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