, J. A., and Riedl, J. (2000). "Explaining collaborative filtering recommendations," in *CSCW00: Computer supported cooperative work*. (Pennsylvania, Philadelphi. doi:10.1145/358916.358995

Holmes, G., Sartor, P., Reed, S., Southern, P., Worden, K., and Cross, E. (2016). Prediction of landing gear loads using machine learning techniques. *Struct. Health Monit.* 15, 568–582. doi:10.1177/1475921716651809

Hoole, J. G. (2020). *Probabilistic fatigue methodology for aircraft landing gear*. Ph.D. Thesis. University of Bristol. Available at: https://hdl.handle.net/1983/8061165f-0a39-4532-bbbd-029e99286706 (Accessed June 3, 2021).

Hunt, S. R., and Hebden, I. G. (2001). Validation of the Eurofighter Typhoon structural health and usage monitoring system. *Smart Mat. Struct.* 10, 497–503. doi:10.1088/0964-1726/10/3/311

International Organization for Standardization (2022). ISO/PAS 21448:2022: Road vehicles-safety of the intended functionality. Available at: https://www.iso.org/standard/77490.html (Accessed August 20, 2022).

Irving, P. E., Strutt, J. E., Hudson, R. A., Allsop, K., and Strathern, M. (1999). The contribution of fatigue usage monitoring systems to life extension in safe life and damage tolerant designs. *Aeronaut. J.* 103, 589–595. doi:10.1017/S0001924000064228

Kaakai, F., Dmitriev, K., Adibhatla, S., Shreeder")Baskaya, E., Bezzecchi, E., et al. (2022). Toward a machine learning development lifecycle for product certification and approval in aviation. *SAE Int. J. Aerosp.* 15, 9. doi:10.4271/01-15-02-0009"

Kaur, H., Pannu, H. S., and Malhi, A. K. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Comput. Surv.* 52, 1–36. doi:10.1145/3343440

Khan, S., Phillips, P., Hockley, C., and Jennions, I. (2014a). No Fault Found events in maintenance engineering Part 2: Root causes, technical developments and future research. *Reliab. Eng. Syst. Saf.* 123, 196–208. doi:10.1016/j.ress.2013.10.013

Khan, S., Phillips, P., Jennions, I., and Hockley, C. (2014b). No Fault Found events in maintenance engineering Part 1: Current trends, implications and organizational practices. *Reliab. Eng. Syst. Saf.* 123, 183–195. doi:10.1016/j.ress.2013.11.003

Khan, S., and Yairi, T. (2018). A review on the application of deep learning in system health management. *Mech. Syst. Signal Process.* 107, 241–265. doi:10.1016/j.ymssp.2017.11.024

Kocielnik, R., Amershi, S., and Bennett, P. N. (2019). "Will you accept an imperfect AI? Exploring designs for adjusting end-user expectations of AI systems," in CHI '19: CHI Conference on Human Factors in Computing Systems, May 4–9, 2019 (Glasgow, Scotland. doi:10.1145/3290605.3300641)

Kursa, M. B., and Rudnicki, W. R. (2011). The all relevant feature selection using random forest. Available at: http://arxiv.org/abs/1106.5112 (Accessed January 7, 2022).

Ladda, V., and Struck, H. (1991). Operational loads on landing gear." in 71st Meeting of the AGARD Structures and Materials Panel (Povoa de Varzim, Portugal), 8th–12th October 1990. Available at: https://ntrl.ntis.gov/NTRL/dashboard/searchResults/titleDetail/N9128158.xhtml (Accessed November 18, 2020).

Le-The, Q.-V. (2016). *Application of multiaxial fatigue analysis methodologies for the improvement of the life prediction of landing gear fuse pins*. MSc Thesis. Carleton University. doi:10.22215/etd/2016-11572

Liu, Y., Loh, H. T., and Sun, A. (2009). Imbalanced text classification: A term weighting approach. *Expert Syst. Appl.* 36, 690–701. doi:10.1016/j.eswa.2007.10.042

Mitchell, T. M. (1997). *Machine learning*. New York, NY: McGraw-Hill.

Montavon, G., Samek, W., and Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.* 73, 1–15. doi:10.1016/j.dsp.2017.10.011

Murphy, C., Kaiser, G., and Arias, M. (2006). *A framework for quality assurance of machine learning applications*. CUCS-034-06. doi:10.7916/D8MP5B4B

Nguyen, G., Dlugolinsky, S., Bobák, M., Tran, V., López García, Á., Heredia, I., et al. (2019). Machine learning and deep learning frameworks and libraries for large-scale data mining: A survey. *Artif. Intell. Rev.* 52, 77–124. doi:10.1007/s10462-018-09679-z

Pascual, F. G., and Meeker, W. Q. (1999). Estimating fatigue curves with the random fatigue-limit model. *Technometrics* 41, 277–289. doi:10.1080/00401706.1999.10485925

Poggio, T., Banburski, A., and Liao, Q. (2020). Theoretical issues in deep networks. *Proc. Natl. Acad. Sci. U. S. A.* 117, 30039–30045. doi:10.1073/pnas.1907369117

Pu, P., and Chen, L. (2006). "Trust building with explanation interfaces," in IUI06: 11th International Conference on Intelligent User Interfaces, 29 January 2006–1 February 2006 (Sydney, Australia). doi:10.1145/1111449.1111475

RTCA (2012). *DO-178C software considerations in airborne systems and equipment certification*. Washington, DC: RTCA, Inc.

RTCA (2000). *DO-254: Design assurance guidance for airborne electronic hardware*. Washington, DC: RTCA, Inc.

Russell, S., and Norvig, P. (2022). *Artificial intelligence: A modern approach*. Global Edition 4th ed. Harlow, UK: Pearson.

Sae International (2021a). AIR6988: Artificial intelligence in aeronautical systems: Statement of Concerns. Available at: https://saemobilus.sae.org/content/AIR6988 (Accessed May 13, 2021).

Sae International (2010). ARP4754A: Guidelines for development of civil aircraft and systems. Available at: https://www.sae.org/standards/content/arp4754a (Accessed May 6, 2020).

Sae International (2021b). ARP6461A: Guidelines for implementation of structural health monitoring on fixed wing aircraft. Available at: https://saemobilus.sae.org/content/ARP6461A (Accessed February 3, 2022).

Schmidt, R. K. (2021). "*The design of aircraft landing gear*," (Warrendale, PA: SAE International), 858

Schwartz-Ziv, R., and Tishby, N. (2017). Opening the black box of deep neural networks via information. Available at: https://arxiv.org/abs/1703.00810 (Accessed February 28, 2022).

Smith-Renner, A., Fan, R., Birchfield, M., Wu, T., Boyd-Graber, J., Weld, D. S., et al. (2020). "No explainability without accountability: An empirical study of explanations and feedback in interactive ML," in CHI '20: CHI Conference on Human Factors in Computing Systems (Honolulu, USA, 1–13. doi:10.1145/3313831.3376624

Staszewski, W. J., and Boller, Chr. (2004). "Aircraft structural health and usage monitoring," in *Health monitoring of aerospace structures: Smart sensor technologies and signal processing* (Chichester: J. Wiley), 29.

Sweeney, D. L. (2015). "Understanding the role of RTCA DO-160 in the avionics certification process," in *Digital avionics handbook* (Boca Raton: Taylor &Francis Group), 194.

Tao, J. X., Smith, S., and Duff, A. (2009). The effect of overloading sequences on landing gear fatigue damage. *Int. J. Fatigue* 31, 1837–1847. doi:10.1016/j.ijfatigue.2009.03.012

Ten Zeldam, S. (2018). *Automated failure diagnosis in aviation maintenance using eXplainable artificial intelligence (XAI)*. MSc Thesis. University of Twente. Available at: https://purl.utwente.nl/essays/75381 (Accessed March 11, 2022).

Viaña, J., Ralescu, S., Ralescu, A., Cohen, K., and Kreinovich, V. (2022). Explainable fuzzy cluster-based regression algorithm with gradient descent learning. *Complex Eng. Syst.* 2, 8. doi:10.20517/ces.2022.14

Wanhill, R. J. H. (2018). "Fatigue requirements for aircraft structures," in *Aircraft sustainment and repair*, Editor R. Jones, A. Baker, N. Matthews, and V. Champagne (Oxford: Elsevier), 17–40. doi:10.1016/b978-0-08-100540-8.00002-9

Wong, J., Ryan, L., and Kim, I. Y. (2018). Design optimization of aircraft landing gear assembly under dynamic loading. *Struct. Multidiscipl. Optim.* 57, 1357–1375. doi:10.1007/s00158-017-1817-y

Xu, Z., Shen, D., Nie, T., and Kou, Y. (2020). A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data. *J. Biomed. Inf. X.* 107, 103465. doi:10.1016/j.jbi.2020.103465

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. Available at: http://arxiv.org/abs/1611.03530 (Accessed January 11, 2022).

Frontiers in
Astronomy and Space Sciences

Explores planetary science and extragalactic astronomy in all wavelengths

Advances the understanding of our universe - from planetary science to extragalactic astronomy, to high-energy and astroparticle physics.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact