Available online at www.sciencedirect.com

ScienceDirect

Advances in Space Research xxx (xxxx) xxx

ADVANCES IN SPACE RESEARCH (a COSPAR publication)

www.elsevier.com/locate/asr

Assessment of space weather modeling capabilities and transition to operations ☆

Maria M. Kuznetsova [a,*], Martin A. Reiss [a,af], Edmund Henley [b], Eric Adamson [c], Suzy Bingham [b], Mario M. Bisi [d], François-Xavier Bocquet [b], Laura Boucheron [e], Min-Yang Chou [a,f], Claudio Corti [a,g,af], Gian Luca Delzanno [h], Richard A. Fallows [d], Biagio Forte [i], Alexi Glover [j], Alexa Halford [k], Carl J. Henney [l], K.D. Leka [m], M.Leila Mays [a], Mark Miesch [c,ag], Karin Muglach [f,k], Mathew Owens [n], Nikolai Pogorelov [o], Lutz Rastätter [a], Michelangelo Romano [k], Pete Riley [p], Christine Verbeke [a,q], Jack Wang [a,f], Kathryn Whitman [r,s], Jia Yue [a,f], Yihua Zheng [a], Charles N. Arge [k], Sean Bruinsma [t], Federico Da Dalt [ah], Katherine Garcia-Sage [k], Joycelyn Jones [a], Maya Levisohn [a,u], Judit Palacios [ah], Tuija Pulkkinen [ae], Aaron Ridley [ae], Liutauras Rusaitis [a,f], Evangelia Samara [f,k], Ioanna Tsagouri [v], Jonathan Vigh [w], Robert Weigel [x], Chiu Wiegand [a], Clayton Allison [y], Ricky Egeland [z], Dinesha V. Hegde [o], Michael Kirk [k], Adam Kubaryk [c], Dibyendu Nandy [aa], Teresa Nieves-Chinchilla [k], Philip Quinn [y], Syed Raza [o], Elana Resnick [a,ab], Talwinder Singh [ac], Luke Stegeman [ad]

[a] *NASA Goddard Space Flight Center, Community Coordinated Modeling Center, Greenbelt, MD 20771, USA*
[b] *Met Office, FitzRoy Road, Exeter, Devon EX1 3PB, UK*
[c] *NOAA Space Weather Prediction Center, Boulder, CO, USA*
[d] *RAL Space, STFC Rutherford Appleton Laboratory, Harwell Campus, Didcot, Oxfordshire, UK*
[e] *Klipsch School of Electrical and Computer Engineering, New Mexico State University, Las Cruces, NM 88003, USA*
[f] *Catholic University of America, 620 Michigan Ave NE, Washington, DC 20064, USA*
[g] *University of Hawaii, Physics and Astronomy Department, 2505 Correa Rd, Honolulu, HI 96822, USA*
[h] *Los Alamos National Laboratory, Los Alamos, NM 87545, USA*
[i] *Department of Electronic and Electrical Engineering, University of Bath BA2 7AY Bath, UK*
[j] *European Space Operations Centre, European Space Agency, 64293 Darmstadt, Germany*
[k] *Heliophysics Science Division, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA*
[l] *Air Force Research Laboratory, Space Vehicles Directorate, Kirtland AFB, NM 87117, USA*
[m] *Colorado Research Associates Division, NorthWest Research Associates, Inc., 3380 Mitchell Lane, Boulder, CO 80303, USA*
[n] *Department of Meteorology, University of Reading, Reading RG6 7BE, UK*
[o] *The University of Alabama Huntsville, Huntsville, AL 35805, USA*
[p] *Predictive Science Inc., San Diego, CA 92121, USA*
[q] *CmPA/Department of Mathematics, KU Leuven, Celestijnenlaan 200 B, 3001 Leuven, Belgium*
[r] *University of Houston, 4800 Calhoun Rd, Houston, TX 77204, USA*
[s] *KBR, 601 Jefferson Street, Houston, TX 77002, USA*
[t] *Centre National D'etudes Spatiales, Space Geodesy Office, 18 Avenue E. Belin, 31401 Toulouse, France*
[u] *Telophase Corp., Lanham, MD 20706, USA*

☆ This article is part of a special issue entitled: 'COSPAR SW Roadmap 2022' published in Advances in Space Research.

* Corresponding author.
E-mail addresses: mmkuznets@gmail.com, Maria.M.Kuznetsova@nasa.gov (M.M. Kuznetsova).

https://doi.org/10.1016/j.asr.2026.03.027
0273-1177/

[v] National Observatory of Athens, Institute for Astronomy, Astrophysics, Space Applications and Remote Sensing, Metaxa and Vas. Pavlou, GR-15236 Penteli, Greece
[w] National Center for Atmospheric Research, Boulder, CO 80307, USA
[x] Space Weather Lab, Department of Physics and Astronomy, George Mason University, Fairfax, VA 22030, USA
[y] Leidos, Inc., 555 Forge River Rd, Ste 150A, Webster, TX 77598, USA
[z] NASA Johnson Space Center, 2101 E NASA Pkwy, Houston, TX 77058, USA
[aa] Center for Excellence in Space Sciences and Indian Institute of Science, Education and Research, Kolkata, Mohanpur 74125, India
[ab] ASRC Federal, 7000 Muirkirk Meadows Drive, Beltsville, MD 20705, USA
[ac] Department of Physics & Astronomy, Georgia State University, 25 Park Place, Atlanta, GA 30303, USA
[ad] Alan Levin Department of Mechanical and Nuclear Engineering, Kansas State University, Rathbone Hall, 1701B Platt St Suite 3002, Manhattan, KS 66502, USA
[ae] Department of Climate and Space Sciences and Engineering, University of Michigan, 2455 Hayward Street, Ann Arbor, MI 48109, USA
[af] Universities Space Research Association, 425 3rd Street SW, Suite 950, Washington, DC 20024, USA
[ag] Cooperative Institute for Research in Environmental Sciences, 1665 Central Campus Mall 216 UCB, Boulder, CO 80309, USA
[ah] Starion Deutschland GmbH for European Space Agency, 64293 Darmstadt, Germany

Received 8 May 2025; received in revised form 5 March 2026; accepted 7 March 2026

Abstract

Space environment models are imperative tools for space weather research and operational forecasting. Two major targets in the space weather enterprise are to advance predictive modeling capabilities and to promptly transition new models to space weather operations. Model assessments are essential to build credibility for potential users, while the approach to model validation depends on user priorities. Precise, transparent and up-to-date evaluations of strengths and limitations of space weather models provide vital guidance toward the progress in understanding of space environment phenomena. It also demonstrates potential of new models to improve state-of-the-art operational space weather forecasting. Therefore, a space weather capabilities assessment is one of the Overarching Activities of the COSPAR International Space Weather Action Teams (ISWAT) initiative. ISWAT Action Teams associated with ISWAT domain-focused (S – Space weather origins at the Sun, H – Heliosphere variability, G – coupled Geospace System, plus the newly established cross-domain Sun-to-Geospace – S2G) Clusters, provide pivotal progress in Assessment Overarching Activity. The ISWAT community-wide validation projects include both Modeling Challenges for historic time periods and pre-event real-time ensemble predictions. At the same time, operational space weather forecasting centers perform pre-operational and operational real-time continuous model validations and provide feedback to developers. This Roadmap paper presents a comprehensive overview of successes and lessons learned from validation projects and the transition to operations activities since the previous Space Weather Roadmap (Schrijver et al., 2015) and outlines recommendations for moving forward.

Keywords: Space weather; Models; Validation; Metrics; Uncertainty; Forecasting

1. Introduction

Space weather consists of complex physical processes characterised by phenomena originating deep within the Sun, traveling through the heliosphere, interacting with the coupled geospace system, and thus impacting critical technological infrastructure on our planet. Violent solar eruptions, primarily manifesting as solar flares and coronal mass ejections (CMEs), can affect the daily life of humans, by impacting, for example, communications, navigation, power supplies, national security, space travel, and more. In addition, protons and heavy ions can be accelerated by solar flares and shocks driven by CMEs to high energies producing so-called solar energetic particles (SEPs) that can expose astronauts and air travelers to dangerous levels of radiation. Consideration of space weather is crucial to the development and operations of space exploration missions, including the future Moon and Mars initiatives. In the past decade, the difficult task of understanding, quantifying and predicting these solar eruptions and their terrestrial impacts has become a strategic national and global priority.

The Sun-to-impact space weather flow has an intrinsically complex nature. Space weather processes cover different space domains with underlying physical phenomena that span a vast range of spatial and temporal scales. To solve the Sun-to-Impact Space Weather puzzle we need to assemble it part-by-part, simultaneously addressing problems focused on specific physical domains, assessing the accuracy and uncertainty of information passed between domains, and connecting all validated solutions from space weather origins on the Sun to their impacts.

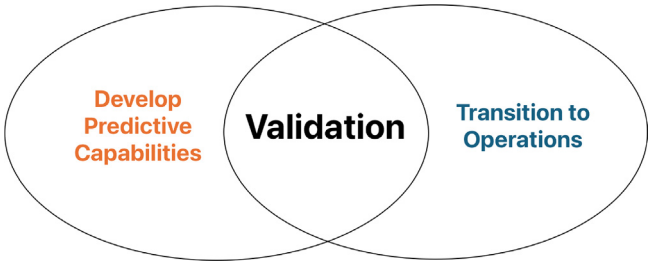

Fig. 1.1. Two major overlapping challenges in building predictive space weather capabilities involve: 1) developing new capabilities based on science, and 2) transitioning them to operations. Progress on both challenges can be supported through ongoing validation activities.

Advancing space weather predictive capabilities should address two major overlapping challenges (see Fig. 1.1):

1. First, any new predictive capabilities should be based on newly developed or improved scientific understanding and should demonstrate improvements to the current state-of-the-art in operational forecasting.
2. Second, the new capabilities should be transitioned to operational use (Research to Operations), and a feedback loop should be established (Operations to Research).

Progress on both challenges can only be achieved through ongoing validation activities during both the development and transition of new capabilities at multiple stages, as well as during ongoing operational use of such capabilities, once they are implemented.

Advances from the current state-of-the-art in space weather prediction through new scientific understanding, models, and data must be quantitatively assessed and tracked through community-coordinated assessment strategies. The importance of coordinated assessment has been emphasized in the original 2015–2025 Space Weather Roadmap commissioned by the Committee on Space Research (COSPAR) and the International Living With a Star (ILWS) (Schrijver et al., 2015). An International Forum on Space Weather Capabilities Assessment (IFSWCA[1]) was established during the International Community Coordinated Modeling Center (CCMC) – Living With a Star (LWS) Workshop in 2017 to address the need to quantify and track progress in space weather modeling and to establish internationally standardized metrics for meaningful model assessment. Outcomes of IFSWCA activities have been summarized in a special issue of the AGU 2018 Space Weather Journal titled "Space Weather Capabilities Assessment[2]" (see, Rastätter et al., 2019; Zheng et al., 2019; Choudhary, 2019; Scherliess et al., 2019; Verbeke et al., 2019; Kalafatoglu Eyiguler et al., 2019; Robinson et al., 2019; Bruinsma et al., 2018; Tsagouri et al., 2018; Roberts et al., 2018; Liemohn et al., 2018; Meier et al., 2018; Hill, 2018; MacNeice, 2018; MacNeice et al., 2018; Tobiska et al., 2018; Welling et al., 2018; Murray, 2018; Richardson et al., 2018; Shim et al., 2018; Wintoft and Wik, 2018). The IFSWCA is a precursor to the COSPAR International Space Weather Action Teams (ISWAT[3]) initiative established in 2018. All active IFSWCA working teams were migrated to ISWAT, and most of the assessment topics introduced by the IFSWCA were adopted by ISWAT action teams. A history of ISWAT and overview of ISWAT activities are included in this special issue (Kuznetsova et al., 2026).

ISWAT action teams work via self-guided topical collaborations on different aspects of space weather. As shown in Figure 1.2, ISWAT action teams are organised into Clusters by domain, phenomena, or overarching activity. Space Weather Capabilities Assessment is an ISWAT Overarching Activity involving teams across all domains. Results achieved by ISWAT action teams focused on the assessment of space weather capabilities are published in the COSPAR Space Weather Roadmap Special Issue 1 "Science and Applications" (Bisi and Shea, 2023). Highlights of validation studies from Roadmap Special Issue 1 are presented in §4.2.

There is a clear need to enhance operational space weather forecasting to achieve an outcome similar to modern terrestrial weather forecasting that provides predictions up to several days in advance (Bauer et al., 2015). For clarity, we emphasize here that the term "operations" (and variants thereof appearing throughout this text) generally refers to the specific role of space weather forecast service providers (e.g., NOAA's Space Weather Prediction Center in the USA, and equivalents in other countries) and the capabilities which they maintain and critically depend on, 24 h per day, in their missions supporting safety and prosperity (see detailed description in Appendix B). Transitioning scientific advancements into operational use, is commonly referred to as R2O (Research to Operations). R2O is a multi-step, multifaceted process that begins in the initial stages of model development and includes assessment through all phases of transition to operations. Demonstration on how new understanding, models, and data can improve operational space weather predictions is a crucial part of the R2O process. Equally important is the feedback from Operations to Research (O2R), which ensures that experience from forecasters and operational users reaches model developers and closes the Research-to-Operations-to-Research (R2O2R) loop. A successful R2O2R process (frequently referred to as R2O2R platforms, frameworks, or pipelines) requires collaboration among all space weather participants, including domain scientists, modelers, software developers, system engineers, instrument designers, space weather service providers, and expert users.

This paper examines different components of both challenges (see Fig. 1.1) related to model validation. These

[1] https://ccmc.gsfc.nasa.gov/iswat/IFSWCA/.

[2] https://agupubs.onlinelibrary.wiley.com/doi/toc/10.1002/(ISSN)1542-7390.SW_CASS.

[3] https://iswat-cospar.org.

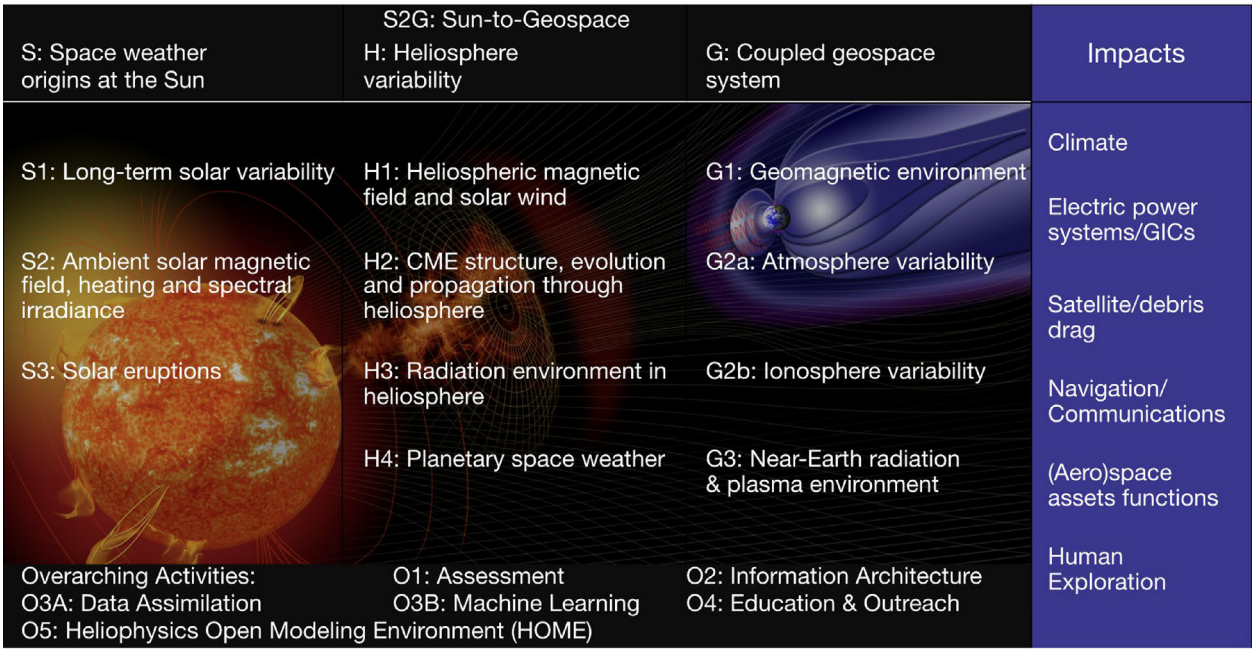

Fig. 1.2. Organization of the COSPAR ISWAT Initiative as of December 2025 (taken from the COSPAR ISWAT home page). The domains are listed in the top row and each domain consists of numbered "Clusters," which are research focus areas within each domain. Each of almost 100 teams has a unique identifier composed of the cluster number and the team number (for instance, H1-01). More information on ISWAT and their individual teams can be found online[3] and in Kuznetsova et al. (2026, this issue).

components include the following topics: intended uses of validation (§2), elements of validation studies (§3), validation across individual domains (§4), how to validate modeling of physical phenomena (§5), demonstration of operational potential through pre-event forecasting activities (§6), tracking frameworks (§7), R2O2R pipelines (§8), real-time validation strategies (§9). In §10, we identify gaps and provide recommendations to address them.

Appendix A lists acronyms used throughout, and Appendix B explains terminology such as "space weather enterprise", "quality assessment", "ensemble", and "users", as well as the definitions of "validation" and "verification" which swap meanings between the terrestrial and space weather conventions. Where the terrestrial weather convention is unavoidably used in the text, the space weather convention is indicated in brackets thereafter. Appendix C provides tables defining levels for the various model maturity tracking frameworks used across the enterprise.

2. Model validation and intended use

Model validation is a process of quantifying and building credibility in numerical models from the perspective of the intended uses of the model. As shown in Fig. 2.1, scientists (researchers) primarily want the quantities derived from models to be scientifically meaningful, while space weather forecasters (operators) want them to be reliable, robust, and produced quickly enough to be useful in operational forecasting. Therefore, the goals of validation studies are divided into quantitatively assessing:

1. Model ability to reproduce observed signatures and physical phenomena (science); and
2. Model usefulness for space weather operations and addressing space weather user needs (applications).

The science of space weather is multi-disciplinary. It comprises physics, chemistry and mathematics, including statistics, programming, and data analysis. Scientists use models to advance the understanding of physical processes and study physical phenomena. The purpose of science validation is to demonstrate what physical phenomena are consistently modeled correctly and what are consistently modeled wrongly. Evaluating the science quality and applicability of modeling approaches is critical for ensuring that conclusions derived from model outputs are scientifically sound and simulated features are not caused by numerical artifacts. Validating the science quality of physics-based models should include testing and verifying consistency of simulation results with used modeling approximation, initial assumptions, and known analytical solutions, where possible. It is extremely important to analyze and determine why different models, versions of the same model, or configurations of the same model give different results. In most cases, a standard suite of idealized boundary conditions would be instrumental for evaluating the science quality of models.

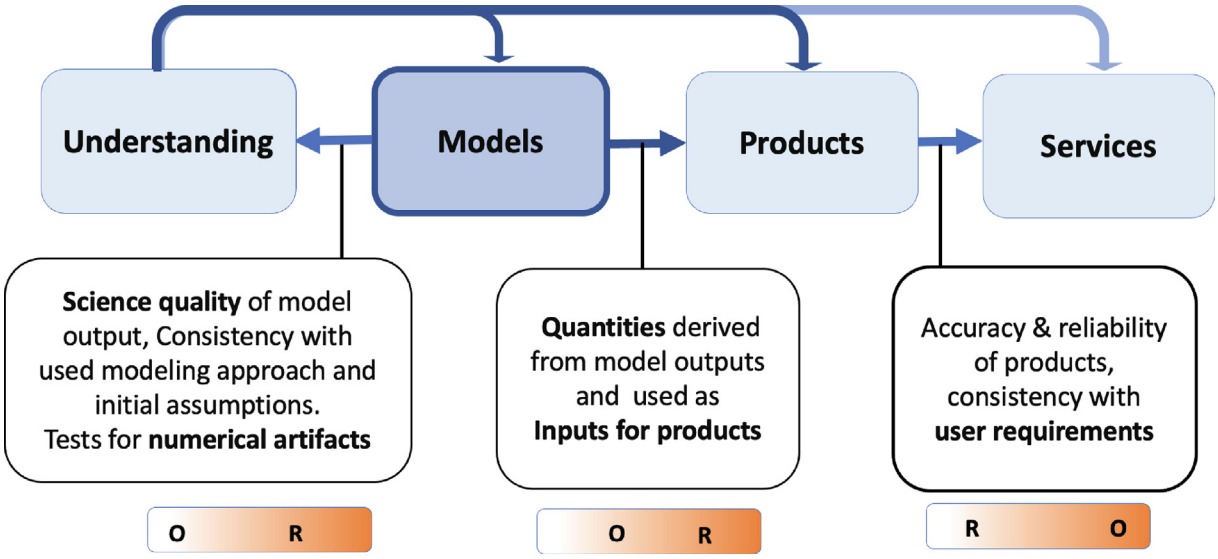

Fig. 2.1. What needs to be validated? Models are used to advance understanding of phenomena and processes and to produce inputs to products and services tailored to address user requirements. Improved understanding triggers development of new models with potential to improve products and services. The white-orange colormap illustrates low (white) and high (orange) levels of interest by researchers (R) and operators (O). Here 'operators' refers primarily to operational agency forecasters, but can also represent expert users.

In scientific validation, identifying and quantifying physical phenomena in data and models, as well as defining an approach to measure model-data and model-model agreements, is a very challenging research project (see §5 for examples). These challenges include sensitivity of model outputs to data that drives them and internal model parameters and assumptions. Even if a scientific model is self-consistent and reproduces available analytical solutions, it still may be unacceptable from the viewpoint of its predictive capabilities.

Unfortunately, so far, there has been only limited engagement from users in advancing scientific understanding, while scientists have not always been guided by operational usage of their models. To achieve progress, there should be mutual exchange, with both parties sharing responsibility. Scientists and the scientific enterprise gain from involvement in the successful transition of models to operations. These successes demonstrate to stakeholders the societal benefits of basic and applied research and allow the body of scientific knowledge to influence user decisions. End-users, similarly, benefit from understanding both the limitations and potential of the models their success depends on. Close collaboration, especially focusing on comprehensive and standardized validation, can help close this gap. Meaningful validation at every stage of the development of predictive capabilities ensures that our operational forecasting models are not only of a high scientific standard but also useful, usable, and used! Key to this is that validation criteria, including metrics, reflect end-user needs and are not only tailored to highlight new scientific developments.

3. Elements of validation studies

This section discusses elements of coordinated validation studies to track improvement of space weather modeling capabilities, including:

- Parameters for validation derived from model outputs and observations
- High-quality observational data
- Algorithms for model-data comparisons (metrics)
- Reference models used as benchmarks
- Repositories and metadata for observational and simulated time series
- Tools for validation

3.1. Selecting parameters for validation

Validation projects for space weather models underpinning applications may see substantial benefit from a focus on key parameters that serve as inputs for multiple space weather products and/or are used by forecasters to issue reports, alerts and warnings that facilitate decision making. These parameters (Essential Space Environment Quantities – ESEQs) can also serve as inputs for models of impacts producing impact quantities (see Fig. 3.1).

Impact-driven validations are a high priority in assessing products provided to end users. However, impact information may not be openly accessible to enable validation without a dedicated agreement between the service developer/provider and the end user. Quantifying space weather

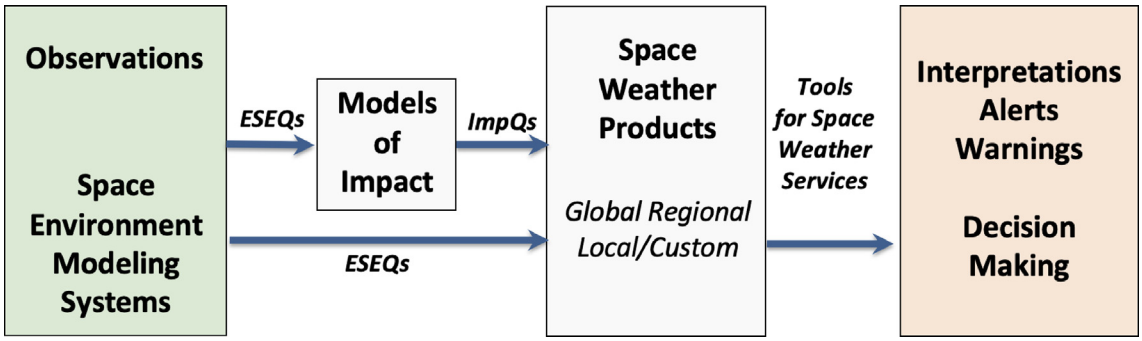

Fig. 3.1. Example flow of space weather information. Essential Space Environment Quantities (ESEQs) and Impact Quantities (ImpQs) are derived from observations and nowcasting / forecasting modeling systems. ESEQs serve as inputs for models of impacts and various space environment products and applications, used by service providers (forecasters, analysts) as a basis for interpretations, alerts, and warnings. To aid decision-making, global and regional products are tailored for user groups, while local and custom products are tailored for specific expert users.

impacts accurately fundamentally requires addressing the tensions involved with sensitivities in commercial or security settings (e.g., Eastwood et al., 2017). Existing space weather impacts quantities and related datasets are opportunistic data, that is, information not originally collected for space weather science but that can nevertheless be used for space weather validation. These opportunistic datasets are valuable due their proximity to operational needs, and they often have the advantage of extensive temporal and spatial coverage. Coupling of the space environment with models of impact introduces additional sources of uncertainties for validation of space environment models.

Approaches to move forward include:

- Collaborating with the broader space weather communities and decision-makers to improve the availability and quality of impact information.
- Identifying ESEQs closely linked to impacts that can be derived from simulation outputs and accessible observations.
- Partnership between scientific and commercial sectors in sharing proprietary data.

Identifying ESEQs indicative of impacts is a challenging task that involves communication with users and conducting quantity-impact correlation studies. In addition to space environment conditions, impacts can be influenced by other factors. For example, the time derivative of the ground magnetic field (referred to as "dB/dt") can serve as an indicator for the level of geomagnetically induced electric field that is the primary physical quantity driving Geomagnetically Induced Currents (GICs). Consequently, although numerous additional complexities such as ground conductivity, conductor system configuration, and engineering details such as high-voltage power transformer design are critical for a detailed threat assessment, dB/dt can indicate a potential GIC hazard (Viljanen et al., 2001; Pulkkinen et al., 2013).

Table 3.1 shows examples of user groups, impacts and corresponding ESEQs linked to impacts. More details on identifying ESEQs for specific impacts can be found in domain-focused reviews in this Special Issue including Boyd et al. (2026), Bruinsma et al. (2026), Tsagouri et al. (2026a,b), Jun et al. (2026), Zheng et al. (2026), Minow et al. (2026), Opgenoorth et al. (2026), Guo et al. (2026), and Ishii et al. (2026a), this issue.

Impacts of the near-Earth radiation environment on satellite functionality and humans in space include surface charging, internal charging, total dose, single event effects, radiation effects for aviation, and radiation impacts on crewed missions. Efforts have been made by the Near-Earth Space Radiation and Plasma Environment ISWAT cluster (G3[4]) to define measurable ESEQs that are directly related to corresponding impact-driven anomalies (Zheng et al., 2026, Table 11, this issue).

Another important group of ESEQs are parameters that are passed between domains and serve as inputs (drivers) for space environment modeling systems. These ESEQs are also used by forecasters in alerts and warnings that facilitate decision making. Examples of drivers are listed in Table 3.2.

Model validation based on ESEQs is the first step in demonstrating the operational potential of a model. Note that Tables 3.1 and 3.2 are not intended to be an all-inclusive list of space weather information used in space weather operations. The focus is on ESEQs that can be derived from modern space weather models to set the scene for discussion on community validation efforts to demonstrate potential of new modeling capabilities.

3.2. Addressing the need for validation-ready observational datasets

An important element of successful space weather model validation is establishing validation-ready observational datasets. Observational datasets agreed upon by the community can be selected based on important geomagnetic events or defined time periods with start and end dates, providing a common ground for comparisons. Future validation studies gain significantly in credibility and usefulness by using such community-agreed datasets instead of individ-

[4] https://iswat-cospar.org/g3.

Table 3.1
Examples of ImpQs and ESEQs indicative to impacts.

User Groups	**Impact Quantities (ImpQs)**	**Essential Space Environment Quantities (ESEQs)**	**References and Relevant Sections in this paper**
Electric power systems	Geomagnetically Induced Current (GICs), Geoelectric field at selected locations/regions)	Ground magnetic perturbationsDelta-B, dB/dt, Geomagnetic indices (Kp, Ap, Dst, Hpo)	Viljanen et al. (2001); Pulkkinen et al. (2013); Opgenoorth et al. (2026) 6.7, 9.1
Navigation, Communications	GNSS positioning quality, signal availability. Difference between predicted and actual radio signal propagation. Degradation of radio wave propagation in different frequency ranges. Maximum useable frequency (MUF), Lowest usable frequency (LUF)	TEC (Total Electron Content), STEC (Slant TEC), ROT (Rate of TEC), ROTI (ROT Index), Plasma density along high frequency (HF) signal propagation, Critical frequency foF2, Peak density NmF2 of F2 layer, Height of peak density hmF2, Phase and amplitude of scintillations, Scintillation index S4, Ionospheric absorption	Tsagouri et al. (2026a,b); Ishii et al. (2026a), this issue; Forte et al. (2026); Fiori et al. (2022) 3.2.4, 3.2.5, 3.2.6, 4.1.7, 9.1, 9.4
Satellite Drag	Difference between predicted and actual satellite location	Neutral density along satellite orbit.*	Bruinsma et al. (2026) 3.2.7, 4.1.6, 9.1
(Aero)space Assets functions **	Surface charging Internal charging Single Event Effects Total Dose in Orbit Dose rate in aircrafts (D-index)	> 10 keV electron flux > 1 MeV, > 2 MeV electron fluence (over 24-hour, or 48-hour,or 72-hour periods) > 30 MeV proton flux 30–50 MeV proton fluence, >1.5 MeV electron fluence(over mission duration) > 300 MeV proton flux	Jun et al. (2026); Zheng et al. (2026); Minow et al. (2026); Boyd et al. (2026) 3.2.3, 4.1.8
Human Exploration ***	Radiation exposure	> 10 MeV, > 100 MeV, > 500 Mev solar proton flux/ fluence ****	Whitman et al. (2023); Guo et al. (2026) 4.1.5, 8.2, 9.1, 9.3

* Requires a debiasing procedure to take differences in aerodynamic coefficient models into account.

** Given the complexity of various impacts on aerospace assets, we focus on the most relevant ESEQs that can be measured/derived from both observations and models, acknowledging some simplifications. Ideally, accurate assessments require energy (or linear-energy-transfer spectra for single event effects – SEEs) across the appropriate energy ranges for each impact. These effects occur on different timescales: surface charging within seconds to minutes, SEEs in microseconds to milliseconds, internal charging over hours, and total dose across the mission duration. Fluence is flux integrated over time.

*** Solar proton flux is also important for ionosphere absorption especially in polar regions.

**** For assessing human radiation exposure, both flux and fluence are relevant. Fluence (integrated over time) determines the total dose that was received, while flux (per unit time) determines the dose rate. Total dose is the main driver of long-term risks like cancer, and dose rate is critical for short-term effects and operational safety.

Table 3.2
Examples of ESEQs that serve as drivers for space environment modeling systems.

ESEQs / Drivers	Space Weather Significance	Relevant Sections
Solar X-ray flux and solar flare forecasts	Inputs for ionospheric absorption models. Important for CME and SEP forecasts	4.1.4, 6.2, 9.2
Solar indices – F10.7: solar radio flux at 10.7 cm – F30: solar radio flux at 30 cm – S10: integrated 26–34 nm EUV irradiance – M10: derived from FUV at 160 nm – Mg II: core-to-wing irradiance ratio	Used as proxies for EUV (extreme ultraviolet) and FUV (far ultraviolet) solar irradiance in thermosphere models	6.4
Solar wind and interplanetary magnetic field (IMF) parameters at Lagrange Point 1 (L1) and other locations	Inputs to geospace models	3.2.2, 4.1.1,6.5, 6.6, 9.1
Coronal Mass Ejections (CME) kinematic parameters (e.g., speed, direction)	Inputs to CME propagation models and Solar Energetic Particles (SEP) models	4.1.2, 4.1.3, 9.2, 9.3
CME Time of Arrival (ToA) at L1	Timing of storm onset, timing of impacts	3.2.2, 4.1.2, 4.1.3, 6.1, 9.2, 9.2, 9.3
Geomagnetic activity indices – Kp: quasi-logarithmic 3-hour geomagnetic activity index (0–9 range) – Ap: linear (nanotesla) daily index derived from Kp (0–400 range) – Dst: Disturbance storm time index – AU, AL, AE: Auroral indices – High resolution counterparts to traditional indices*	Inputs to ionosphere-thermosphere models Characterize geomagnetic storm strength	6.7, 9.1, 9.4

* The newer indices such as open-ended Hpo[a] (Hp30/Hp60) and SuperMAG[b] indices (SYM-H, SME=SMU-SML) use data from numerous observatories and provide faster, more spatially detailed views of magnetic disturbances than older, slower indices like Kp (based on the data of 13 mid-latitude geomagnetic observatories), Dst (based on four near-equatorial observatories), and AE = AU-AL (based on 12–13 magnetometers in the northern auroral zone). [a]https://kp.gfz.de/en/hp30-hp60. [b] https://supermag.jhuapl.edu/info/.

ually selected subsets, which make comparisons to past and future studies challenging, or even impossible. Validation-ready datasets ensure a level playing field and support rigorous model validation, particularly for data-driven predictive models of space weather phenomena, such as the ambient solar wind, coronal mass ejections (CMEs), and solar flares. For validation based on historic time periods (see §4), it is important that the observational dataset is accurately selected and prepared with quality control applied. Developing and curating such datasets is complex and requires special consideration. Note that for real-time validations data are not necessarily accurately selected and prepared with quality control. The following sections provide examples of efforts to create datasets prepared for validation.

3.2.1. Coronal hole dataset

Reiss et al. (2024) discuss the efforts of the Coronal Hole Boundary Working Team[5] (S2-01) in the "Ambient solar magnetic field, heating, and spectral irradiance" S2 cluster within the COSPAR ISWAT initiative. Coronal holes are dark regions in extreme ultraviolet (EUV) observations of the Sun due to their lower density and temperature compared to the surrounding coronal plasma. They play an important role in space weather research because they are associated with the origin of fast solar wind streams that continually interact with planetary environments, causing geomagnetic disturbances (Krieger et al., 1973; Tsurutani et al., 2006). Furthermore, they influence the ambient solar wind conditions through which interplanetary coronal mass ejections (ICMEs), the driver of the most severe forms of space weather, propagate. These large-scale structures in ambient solar wind can distort and deflect ICMEs, affecting their geomagnetic effects (Riley et al., 1997; Odstrčil and Pizzo, 1999b; Zhou and Feng, 2017).

Comparative studies show that there is a need to enhance our capabilities for automatically locating and assessing coronal holes (Linker et al., 2021; Reiss et al., 2021a). The S2-01 team collaborated to create a dataset of selected Solar Dynamics Observatory (SDO) observations to test future solar feature detection schemes. The authors identified the need for a standardized dataset for developing, testing, and improving automated coronal hole detection schemes. This community dataset comprises 29 manually selected SDO images, accompanied by the coronal hole detection results from widely-applied automated detection schemes. The SDO images were observed between 2014 and 2019, spanning the time from maximum solar activity to the subsequent minimum. The dataset includes all seven EUV wavebands ranging from 9.4 nm to 33.5 nm captured by the Atmospheric Imaging Assembly (AIA) instrument and the line-of-sight measurements of the photospheric magnetic field from the Helioseismic and Magnetic Imager (HMI) instrument for each date.

The images capture a wide range of appearances of coronal holes and other solar features, and were selected to chal-

[5] https://iswat-cospar.org/s2-01.

lenge automated detection schemes. Each image in the dataset is accompanied by coronal hole labels manually assigned by an experienced observer (see, Reiss et al., 2024). These labels serve as the ground truth for evaluating the accuracy of automated detection schemes in terms of an event-based validation. After performing basic preprocessing steps on level 1.0 data from the SDO data platform, the data was provided as input to the 14 participating research teams. All the collected coronal hole detection results from these research teams are publicly available for future assessment and comparison as part of this community dataset.[6]

3.2.2. Interplanetary scintillation/inner heliosphere data

For the heliosphere, there are several radio techniques available for gleaning the information needed for possible identification of ESEQs in the future. The main foci, currently, are on a scientifically well-validated technique of Interplanetary Scintillation (IPS) (Clarke, 1964; Hewish et al., 1964; Fallows et al., 2023 and references therein) and that of Faraday rotation. IPS can provide velocity information (e.g., Bisi et al., 2010), density proxies (e.g., Gapper et al., 1982), and inclinations of magnetic-field orientation (Fallows et al., 2023) but not magnetic-field direction (as yet). For the latter part as the second of the two main foci, Faraday rotation; this is being explored in coronal regions (e.g., Kooi et al., 2022) at high radio frequencies, with modelling across the heliosphere (e.g., Jensen et al., 2010), and with experiments underway to test Faraday rotation capabilities in the heliosphere using the LOw Frequency ARray (LOFAR, e.g., Bisi et al., 2016) and Murchison Widefield Array (MWA) radio-telescope systems.

On the IPS side, where verification and validation work beyond just the science is being undertaken, the first major step was to unify a usable dataset and data format for the IPS data taken from multiple IPS-capable radio observatories around the globe. This was accomplished through: 1) setting up of the Worldwide IPS Stations (WIPSS) Network (Bisi et al., 2016; 2017); and 2) implementing the IPS Common Data Format (IPSCDFv1.0 and IPSCDFv1.1) which all IPS data providers have agreed to adhere to when sharing their data.

The end IPS data products include the essential velocity determinations and a normalised scintillation level, known as g-level, that is used as a proxy for density. To demonstrate the value of radio data to improve heliospheric modeling these IPS data are included in the University of California, San Diego, three-dimensional (3D) Heliospheric Tomography model (e.g., Jackson et al., 2023) that is used to drive the 3D Magnetohydrodynamics (MHD) heliosphere Enlil model (for more details on the IPS-Enlil see Jackson et al., 2015; 2020 and references therein). A separate avenue of using IPS data with MHD simulations is described in Iwai et al. (2023). It is from determining improvements in tomographic and MHD forecast simulations that a true assessment of the IPS capability can be ascertained. IPS also has the added advantage that it is capable of detecting all heliospheric structures, including CMEs, so it does not need the addition of a CME to be introduced when using MHD simulations with IPS data driving them at the inner boundaries.

An alternative approach, which has not yet been implemented, uses observations of IPS as "ground truth" measurements against which space-weather forecast ensemble runs can be compared and pruned to improve the forecast. This approach would integrate IPS lines of sight through the modelled heliospheric structure to produce modelled IPS results for direct comparison with raw observations. A similar approach when modelling a CME that occurred in May 2005 (Bisi et al., 2010) used 3-D tomography to constrain the location of the CME in an IPS line of sight, improving the modelling of single IPS measurement to determine the CME speed, and speed and direction of both perturbed and unperturbed fast solar wind streams in the line of sight (Breen et al., 2008), demonstrating the potential value of this approach.

Thus, the IPS observing technique (as applied to the heliosphere) holds great potential to establish a systematic link between observations across the inner heliosphere and the response of the magnetosphere-ionosphere system to different space-weather conditions. The technique now benefits from the possibility of using a higher number of radio sources which can be observed from different radio-telescope systems around the world. This capability helps refine tomographic reconstruction of the plasma density and velocity of structures in the heliosphere (and MHD simulations also), estimation of which is crucial to the improvement of space-weather forecasts. As part of the UK Research and Innovation (UKRI) Natural Environment Research Council (NERC) Radio Investigations for Space Environment Research (RISER) project (e.g., Bisi et al., 2023)[7] this validation work is being undertaken via a R2O2R process in conjunction with the UK Met Office. A future challenge related to this type of approach is whether a similar attempt can be expanded to estimate both the magnitude and direction of the magnetic field in the heliosphere.

3.2.3. Near-earth space radiation and plasma environment data

For the near-Earth space radiation and plasma environment, data scarcity remains one challenge. In general, we need more flux measurements for different energy ranges, for different species, and at different orbits. This is particularly true of highly elliptical orbits which are becoming increasingly important for Earth observation, as these

[6] https://figshare.com/articles/dataset/Coronal_Hole_Detection_Comparison_Dataset/23997993/3.

[7] https://gotw.nerc.ac.uk/list_full.asp?pcode=NE%https://gotw.nerc.ac.uk/list_full.asp?pcode=NE%2FX019004%2F1.

satellites travel through the Earth's radiation belts. For radiation effects at aviation altitudes, we need high quality dose measurements (e.g., Meier et al., 2018), especially on north – south and cross-polar flight paths, and especially during SEP events. This is particularly relevant now, as the International Civil Aviation Organization (ICAO) has recently acknowledged the need for operational space weather services for aviation with impacts to both humans and avionics due to radiation (e.g., ICAO, 2019).

While performing model validation under the ISWAT internal charging effort, we have found that, even with the same instrument on the same mission, data from different user groups exhibit some subtle differences (different interpolation schemes could be one of the contributing factors). Cross-calibration of similar measurements (e.g., energetic electron data from multiple satellites) over a long period of time from different missions/sources is often time-consuming and nontrivial. Results from such efforts should be shared as community resources. Turning measurements into useful and meaningful quality data (whether for science or space weather operations) is also challenging. A prime example is mutual contamination of energetic electrons and protons during interesting space weather intervals. Preparation of data (that are relevant for aurora, ring current, radiation belts, and SEP studies) are often cleaned by different groups in different ways, sometimes incompatible with each other.

For impact- or anomaly-driven validation, the situation is even worse due to restricted access to critical information, often caused by inherent limitations (e.g., spacecraft lacking diagnostic information) or commercial sensitivities. As a result, validation studies require considerable effort, including remote root-cause analyses that attribute reaction wheel failures to space weather-induced discharges rather than lubricant issues (e.g., Bialke and Hansell, 2017; Bialke, 2018), rigorous statistical correlations between space environment conditions and publicly reported anomalies on government-funded satellites like GOES (e.g., Kress et al., 2024; Rodriguez et al., 2025), and creative approaches for commercial fleets, such as inferring anomalies from financial filings – a method that remains contentious due to the risk of causal misattribution (Choi et al., 2011; Mazur and O'Brien, 2012).

Opportunities exist to broaden the effects considered in impact-based validation. Satellite lubricants, for example, are susceptible to ionizing radiation (e.g., Fusaro, 1994; Jones and Jansen, 2006; Macha et al., 2025), yet remain largely unexplored compared to well-documented radiation impacts on electronics and solar cells (e.g., Hands et al., 2018). Investigating lubricant degradation—through cumulative exposure to radiation belt enhancements, galactic cosmic rays, or SEP events—could enable short- to medium-term forecasting. Longer-term statistical approaches, extending to space climate timescales, may further provide actionable insights via scoreboards (§6.9).

Observational limitations discussed in this section should not prevent the community from moving the validation efforts forward by leveraging whatever is currently available. To the contrary, the ongoing validation initiatives and progress made will hopefully drive improvements and can substantiate the case for crucial observational needs in future mission planning.

3.2.4. Ionosonde foF2 and hmF2 measurements

The success of Ionosphere-Thermosphere-Mesosphere (ITM) model validation activities highly relies on the availability and readiness of high-quality long-term observational datasets. Ideally those data span over more than one solar cycle, cover numerous geomagnetic storms, and are distributed spatially so as to sample globally across multiple scales. In addition, the validation-ready data should be easy to access, calibrated, and have detailed documentation with metadata. In identifying potential validation datasets, there is typically a trade-off between dataset quality and comprehensive spatial and temporal distribution, with opportunistic data, including impact data, being vastly more extensive but often much noisier and lacking in details that are required for high confidence in individual data sources. Impact data and other opportunistic data additionally requires processing of model output using tools such as ray tracing or orbit propagation. These tools have their own assumptions that should be recorded in validation documentation. For these reasons, opportunistic data sources are promising but cannot be used in isolation, and although limited in data volume, high-quality data sources are crucial. Here we introduce the datasets that are commonly used for the ionosphere and thermosphere model validation.

The most readily available and low latency ionosphere data is the ionosonde foF2 and hmF2 measurements provided by the Global Ionosphere Radio Observatory (GIRO) portal[8] (Reinisch and Galkin, 2011). More than 17 million ionograms and ionospheric characteristics such as foF2 and hmF2 and vertical electron density profiles are publicly accessible. The critical frequency of the F2 layer, foF2, indicates the highest frequency the ionosphere can reflect. hmF2 is the height of the peak density NmF2 of the F2 layer, affecting the ray path of radio wave signals through the ionosphere. The automatic real-time ionogram scaling with true-height, or automatic scaling system is used by GIRO (Huang and Reinisch, 1982; Reinisch and Huang, 1983). This enables the mass processing of ionosonde data. There are limitations, for example, during active ionospheric conditions and during severe geomagnetic storms. During active ionospheric conditions, spread F can occur limiting the ability to retrieve information about peak parameters. On the other hand, during severe geomagnetic storms, the auto scaling software leads to errors including sporadic E layer misidentified as F2, F1, or E layer; second reflection scaled instead of first reflection (frequency might be correct but not the virtual height); F1

[8] https://giro.uml.edu/.

layer misidentified as F2 layer, etc (Stankov et al., 2012, 2023). As a result, manual scaling is necessary to retrieve the correct electron density profiles and foF2/NmF2 during disturbed periods. Additionally, the absorption and blackout of radio waves due to solar flares and particle precipitation can lead to the absence of echoes in ionosondes. The limitation of manual scaling data and the absence of ionosonde echoes make the storm time model foF2 and hmF2 validation challenging. The limitation of ionosonde data used as validation datasets is due to its limited spatial coverage, in particular, over the ocean and polar regions. High-latitude data sets, which are increasingly in demand, such as the Canadian High Arctic Ionospheric Network (CHAIN), are notably missing from GIRO.

There are on-going efforts to improve the gaps in data coverage and data quality through the Global Ionosonde Observation/Operation Network (GION), established as the global comprehensive representative of ionosonde operators and users. This is a joint project of the International Space Weather Coordination Forum[9] (ISWCF) facilitated by the WMO-ISES-COSPAR Coordination Team[10] (WICCT). This coordinates between the World Meteorological Organisation (WMO), the International Space Environment Services (ISES), and COSPAR. Discussions of the ISWCF and WICCT initiatives are presented in Ishii et al. (2026b, Section 4.2, this issue). The purpose of GION is to improve communication among ionosonde and observation data users and also to serve as the representative of the global ionosonde community within the ISWCF. As its first initiative, GION is engaging in discussions to raise the visibility of ionosonde observations through WMO Integrated Global Observing System (WIGOS) and WMO Observing Systems Capability Analysis and Review Tool (OSCAR).

3.2.5. Global Navigation Satellite System (GNSS) for ionospheric monitoring

Global Navigation Satellite Systems (GNSS) provide another popular dataset that can be utilised for ionospheric monitoring and, hence, for the validation of ionospheric models. From various networks of ground-based GNSS receivers, it is possible to estimate the TEC. TEC is the integration of the ionospheric electron density along the ray path between a GNSS satellite and a receiver. By combining GNSS observations from different frequencies (e.g., code, phase) it is possible to deduce an estimate for TEC along the lines of sight of each satellite in view from a given ground receiver. Slant TEC (STEC) estimates can be calibrated and combined to provide an indication of the ionospheric vertical TEC at the zenith of each ground receiver. The GNSS data in Receiver INdependent EXchange (RINEX) format is provided by several centers at the International GNSS Service (IGS), University NAVSTAR Consortium, and other various regional networks. Madrigal TEC developed by the Massachusetts Institute of Technology Haystack Observatory is publicly accessible[11] to evaluate the performance of ionosphere models on 2D maps (e.g., Chou et al., 2023). The Madrigal TEC has standard spatial and temporal resolutions of 1° latitude by 1° longitude with 5 min cadence. The limitation of TEC data is the lack of data over oceans and some continents such as Africa, Antarctica.

In addition, GNSS RINEX data allow the estimate of temporal and spatial variations in the background ionisation, through the calculation of spatial and temporal gradients in TEC (Forte et al., 2026, this issue; Jakowski and Hoque, 2019). Dedicated GNSS networks also provide scintillation information at both equatorial and high latitudes: the data can be retrieved from institutes and observatories that operate these networks (in various latitude and longitude sectors). For example, CHAIN use specialized Ionospheric Scintillation Monitoring Receivers that are capable of collecting high-rate raw GNSS data (e.g., at 50 or 100 Hz), which is necessary to resolve the rapid signal fluctuations associated with scintillation. A comprehensive review of ground-based networks including scintillation monitoring available world-wide is presented in Ishii et al. (2026a,b,c) in this special issue.

3.2.6. GNSS radio occultation (RO) foF2 and hmF2 measurements

The spaceborne GNSS radio occultation (RO) foF2 and hmF2 measurements provide global data coverage without continental limitations, e.g., (Lin et al., 2007a,b). The RO remote sensing technique uses low-earth orbit (LEO) satellites to track radio signals from GNSS transmitters. The radio wave passing through the ionosphere can provide TEC measurements and vertical profiles of electron density through the Abel inversion (Schreiner et al., 1999, 2007; Syndergaard, 2002; Lei et al., 2007) or one-dimensional variational retrieval methods (Culverwell et al., 2024; Elvidge et al., 2024). Several RO missions have been launched, including GPS/Meteorology (GPS/MET), CHAMP, FORMOSAT-3/COSMIC, FORMOSAT-7/COSMIC2, MetOp, GRACE-FO, Korean Multi-Purpose Satellite (KOMPSAT-5) (e.g., Hajj and Romans, 1998; Jakowski et al., 2002; Schreiner et al., 2020; Hoque et al., 2023) and commercial RO missions from GeoOptics Inc. and Spire Global Inc (e.g., Angling et al., 2021; Chang et al., 2025; Zakharenkova et al., 2025). The RO-derived foF2 and hmF2 have been proven to provide reliable and high-quality data for global ionospheric specification (e.g., Kelley et al., 2009; Cherniak et al., 2021; Lin et al., 2020). One caveat is that the spherical symmetry assumption of the Abel inversion can lead to significant retrieval errors to the ionospheric E region and equatorial ionosphere (Yue et al., 2010; Chou et al., 2017). RO missions can also provide information about scintillation on occultation links (e.g., Wu, 2020).

[9] https://www.iswat-cospar.org/iswc-forum.
[10] https://iswat-cospar.org/wicct.
[11] https://cedar.openmadrigal.org/index.html.

3.2.7. Neutral density

The most reliable data to validate against thermosphere models is the in-situ accelerometer-inferred neutral density onboard satellites including CHAMP, GRACE, GOCE, GRACE-FO and Swarm. The variations in density depend on position, season, and the level of solar and geomagnetic activity. These instruments provide precise measurements from pole to pole and are also appropriate for storm-time model evaluation.

Neutral density can also be derived from GPS tracking by means of a Precise Orbit Determination (POD) process, which offers an alternative yet reliable approach for estimation of in-situ thermospheric density (van den IJssel and Visser, 2007; Siemes et al., 2016) and validation of thermospheric models (Waldron et al., 2024).

Accelerometer or POD-derived densities have different merits. The accelerometer inferred densities have a resolution of about 80 km along the orbit, whereas the resolution of a POD-derived dataset is much coarser. Presently, only two accelerometer satellite missions are in operation, GRACE-FO and Swarm while there are more satellites providing the POD densities.

However, each density dataset derived from acceleration or orbit data relies on assumptions about drag coefficients, which remain the largest source of uncertainty in retrieved neutral density. The choice of drag coefficient can result in substantial differences (10–20%) in density estimates at ~500 km altitude during periods of low solar activity (Mehta et al., 2023, Bruinsma et al., 2023). Determining which drag coefficient model best represents physical reality remains an open question. These in-situ measurements inherently provide poor spatial and temporal distribution. Because measurements are only available along the orbit, they are limited at the altitude of the satellite, and the local times of ascending and descending pass for a given day.

3.2.8. Neutral temperature

Neutral temperature disturbances in the lower thermosphere can have a notable impact on density at higher altitudes (e.g., Laskar et al., 2023). A 1 K temperature perturbation over a 10 km height range around 100 km can cause about 1% density change at higher levels due to hydrostatic adjustment (Bruinsma et al., 2023). Thus, improved representation of temperature in a modeled lower atmosphere can lead to a more accurate density at higher altitudes. However, neutral temperature in the modeled lower thermosphere has not yet been systematically validated against long-term space-based observations. The recent launch of the Global-scale Observations of the Limb and Disk (GOLD) mission provides the first full-disk retrieval of thermospheric temperature (Eastes et al., 2017, 2020; Evans et al., 2024), presenting a unique and indispensable opportunity to rigorously assess and improve model representations of neutral temperature dynamics in the lower thermosphere (e.g., Liu et al., 2023).

3.3. Selecting metrics

Measures that quantify the difference between a specific physical quantity simulated by a model and observations of this quantity are often called 'metrics'. Fig. 3.2 illustrates three broad categories of metrics for time series of physical quantities: point-to-point comparison metrics, threshold-based metrics, and event-based metrics.

Point-to-point metrics are widely-applied for model-data time series comparisons which directly compare a collection of paired values, where one value in each pair is measured, and the other is modelled at the same time and location (see vertical arrows in Fig. 3.2a).

- Metrics such as mean error (ME), mean absolute error (MAE), mean squared error (MSE), root mean square error (RMSE), mean relative error (MRE), mean ratio, standard deviation (STD) quantify the magnitude of the error, and are used to quantify model accuracy.
- Metrics such as correlation coefficients (e.g., Pearson correlation coefficient), R, and the coefficient of determination, (R2 or R^2 / R-squared), measure the linear relationship between modeled and observed quantities.

Threshold-based metrics quantify how well a model captures occurrences in the time series where a physical quantity exceeds (or falls below) a defined threshold value (Owens, 2018**).** Owens (2018) argues that labeling each time step as an event or non-event based on a threshold value and computing binary metrics from it offers two main advantages. First, binary metrics can assess the ability of a model to produce actionable predictions, an aspect that point-to-point metrics lack. Secondly, error functions and correlation coefficients are more sensitive to outliers than binary metrics. In this way, we can validate a model based on the number of correctly predicted events (true positives, TPs), forecasted events that are not observed (false alarms, FPs), observed events that are not predicted (misses, FNs), and correctly predicted non-events (true negatives, TNs). A contingency table summarizes these numbers, which allows for the computation of various skill measures afterwards. For example, Reiss et al. (2023) outlines a range of such skill measures, one of them being the True Skill Statistic (TSS) which is defined on a scale of [1, −1]. According to TSS, a perfect forecast model would score 1 while a model without skill would score 0. A perfect inverse forecast, on the other hand, would score −1. The TSS is popular in the validation community because it uses all entries in the contingency table and is not biased by the ratio between observed and forecasted events. We refer the reader to (Pulkkinen et al., 2013; Verbeke et al., 2019, Kahler and Darsey, 2021; Reiss et al., 2023) for a more thorough review of scores that are associated with the usage of contingency tables, such as the Probability of Detection, the False Alarm Rate (FAR), the Success Ratio, the Bias Score, the Heidke Skill Score (HSS), the Critical Success Index (CSI), and the TSS.

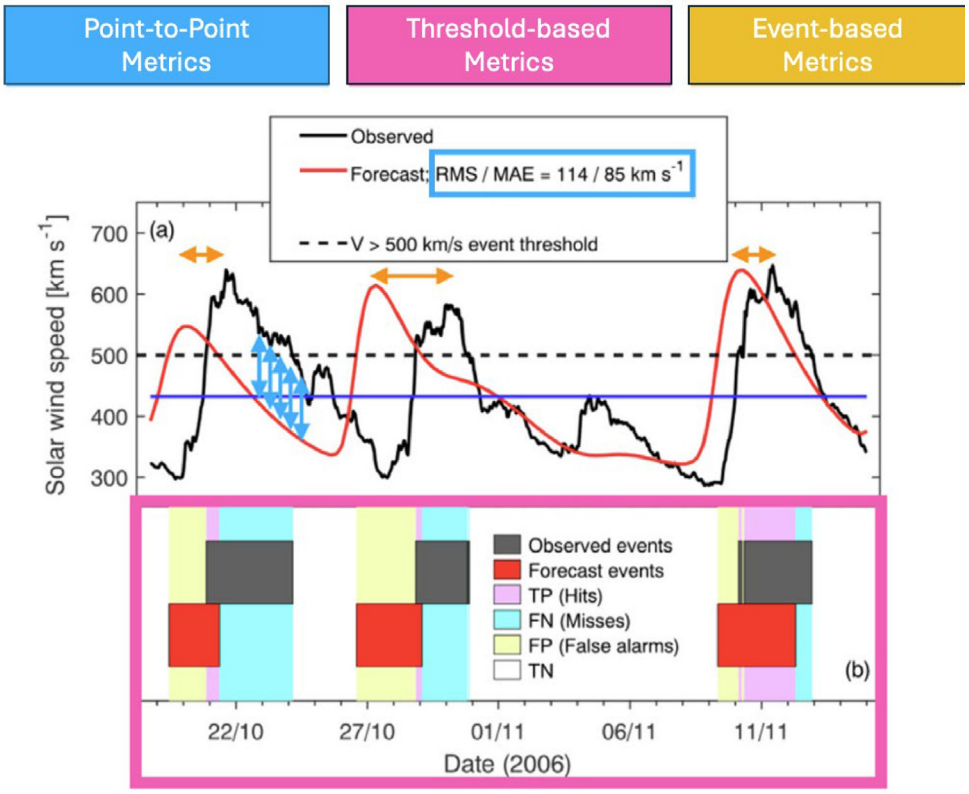

Fig. 3.2. Three categories of metrics commonly used for time series comparisons of physical quantities such as solar wind speed at L1: a) Observed (black) and modeled (red) solar wind speed as a function of time, compared through point-to-point (blue arrows) and event-based metrics (orange arrows). The blue line shows the climatological mean as a benchmark. b) Threshold-based metrics (pink box) defined through a user-defined event threshold, 500 km/s in this case as indicated by the horizontal black dashed line in Panel (a). Modified from Owens (2018) and Reiss et al. (2023).

Event-based metrics identify discrete events in a time series. These events are typically defined by an increase/decrease in a physical quantity within a given time interval. When such a criterion is applied to both model and observation, a statistical analysis of errors in event timing, amplitude, and duration can provide valuable insights. For example, even when the temporal evolution of solar wind properties is captured reasonably well, differences may still arise between the predicted and observed arrival times. Several studies have proposed a strategy involving three steps to address this problem (Owens et al., 2005; MacNeice, 2009a,b; Reiss et al., 2016). This includes detecting abrupt changes from slow to fast solar wind speed, associating these "events" in measurements and forecasts, and computing binary metrics (Fig. 3.2). We refer the interested reader to Reiss et al. (2023) for more details on event-based metrics.

Exact definitions of events and thresholds for threshold-based and event-based metrics are domain-specific. For instance, in magnetospheric research, a ground magnetic perturbation is defined by |dB/dt| exceeding a certain threshold (e.g., Welling et al., 2018). In solar-heliospheric research, a high-speed enhancement is defined by how abruptly the solar wind bulk speed changes from slow to fast (e.g., Owens et al., 2005).

Fig. 3.2 illustrates these three types of metrics commonly used in model-data comparisons on the example of solar wind bulk speed at L1. However, many more metrics are used by the space weather community. For instance, probabilistic-based metrics, such as the Brier Skill Score and Relative Operating Characteristics Skill Score, are often used for solar flare forecasts (see § 4.1.4). Other examples, inspired by terrestrial weather, include studying the probabilistic information from creating large ensembles of model solutions. Here, a so-called "cost-loss analysis" provides valuable insights for users with different risk tolerances (Owens and Riley, 2017; Henley and Pope, 2017). Another technique that quantifies timing and amplitude errors is Dynamic Time Warping (DTW, Samara et al., 2022). DTW is particularly effective when comparing time series with similar patterns but temporal offsets, which is sometimes the case between solar wind observations and model solutions. The assessment is conducted point-by-point between the time series, considering not only points at the same timestamp but also neighboring points. In this way, dynamic time warping can assess both amplitude and time differences between two time series. Furthermore, statistical approaches to model validation, which study the overall behavior of a system, are discussed in Section 4.3. Liemohn et al. (2021, 2025) provide a comprehensive summary of key metrics used for magnetosphere model

validation, that are applicable to other heliophysics domains as well.

Although a wide variety of metrics exist in the literature, it is important to recognize that each metric focuses on different aspects of model performance. The relative importance of metrics depends on the specific application. Consequently, model ranking significantly depends on the selected metrics. Models that rank highly for metrics selected for one specific ESEQ may perform poorly when evaluated with a different metric or when a different physical quantity is studied.

It is important that validation studies for space weather applications produce scores and model rankings that are understandable to forecasters and users of specific applications. On the other hand, narrowly defined parameters required for operational space weather forecasting cannot be used to rank the overall capabilities of a model. Moreover, narrowly defined parameters cannot assess the quality of advanced physics-based models and their ability to provide insight into underlying physical processes. Debates on approaches to standardizing metrics started more than 15 years ago. Defining metrics standards was one of the major discussion topics in the IFSWCA. A big part of the space weather modeling community has now realized that no single metric can quantify the performance of a model, but a combination of them is necessary for a complete and comprehensive assessment. But how many of those metrics are enough?

Various research teams and authors around the world apply metrics that they are the most familiar with, or prefer in terms of easier applicability, serving specific purposes, or convenience of clarity of results. The question remains: which subset of those metrics is the most adequate for the specific phenomenon of interest that will provide the most complete evaluation and can be adopted by the community for scientific research and publications? A recently formed overarching activity O1-03 under COSPAR ISWAT "Toward a community consensus for metrics in solar physics and space weather" is addressing this question.

To facilitate space weather community validation studies and enable flexibility in metric selection, the CCMC created an open-source software that includes community-developed algorithms for model-data comparisons (Validation Skill Score Library[12]). The Validation Skill Score Library is used in the backend of a web-based automated assessment system CAMEL (Comprehensive Assessment of Models and Events Using Library Tools) but can also function as a stand-alone tool (see §3.6.1).

To facilitate coordinated validation activities within ESA's Space Weather Service Network of more than 50 participating expert groups, a set of validation guidelines has been developed which includes a comprehensive set of metrics along with guidelines for their use in validation studies. These are used as the basis for all network related developments and are made available online to support wider community initiatives and collaborations.[13] Ongoing developments are currently working to implement these as the basis of an online validation platform geared towards supporting further validation initiatives.

The space weather community can also take advantage of comprehensive libraries of metrics in Model Evaluation Tools (MET[14]) developed for evaluating the performance of forecasts from numerical weather models used by the terrestrial weather community. MET includes

- standard verification scores comparing gridded model data to point-based observations
- spatial verification methods comparing gridded model data to gridded observations using neighborhood, object-based, and intensity-scale decomposition approaches
- ensemble and probabilistic verification methods comparing gridded model data to point-based or gridded observations

3.4. Reference models

Reference models serve as valuable benchmarks for putting new model development into context. These reference models can be commonly accepted or current state-of-the-art. Using operationally-implemented models as reference can be useful for assessing the potential of new models with enhanced capabilities. Alternatively, it is often helpful to consider much simpler models, typically defined by straightforward mathematical or empirical relationships. Simple reference models are useful as they provide a minimum threshold that more sophisticated models should surpass, but can often be irritatingly difficult to beat. By comparing newly developed models to reference models, scientists can determine whether new developments offer meaningful improvements.

In validation studies of solar wind models, for instance, a commonly used reference model is the climatological mean, defined as the mean value of observations (see, e.g., Owens, 2018). Another reference is the recurrence model, which assumes that solar wind conditions at Earth will repeat after each Carrington rotation due to coronal holes often only evolving slowly relative to the 27-day solar rotation timescales. A variant of this is the persistence model, which assumes that solar wind conditions will remain the same as they were a defined number of days ago, e.g., due to similar solar wind conditions from coronal holes (Owens et al., 2013; Reiss et al., 2016). Sometimes the recurrence model is referred to as the "27-day persistence model", as the conceptual difference between recurrence

[12] https://github.com/nasa/camel.

[13] https://swe.ssa.esa.int/documents/d/guest/ssa_swe_escdef_validation_guidelines.

[14] https://dtcenter.org/community-code/model-evaluation-tools-met.

and 27-day persistence is frame-dependent (from the perspective of the Earth having similar conditions re-occur when re-encountering the same solar wind stream, or a given location on the solar corona having similar conditions persist). A more sophisticated yet still physically-trivial reference model which can also be hard-to-beat is the "analogue" or "similar day" model, which simply performs pattern-matching of current conditions against the past (e.g., Owens et al., 2017, and references therein to terrestrial heritage, and other space weather uses). Using such reference models in validation studies provides a consistent benchmark that enables future research to track progress over time.

An advantage of reference models is that they allow for the calculation of metrics that show relative improvements of new developments. One approach for comparing a new model to a reference model is using the so-called skill score, which compares the MSE of the forecast to the MSE of the reference model. A model that is less skillful, similar, or better than the reference model results in a negative, zero, or positive (between 0 and 1, with 1 being the best possible forecast of the model) skill score value, respectively (see e.g., Owens et al., 2005; 2008).

3.5. Repositories and metadata for observational and simulated timelines

Validation projects involving comparisons of time series for selected parameters use data obtained from different resources (models and observations) for the same locations and time periods. Summarising results of model-data comparison in publication and presentation, however, is not sufficient for coordinated community-wide validation studies and tracking progress over time. All observational and simulated timelines (time series) with comprehensive metadata should be stored in open repositories preferably in standard format with an Application Programming Interface (API) access. The ILWS-COSPAR Roadmap (Schrijver et al., 2015) has recommended standardizing metadata and harmonizing access to observational data and model outputarchives including repositories of time series for validation projects.

3.5.1. SPASE and HAPI standards

Over the past two decades, the heliophysics community developed the Space Physics Archive Search and Extract (SPASE) metadata model for recording descriptive information about observational data products (Roberts et al., 2018). Today, SPASE is maintained by the international SPASE working group.[15] For a detailed discussion on SPASE metadata, we refer to Fung et al. (2023; Roadmap Special Issue 1) and Masson et al. (2026; Roadmap Special Issue 2, this issue). Modeling centers such as the Virtual Space Weather Modeling Center (VSWMC) in Europe and the CCMC in the US use SPASE metadata to describe models. Implementation of SPASE for derived space weather products (including timelines for validation) is an on-going effort at the CCMC.

The Heliophysics Application Programmer's Interface (HAPI[16]) is a standard developed by the heliophysics community for accessing distributed time series data to increase interoperability (Weigel et al., 2021). HAPI specification allows data providers to use a standard set of conventions for returning data in response to a URL based request. HAPI metadata includes less details than SPASE metadata. However, in contrast to HAPI, SPASE does not include a specification for the access of data. HAPI metadata can reference any existing SPASE record. All HAPI metadata uses JSON (JavaScript Object Notation), which has near-universal support in modern scientific programming languages. COSPAR Panel on Space Weather (PSW) 2018 Resolutions recommended SPASE be the metadata standard and HAPI be the common data access API for space science and space weather data.

The CCMC Integrated Space Weather Analysis (ISWA[17]) system supports the HAPI Data Access Specification[18] (Version 2.0) for delivery of time series data. A HAPI compliant ISWA-HAPI server catalog[19] lists all registered datasets with unique identifiers (IDs) and descriptions of all parameters in every dataset record. Currently ISWA – HAPI includes all data resources/parameters that are shown in ISWA's timeline. More timelines are planned to be available via ISWA-HAPI in the future.

3.5.2. Time periods

Validation studies for time series are typically focused on a set of time periods (sometimes referred to as events). Time periods used for community-wide studies should have IDs for easy references. To facilitate community-wide validation projects, the CCMC initiated a database of Time Periods of space weather phenomena. Currently there are more than 150 Geospace Storm Time Periods in the CCMC Metadata Registry[20] (CMR). Start/end times for geospace storm time periods are selected to include all storm phases including a relatively quiet pre-storm phase. Metadata for geospace storm time periods includes global storm characteristics (e.g., max Kp, min Dst, number of consecutive storms), drivers of geospace (solar wind parameters at L1, solar indices), run IDs for all simulations in the CCMC database with links to results, and related publications. Time Period Identifiers include year and month of the Time Period start time and a unique number to distinguish time periods with the same start year/month. Subsets of time periods used in publications and/or community validation campaigns are tagged to facilitate track-

[15] https://spase-group.org/.

[16] https://hapi-server.org/.

[17] https://ccmc.gsfc.nasa.gov/tools/ISWA/.

[18] https://github.com/hapi-server/data-specification

[19] https://iswa.ccmc.gsfc.nasa.gov/IswaSystemWebApp/hapi/catalog.

[20] https://kauai.ccmc.gsfc.nasa.gov/CMR/TimeInterval/viewAllTI.

ing progress over time by modeling the same set of time periods with upgraded or new models.

3.5.3. Timelines for validation studies: quantities, locations, resources

Timelines for time series validation studies can be described by *quantities, locations, and resources.* Each of these timeline characteristics need to have an ID and metadata.

3.5.3.1. Quantities. A basic challenge in space weather validation is the difficulty in associating nominally equivalent quantities between models and observations, due to the comparative lack of standards, notably for metadata on quantities. It is typical that the same physical quantity (e.g., proton density) has different names in different space weather data sets even at the same data center. This arises as space weather lacks a holistic equivalent or support for the *standard_name* quantities descriptor metadata of the Climate and Forecast (CF) Metadata Conventions[21] adopted in terrestrial weather, typically within files in NetCDF (Network Common Data Form) file format (Rew et al., 1989; Hassell et al., 2017). In terrestrial weather, this *standard_name* identifier enables powerful automated identification of equivalent observed or modelled variables (terrestrial equivalents of ESEQs) across all the relevant terrestrial spheres (cryosphere, hydrosphere, biosphere, atmosphere), providing a unified solution useful for comparisons of multiple models and multiple data sets of the coupled Earth system, despite the inherent heterogeneity of the component spheres and the variables concerned, to good effect (Henley and Andries, 2025).

A similar effort in the astronomy community, has taken a slightly different approach, providing unification where feasible (e.g., the International Virtual Observatory Alliance controlled vocabulary[22]), while also maintaining a list of vocabularies,[23] to allow for multiple overlapping pre-existing vocabularies, and avoiding incurring the time and effort to create a single consensus vocabulary. Both approaches have their merits; here we concentrate on the CF approach.

The CF conventions include a standard name table,[24] which defines strings that identify physical quantities. The CF conventions include guidelines[25] on how standard names should be constructed, allowing embedding of "qualifications" giving richer information. Example standard names from the table range from the simple (e.g., "air_temperature") to the more qualified (e.g., "air_temperature_at_effective_cloud_top_defined_by_infrared_radiation").

Further coordination work is needed to examine the interplay of existing enterprise metadata standards, notably International Solar-Terrestrial Physics (ISTP) dictionary keyword guidelines, SPASE vocabulary (see Fung et al., 2023; Masson et al., 2026), and terrestrial compatibility considerations for operational space weather actors (e.g., with systems based on the CF conventions), to see what-if-any unification is desirable and achievable (see also discussion in Corti et al., 2026). Implementing NetCDF file format and CF convention specifically is not a requirement for space weather. Rather, the important action is to reach an agreement on naming convention and unique IDs for physical quantities in space weather (including ESEQs) as a necessary step for harmonising model-data comparisons for the same quantity from different simulations and observational datasets. Databases of quantities (parameters, variables) and lookup tables linking alternative names and identifiers for the same quantity would be helpful for interconnecting different repositories and to ensure that we are comparing the same quantities (parameters, variables).

3.5.3.2. Locations. Parameters derived from observations and model outputs are compared at the same location. Examples of locations are ground stations (e.g., magnetometer or ionosonde) or orbits (e.g., GOES). For global parameters (such as global indices) or other integrated quantities location is not applicable. Metadata for most orbits are available at Heliophysics Data Portal.[26] Other valuable resources include the WMO OSCAR database: its queryable satellites table[27] in the space-based capabilities section provides useful locations on many other (mainly operational) environmental satellites worldwide. For example, space weather monitoring payloads are carried by spacecraft such as the Korean GEO-KOMPSAT-2A,[28] for which the longitude in geostationary orbit is relevant, and the sun-synchronous polar-orbiting Europe's Meteorological Satellite Agency (EUMETSAT) Metop-B,[29] for which the equatorial crossing time is key. Pilot projects on space weather uses of the WMO Information System (WIS2.0) for global data sharing (discovery and transport) are ongoing,[30] in the space-based case for GEO-KOMPSAT-2A magnetometer data by ESA; in the ground-based case for ionosonde data by Brazil's National Institute for Space Research (INPE). The WMO Core Metadata Profile Version 2 (WCMP2)

[21] https://cfconventions.org/.
[22] https://www.ivoa.net/documents/UCD1+.
[23] https://www.ivoa.net/documents/latest/Vocabularies.html.
[24] https://cfconventions.org/Data/cf-standard-names/current/build/cf-standard-name-table.html.
[25] https://cfconventions.org/Data/cf-standard-names/docs/guidelines.html.
[26] https://heliophysicsdata.gsfc.nasa.gov/websearch/dispatcher.
[27] https://space.oscar.wmo.int/satellites.
[28] https://space.oscar.wmo.int/satellites/view/geo_kompsat_2a.
[29] https://space.oscar.wmo.int/satellites/view/metop_b.
[30] https://community.wmo.int/en/meetings/5th-meeting-expert-team-space-weather-30-october-1-november-2024.

Fig. 3.3. Simulation Run is a resource for Simulated Data. Run metadata should include Time Period ID and Simulation Settings ID. Run ID and Post-processing tools metadata should be included in Simulated data metadata. Validation studies are usually test simulation settings and how changes in simulation settings are affecting model performance. The same simulation settings can be applied to different time periods to generate run series for a specific simulation setting.

discovery metadata[31] underpinning WIS2.0 includes spatial extent information[32] (in the Earth coordinate reference system, i.e. GEO), which promises powerful discovery-by-location of relevant observations which can be bounded in this way (e.g., those from ground stations), provided data providers populate this metadata appropriately.

3.5.3.3. Resources. Resources for observational timelines are datasets corresponding to an instrument (on a spacecraft or on the ground) and data processing techniques. SPASE metadata and identifiers are available for most observational datasets.There are on-going community efforts to harmonize access for ground observatory networks. Observations may not be available for all resources for all time periods. Inventory tables for data availability are important.

Resources for simulated timelines (solutions) are simulation runs that frequently include multiple modeling components, pre-processing (preparation of model-input-ready data from observations) and post-processing (deriving timeline solutions from model outputs) tools. Simulation runs should have comprehensive metadata with unique run identifiers. Simulation settings include the version of the modeling system and a broad range of internal parameters. Applying the SPASE metadata schema to complex multi-component simulation runs is challenging and may be not practical. Complexity of run metadata is discussed in detail in Corti et al. (2026).

Validation studies usually use run series for multiple time periods (events) with the same internal simulation settings. Relations between Simulation Run, Time Period, Simulation Setting, and Simulated Data are illustrated in Fig. 3.3. Parent run and simulation settings IDs should be included in metadata for all simulated data (timelines / solutions) used in validation studies. Tracking and tagging changes in simulation settings (including minor/major model upgrades, and how input data are pre-processed and how timelines of simulated ESEQs are derived) are important for tracking progress and to avoid confusion with interpretation of metrics studies. Note that timelines for validation can be derived from run outputs using different post-processing tools that may produce different solutions.

3.5.4. Metadata for community validation campaigns

Efforts to provide metadata for community validation campaigns are important because they describe how models are configured, executed, and compared to observations. As an example, Reiss et al. (2023) outlines a comprehensive metadata description for the entire validation campaign of large-scale models of the solar corona and solar wind. The authors propose eight types of metadata needed to register a model into their validation effort as part of the ISWAT Team H1-01.[33] This includes all necessary metadata components of the validation study, such as metadata about observations, data preprocessing, model description, model settings, model output, model solutions, and the model chain, focusing on how models from each domain are coupled. Providing this complete metadata description is an important component of the capability to track progress over time.

3.6. Tools for community validations

Establishing validation standards in the community, including the selection of events or time periods, metrics, file formats and metadata, is an important foundation. An equally important step is to use these agreed-upon standards to continuously examine model performance and track progress over time. To draw robust initial conclusions, which can be critically re-examined by others, and stratified to check for potential pitfalls, the validation framework needs to be flexible and easily accessible to the community.

This can be achieved with open tools and interactive archives that support the assessment of space weather models. As discussed in MacNeice (2018), it is recommended that community-agreed-upon validation procedures are automated as much as possible because quality assessment 1) is labor-intensive, 2) suffers from inconsistencies, and 3) requires an infrastructure that can keep up with the pace of new model developments. Here, interactive web-based systems can support the community in meaningful quality assessment efforts. Ideally, such web-based systems have

[31] https://docs.wis2box.wis.wmo.int/en/1.0.0/reference/running/discovery-metadata.html.

[32] https://wmo-im.github.io/wcmp2/standard/wcmp2-STABLE.html#_1_11_geospatial_and_temporal_extents.

[33] https://iswat-cospar.org/h1-01.

standardized formats for model solutions that are uploaded, and compute reports showing the abilities of the model under scrutiny.

Adoption of a common validation tool across the space weather enterprise is likely to show "network effects": once a few in the space weather enterprise adopt a given tool, and do some hard work writing Input/Output modules which can be reused by others (for same use-case, or to serve as examples for other use-cases), the ease-of-use and benefits to them and others accrue. As such, it is worth having wider discussions going forward, and experimentation, to see if this is true.

In this section, we discuss examples of currently available tools and frameworks for community validations.

3.6.1. CAMEL

The CAMEL tool (Rastätter et al., 2019) is a web-based framework developed and maintained at the CCMC to introduce transparency to space weather validation activities and to support community-wide validation projects. The backend of CAMEL leverages CCMC services and infrastructure such as Run-on-Request and CMR. The CAMEL web application is built on the Plotly Dash framework and makes heavy use of Pandas DataFrames, NumPy and other Python modules.

CAMEL is designed for on-the-fly skill score calculation with minimal pre-calculation caching. This allows the user to apply any combination of interpolation types, gap filtering, binning, algorithms for score calculations, etc. Simulated timelines are tagged by simulation setting IDs that indicate the selected data from each category. This enables comparisons for different simulation settings (including different model versions, resolutions, type of drivers, etc) for the same model. An inventory of available timelines can be reviewed before configuring or submitting skill score generation requests. Time periods, simulation settings, and locations (stations/orbits) for score calculations can be assembled manually or tagged subsets linked to publications/campaigns can be selected. Skill scores can be calculated for specific phases of the phenomena (e.g., storm growth phase), a given parameter requirement (e.g., filtered by Kp index), or for a subset of locations (e.g., based on latitude).

Scorecards can be generated for a single simulation setting or for multiple models/settings. Any metrics (algorithm for timeline comparisons) available in the Validation Skill Score Library can be used for ranking. Scorecard designs are specified by the community (e.g., ISWAT team leading a campaign).

Currently most of the simulated timelines for CAMEL are derived from runs generated at the CCMC. Timelines generated outside of CCMC can also be added if simulation settings are registered in the CMR. This enables modelers to check how new developments compare to other solutions across the many different metrics.

In addition to using the CAMEL interactive web analysis tool, the CAMEL framework provides an API allowing users to download approved timelines available on the CAMEL backend database into their own environment. "Approved timelines" refer to data sets (observations or simulations) for which the original owner/creator *gives full permission to the CCMC to share and redistribute the data set.* Moreover, the API allows users to access the full list of ground stations, satellites, model settings, and parameters that are available for each study.

Validation campaigns and ISWAT validation projects currently supported by CAMEL include:

- Thermosphere Neutral Density (ISWAT Team G2A-01)
- Assessing Thermospheric Neutral Density Models using GEODYN's Precision Orbit Determination (ISWAT Team G2A-04)
- Ionosphere Plasma Density (ISWAT Team G2B-05)
- Internal Charging Effects (ISWAT Team G3-04)
- Ambient Solar Wind (ISWAT Team H1-01)
- Ground Magnetic Perturbations (Operational Geospace Model Selection Campaign, GEM Geomagnetic Disturbance Modelling Challenge, ISWAT Team G1-08)

3.6.2. NSF NCAR METplus tool

METplus[34] (Jensen et al., 2024; Jensen, 2023) is a framework of verification (validation) and diagnostic tools centered on Model Evaluation Tools (MET) that has a broad user-base in the terrestrial weather community. In addition to the MET core component (see §3.3), METplus includes.

- Python wrappers to provide low-level automation and time-looping of MET tools
- METplus use cases: configuration files and sample data to show how to invoke METplus wrappers to build workflows ranging from simple to complex
- METcalcpy: calculates statistical output that can be used by other METplus components for analysis
- METplotpy: plots verification (validation) results
- METdataio: reads METplus output and can load them into a database for storing results
- METviewer: web-based interface for deep analysis and display of results (note: this can also be run on the command line for plotting in batch mode)
- METexpress: a display system for quick analysis via predefined queries

METplus has good potential for use in the space weather domain. It can be used for private validation in-house, and for open community validation on the web. This adaptability stems from it having a nicely-decoupled set of component tools to support the complete analysis end of the full validation chain, allowing web-based exposure and exploration of results in automated validation pipelines.

[34] https://metplus.readthedocs.io and https://dtcenter.org/community-code/metplus.

METplus validation tools go beyond timeline comparisons. METplus has already been used by NOAA's Space Weather Prediction Center (SWPC) for two-dimensional validations of ionosphere and for evaluation and quality monitoring of commercial space weather RO providers. Some key adaptations to the space weather domain include stratification of results based on a time-dependent quality flag (e.g., the number of RO rays in a voxel) and the addition of time-dependent masking capabilities for solar azimuthal and Local Solar Time. Approximately 25 space weather use cases have been created so far, including gridded comparisons, point-based validation, validation of human forecasts, and object-based verification. A key innovation has been to develop a containerized system capable of running all of these use cases for retrospective and real-time evaluation of SWPC's space weather models and forecasts. Good presentations (recordings and slides) on this can be found in Vigh et al. (2020, 2021) and Vigh (2023). CCMC has started a collaboration with NSF NCAR to incorporate components of METplus into the CAMEL framework to support community validation studies focusing on comparison observations from two-dimensional images with gridded model outputs.

METplus support for advanced validation capabilities include "object-based validation", useful for validating physical phenomena (see §5). For example one could in principle use METplus to isolate an object such as a 3D CME in model output, and compare this to reconstructed 3D equivalents derived from tomographic reconstructions of IPS or heliospheric imagery (HI) observations, and validate the equivalence over a time sequence covering an event. Additionally, METplus could scale the preceding analysis over multiple events and multiple models to evaluate how well various heliospheric models perform at forecasting CME dynamics. This direct approach can be challenging (Barnes et al., 2025). Alternatively, a common strategy to compare model and observations for validation or assimilation purposes would be to "forward model" the numerical output to cast it into observation space – e.g., "synthetic white light" equivalents of HI observations, explored for various models (e.g., Odstrčil et al., 2005; Odstrčil, 2023; Chen et al., 2024; this may be challenging in reduced-physics contexts, but workarounds can be found – e.g., Barnard et al., 2023), or equivalents for IPS (e.g., Morgan et al., 2023; Waszewski et al., 2023). Note forward modelling is a core component of terrestrial weather forecasting – see e.g., Saunders et al. (2018) for forward modelling of radiances from Earth-pointing satellite radiometers, a key component of observations assimilated into terrestrial weather models to improve their forecasts (for more details, see Henley et al., 2026, this issue).

Another advanced validation capability available in METplus is "stratification", to perform more nuanced conditional validation, isolating to given conditions – "how do the models perform in this observed regime" (e.g., slow solar wind, during substorms, or complex combinations of these). Advanced terrestrial weather examples of the latter include "flow-dependent verification (validation)", e.g., Ferranti et al. (2015), Rodwell et al. (2018). An equivalent familiar to space weather would include superposed-epoch analyses (examples from various domains include (Carter et al., 2014; Manu et al., 2023; Owens, 2020; Prikryl et al., 2014; Rasca et al., 2021) and tools to aid this, e.g., AMDA (Génot et al., 2021) or SpacePy (Niehof et al., 2022).

While the optimum tool for community validation projects remains to be determined, METplus is a competitive candidate. Any alternatives would need similar features, support, and governance characteristics. For instance, the UK Met Office is exploring avenues for their near-real-time validation strategy (Henley et al., 2024). To move forward, initiate discussions on whether a common validation framework has merit, identify any blockers to collaboration, and explore pathways for advancing. Adoption of METplus by the space weather enterprise has been limited to date. However, there is a strong potential here, and further exploration is warranted. Opportunities and challenges in implementation of tools and practices from terrestrial weather for space weather are discussed in the paper by Henley et al. (2026) in this issue.

3.6.3. ESA's space weather service network online validation platform

To support the increasing need for coordinated validation activities both within the ESA Space Weather (SWE) Service Network and in the wider community, a new online validation platform known as SWEVEn (SWE Service Network Validation Environment) is expected to be deployed in 2026. This platform will provide a flexible framework allowing model developers and service providers to compare both model and service product outputs with reference datasets and/or models. Building on the SWE Service Network validation guidelines[13], it will provide access to an extensive validation toolkit to support online analysis, visualisation and sharing of results.

The framework will interface with existing applications and databases hosted within the ESA Space Weather Data Centre to support analysis. Product verification within the ESA SWE Network is ensured during the preparation phase of product integration in the portal. As part of the verification process, all products are formally undergoing a live acceptance test procedure. This testing verifies the correct functioning of the products and their compliance with the ESA SWE requirements baseline including standards from the European Cooperation for Space Standardization (ECSS). Product integration activities that envisage major software development follow the ECSS protocols according to document ECSS-E-ST-40C.[35]

[35] https://ecss.nl/standard/ecss-e-st-40c-rev-1-software-30-april-2025/.

4. Validations based on historic time periods

Model validation for a set of historical time periods or events is currently the most popular approach in the space weather community for evaluations of current capabilities and tracking progress over time. Getting rich insights requires going beyond the typical case of single model validation, using bespoke metrics or sets of historic events. Such efforts can be a first step, but to provide insights which are truly useful to the modeller and the wider community, it is vital to go beyond and robustly situate a given model's performance with respect to others. Having "apples-to-apples" comparisons between models, using equivalent metrics and equivalent comparison periods gives modellers insights into which approaches are effective; operations insights into how operationally-implemented models compare, and potential pre-operational models to consider further; and end users insights into model reliability, helping them judge how to weight any model forecasts or derived products they may include as one of the inputs into their decision processes.

Community-wide validation studies based on historical time periods are frequently called "Community Modeling Challenges". Community Challenges tie the domain-specific research communities together and thereby create a shared vision for assessment strategies. Community Challenges have been initiated by the Geospace Environment Modeling (GEM) program (see Liemohn et al., 2025 for an overview) and are now regularly organized by GEM, Coupling Energetics and Dynamics of Atmospheric Regions (CEDAR), Solar Heliospheric and Interplanetary Environment (SHINE), and by ISWAT Action Teams.

GEM Challenges not only improved our fundamental understanding of space weather but also helped identify models with the potential for operational space weather prediction. An example is the study by Pulkkinen et al. (2013), which led to selection of the Space Weather Modeling Framework (SWMF) developed at the University of Michigan (Tóth et al., 2005, 2012) for operational deployment at NOAA SWPC. Another GEM Challenge has focused on comparing MHD and empirical magnetic field models (Gordeev et al., 2015). A joint GEM-CEDAR Challenge organized an effort to focus on high-latitude ionospheric data-model comparison (Rastätter et al., 2016) and, later, on global ionospheric state data-model comparison (Shim et al., 2017, 2018, 2023; Özturk et al., 2020). As summarized in Liemohn et al., 2025, other GEM organized Challenges have focused on spacecraft surface charging (Yu et al., 2019), inner magnetosphere cross energy-population interactions (Yu et al., 2019), radiation belt modeling (Tu et al., 2019), dayside kinetic processes (Dimmock et al., 2020; Hietala et al., 2020), and the magnetotail at lunar distances (Runov et al., 2023). Even after such community Challenges end, the shared vision on standards will continue to exist in the future validation approaches of participating groups, which is valuable in itself.

4.1. Examples of on-going community validation studies

"How well does it work?" "Can we trust the predictions?" These are fundamental questions to which significant effort has been spent, and yet, for which significant effort is still needed. The different phenomena relevant to space weather forecasting require different approaches and validation methodologies, as discussed below. Overall, more focus on validation is needed to improve benchmarks and develop new metrics that accurately represent performance, enabling improvement and meeting the needs of end users.

4.1.1. Ambient solar wind

At any given time, the heliosphere is filled by ICMEs that interact with the ambient solar wind. Research shows that understanding the prevailing ambient solar wind conditions is essential for predicting when ICMEs will reach any location of interest within the heliosphere such as Earth (see, e.g., Cash et al., 2015; Riley and Ben-Nun, 2021). Over the past few decades, solar wind research has made significant progress enabled by physics-based models such as MAS (Linker et al., 1999; Mikić et al., 1999), MS-FLUKSS (Hegde et al., 2025), and AWSoM (Tóth et al., 2005; van der Holst et al., 2014), and operational models such as Enlil (Odstrčil and Pizzo, 1999a; 1999b; Odstrčil 2003), the Wang-Sheeley-Arge (WSA) model (Arge and Pizzo 2000; Arge et al., 2003), EUHFORIA (Pomoell and Poedts, 2018), and others. Despite these advancements, questions such as "How well does a model perform over a given time period?" and "How much has a model improved over the past decade?", remain. In recent years, action teams have formed within the frame of the COSPAR ISWAT initiative through a bottom-up approach from the space weather community to foster collaboration and develop standardized strategies for model validation.

4.1.1.1. Ambient solar wind validation campaign. The Ambient Solar Wind Validation Team (H1-01) in the "Heliospheric Magnetic Field and Solar Wind" H1 cluster under the COSPAR ISWAT initiative constitutes an example of community validation studies. This team was formed to assess the state-of-the-art in solar wind modeling. To achieve this objective, the ISWAT H1-01 team united solar and heliospheric domain experts, model and application developers, data providers, forecasters, and end-users of space weather products with the following objectives:

- Develop a comprehensive metadata architecture, including metrics, to enable validation of the state-of-the-art modeling and progress assessment over time
- Implement an open online platform in collaboration with the CCMC to automatically assess the solar wind models with community-agreed metrics and validation procedures

- Quantitatively assesses the state-of-the-art in forecasting the solar wind conditions at Earth
- Use the developed infrastructure to generate "Solar Wind Scorecards" for a single or multiple models

During numerous sessions at international workshops and conferences, the H1-01 action team gathered feedback from the space weather community. As discussed in Reiss et al. (2023), this enabled the team to achieve consensus from a large part of the community on forecasting goals, metrics, and metadata infrastructure.

Forecasting goals are essential because they define the physical properties modeling assets should accurately reproduce. Two forecasting goals that almost all models produce are: 1) the ability to forecast the temporal evolution of the solar wind speed, as well as abrupt changes from slow to fast solar wind and 2) the ability to forecast the magnetic polarity and magnetic sector boundary crossings. These essential solar wind properties are measured by missions such as NASA's Advanced Composition Explorer (ACE; Stone et al., 1998), the Wind spacecraft (Ogilvie and Desch, 1997) and the Deep Space Climate Observatory (DSCOVR; Burt and Smith, 2012) positioned near the L1 point. Moreover, solar wind properties are also measured by missions like the Solar TErrestrial RElations Observatory (STEREO; Kaiser 2005), Parker Solar Probe (PSP; Fox et al., 2016) and Solar Orbiter (SolO; Müller et al., 2020). The locations of current spacecraft missions can be found at the Austrian Space Weather Office (ASWO) website[36] and at the CCMC ISWA system (ASWO-Heliocentric Plots of Satellite Positions Cygnet.[37]).

The team agreed to focus on three types of metrics shown in Fig. 3.2 and discussed in Section 3.3.: point-to-point metrics, binary metrics, and event-based validation, which are widely applied for assessing solar wind models (MacNeice 2009a,b; Owens et al., 2005; Reiss et al., 2016). To validate ambient solar wind models, regardless of the type of metric, an added complexity is distinguishing the ambient solar wind from solar transients. Typically, when assessing the accuracy of ambient solar wind models, periods with ICMEs are excluded based on existing ICME event lists (see, for instance, Reiss et al., 2016, Samara et al., 2025).

Interpreting results with complementary types of metrics is crucial for meaningful model validation. Working together with the community on a unified vision for validating solar wind models, the H1-01 Ambient Solar Wind Validation Team developed an open-access tool as a part of the CCMC CAMEL framework for validating ambient solar wind models by comparing their solutions with in situ measurements at L1.[38] This online tool allows the space weather community to test the abilities of solar wind models with the metrics highlighted above (§3.6.1). By doing so, we aim to create an infrastructure that provides an unbiased assessment of progress in solar wind modeling over time.

4.1.1.2. Quantifying uncertainty in solar wind modeling. Efforts to quantify and reduce uncertainties in solar wind forecasts must begin at the Sun. The solar magnetic field, observed through photospheric magnetograms, provides the fundamental boundary condition for coronal and heliospheric simulations and therefore plays a central role in determining solar wind structure. However, current observations, such as those from HMI (Schou et al., 2012) onboard the SDO (Pesnell et al., 2012), the Vector Spectromagnetograph (VSM) (Keller et al., 2003), part of National Solar Observatory/Synoptic Optical Long-term Investigations of the Sun (NSO/SOLIS) (Keller et al., 2003) and the NSO Global Oscillation Network Group (GONG) network (Hill, 2018) are restricted to the Earth-facing hemisphere, leaving the far-side and polar regions unobserved. To compensate for these gaps, data-assimilative and surface flux transport models (DeVore et al., 1984; Hathaway et al., 2022; Roudier et al., 2009; Schrijver and De Rosa, 2003; Sheeley et al., 1985; Upton and Hathaway, 2014; Vögler et al., 2005; Wang et al., 1989), including ADAPT (Arge et al., 2013) and the Open-source Flux Transport (OFT) framework (Pogorelov et al., 2024), are employed to reconstruct the global magnetic field and enhance full-Sun coverage. Despite these advances, incomplete or imperfect magnetic field observations remain one of the primary sources of uncertainty in solar wind models. The resulting uncertainties in boundary maps propagate through flux-transport, coronal, and heliospheric domains, ultimately influencing the fidelity of simulated solar wind conditions throughout the inner heliosphere.

To evaluate the impact of uncertainties due to incomplete boundary conditions, Hegde et al. (2025) conducted a comprehensive validation study using ensemble time-dependent simulations of the ADAPT-WSA-MS-FLUKSS solar-wind model chain, based on publicly available multiple realizations of HMI-ADAPT maps. The analysis focused on the rising phase of Solar Cycle 25 and examined six intervals (P4-P9), each spanning approximately one solar rotation and corresponding to the fourth through ninth PSP perihelia. This period provided favorable conditions for validation through multi-point observations from PSP, Earth, STEREO-A, and Solar Orbiter, distributed across a broad range of radial distances, heliolongitudes, and heliolatitudes. The ensemble simulations revealed that uncertainties introduced at the photospheric boundary produce spatially and temporally variable uncertainty levels within the inner heliosphere. While the model successfully reproduced major global solar-wind features, including the heliospheric current sheet and high-speed streams, it also produced unobserved structures during certain intervals, with performance varying across spacecraft

[36] https://helioforecast.space/now.
[37] https://iswa.ccmc.gsfc.nasa.gov/app/info?cygnetId=743&dataId=2824.
[38] https://ccmc.gsfc.nasa.gov/tools/CAMEL/.

locations. These discrepancies and the uncertainty propagation study underscore the sensitivity of inner heliospheric simulations to uncertain boundary conditions. Systematic uncertainty assessment is therefore essential for establishing the reliability of solar wind modeling and for guiding future developments in data assimilation, model refinement, and operational solar wind forecasting.

Studies on how uncertainties in the ambient solar wind affect CME arrival time forecasts are discussed in Section 4.1.2 below.

4.1.2. CME arrival time

To provide forecast or hindcast arrival times for CMEs, forecasting agencies and researchers use CME propagation models. There are three main components that influence the predicted arrival times: (1) the ambient solar wind through which the CME is traveling in the heliosphere (see section 4.1.1), (2) the kinematics and inferred magnetic structure of the observed CME, which translate to input parameters for the propagation models, and (3) the CME model that is chosen to simulate the CME and its propagation.

Over the past decades, different models for the propagation of CMEs in the inner heliosphere have been developed. These models span a wide variety of different types ranging from computationally inexpensive models such as empirical models (Paouris and Mavromichalaki; 2017) and drag based models (Núñez et al., 2016, Vršnak et al., 2013, Dumbović et al., 2018, Amerstorfer et al., 2018) to computationally expensive models, typically physics-based, 3D MHD models (Odstrčil et al., 2004; Pomoell and Poedts, 2018; van der Holst et al., 2014; Shiota and Kataoka, 2016; Verbeke et al., 2022). The constant development of new models and the upgrading of existing models has created a need within the community to effectively track progress over the years. The ISWAT CME Arrival and Impact team H2-01 has been active since April 2017. Its main goal is to evaluate how well different models and techniques can predict CME arrival time for a set of predetermined events, based on an agreed-upon time period, with open communication with the community.

It is important to distinguish three categories of forecasts: (1) event based forecast, where a forecast is made on whether the CME arrives at a given location or not (binary yes/no arrival forecast), (2) prediction of arrival time and associated parameters such as the CME arrival peak speed, and (3) prediction of the observed solar wind profile over time (point-based measurements). This is analogous to a weather forecast: (1) will it rain (yes/no) (2) what time will the rain start and how much rain will fall, and (3) how does the rainfall forecast vary over time.

Firstly, when a newly erupted CME is observed in coronagraph imagery, a forecast may be made. Subsequently, a CME may be observed to arrive (or not) at the location of interest such as Earth. The final result categorizes into four options: predicted and observed arrival (hit), predicted but no arrival (false alarm), no predicted arrival and no arrival (correct rejection) or no predicted arrival and unexpected arrival (miss) occurs. As this is a categorical forecast, we can therefore create a contingency table. One important key detail here is that one needs to define what is considered to be a "hit". This definition highly depends on the users and any validation study should be able to consider different contingency tables for different arrival time error thresholds used to define a hit (e.g., 3 h, 6 h, 12 h, 24 h, etc.). Skill scores and references that are related to contingency tables are presented in section 3.3.

Secondly, once hit events have been determined, we can also compute metrics related to the predicted CME arrival time and other parameters of importance. Determining the time error as the difference between the forecasted and the observed arrival time, we are able to determine associated metrics, such as ME or MAE (Verbeke et al., 2019; Jolliffe and Stephenson, 2011).

Several large scale validation studies have been performed over the past years, where the above mentioned metrics track CME arrival time forecast accuracy over time. Wold et al. (2018) found a MAE in arrival time of 10.4 ± 0.9 h for 273 CMEs predicted and observed to arrive at Earth, STEREO A, or STEREO B. These results compare well with Riley et al. (2018), who studied 32 models from 2013 through mid-2018. Focusing on 28 events in the CME Arrival Time Scoreboard (see section 6.1), which were forecasted by 6 models, the authors found that the models could predict CME-shock arrival times within ±10 h, but with STD often larger than 20 h. Kay et al. (2024) revisited the analysis of the CME Arrival Time Scoreboard (see section 6.1) using a sample 3.5 times larger than the original study. They found similar overall results, namely a MAE of 13.2 h, a STD of 17.4 h, and a bias of −2.5 h. They also found that the frequently submitted models tend to outperform those from infrequently contributed models. This indicates a need for the community to reflect on the current state of modeling and propose areas of research to improve model predictions, including observational advancements. There are community supported plans to build a companion validation platform for the CME Scoreboard based on the validation codes developed by Kay et al. (2024). Furthermore, within a research context, model developers have tried to consider the same set of hit events such as Dumbović et al. (2018). A collaboration between SWPC and CCMC has also resulted in a validation set of 33 hit events available.[39]

Lastly, a prediction of the observed solar wind profile over time can be made. We note here that the community's focus so far has been mainly on arrival time predictions, and not as much on time profiles. With the rise of magnetised CME models such as Shiota and Kataoka (2016), the new H2-02 team on Magnetic Profiles of Interplanetary CMEs aims to quantify a model's fit to an observational

[39] https://ccmc.gsfc.nasa.gov/static/files/CCMC_SWPC_annex_final_report.pdf.

profile and to provide this quantification for participating models. The team's main challenge is to determine quantities of use to the end user e.g., Maximum B, most negative Bz, timing of min Bz, duration about some critical value.

Over the past years, the community has encountered a few struggles while trying to create a benchmark for CME arrival time forecasting. While focusing on just hit events gives crucial information about how well we are able to predict arrival times (Riley et al., 2018, Kay et al., 2024), it does not tell us about how well we are able to predict whether a CME will hit or not. For this purpose, studies such as Wold et al. (2018) are important as it includes not only hit events, but all events over a longer time period. Furthermore, some models do need information about preceding CMEs to be able to make their predictions, as the preceding CMEs affect the solar wind through which the CME of interest is travelling. For this reason it is important to consider longer periods of time such as suggested by Verbeke et al. (2019), where the time period of 1 January 2011 to 31 December 2012 and 1 January 2015 to 31 December 2015 have been chosen. These chosen time periods allow for intervals of both high and low activity, as well as a difference in the availability of different observational viewpoints.

While model development has flourished over the past decade, resulting in a wide variety of different models, the methods to derive the inputs used to drive those models have stayed very similar and contain a human-in-the-loop component that adds subjectivity to the derived parameters. While this may currently be required operationally (see §9.2), it complicates attribution, and discerning useful model improvements. Unlike is currently established to be the case for flare forecasts (§4.1.4), the lack of stratified validation of CME forecasts with and without human intervention means it is also less clear that subjective modifications consistently add value to CME forecasts. There are causes for concern on the initial constraining of CME parameters with subjective approaches. A recent study by Verbeke et al. (2023) explored how the subjectivity of the user affects the 3D CME parameters that are obtained from the Graduated Cylindrical Shell (GCS; Thernisien, 2011) reconstruction technique. The GCS croissant-like shape is meant to emulate the geometrical appearance of CMEs by projecting the 3D GCS wireframe onto one, two or more nearly-simultaneous coronagraph images. The parameters of the GCS model are used in heliophysics and space weather research to feed into CME propagation models (e.g., Singh et al., 2018, 2019, 2020; Palmerio et al., 2019; Scolini et al., 2019). Different synthetic scenarios were designed where the "true" CME geometric parameters were known. This allowed Verbeke et al. (2022) to quantify minimum uncertainties associated with the GCS technique, especially in relation to the subjectivity of the particular user performing a fit. This information can then later feed into how we construct a list of ensemble members for CME ensemble forecasting (Mays et al., 2015; Dumbović et al., 2018), a technique which is not commonly used so far and should be explored in more detail for operational purposes. Singh et al. (2022, 2023) performed ensemble modeling based on a constant-turn flux rope model and propagated uncertainties in GCS-derived parameters into simulations. Their results indicated that much of the time-of-arrival uncertainty stems from GCS fitting variability itself. Complementary to ensemble modeling, Singh et al. (2023) also applied ML techniques, including lasso regression and neural networks, to HI J-maps from STEREO. These data-driven methods dynamically refine CME propagation predictions and reduce the MAE of arrival time forecasts by roughly a factor of two (to ~4–5 h). Such hybrid ensemble–machine-learning approaches represent promising pathways to operationally mitigate subjectivity and improve CME arrival time forecasting. Two follow-up studies have been designed, where a set of 23 real events have been fitted multiple times and consequently have been modelled by a large variety of CME arrival time models. This effort has been part of the International Space Science Institute[40] (ISSI) team #480 Understanding Our Capabilities In Observing And Modeling CMEs (Verbeke et al., 2023).

Circling back to the three main components that influence the CME predictions, the community has focused on model development and providing a wide range of different CME models. The community has started to investigate the influence of CME kinematics and magnetic structure and its effects on the CME arrival times. We refer the reader to Temmer et al. (2026) in this issue for a broader review of the status and future of CME observations and model development.

Distinguishing the influence of the ambient solar wind through which the CME is travelling in the heliosphere is a difficult task, which has not been studied in deep detail so far. Raza et al. (2025) provided the first attempt to quantify how uncertainties in the ambient solar wind affect CME arrival time forecasts. Their study applied several machine-learning (ML) models: k-nearest neighbors, support vector machines, and linear regression to discrepancies between observed and simulated solar wind parameters at L1 for 122 CMEs recorded between March 2012 and March 2023 in the CCMC Database Of Notifications, Knowledge, Information[41] (DONKI) catalog. Model inputs were derived from WSA-Enlil-Cone (Odstrčil, 2003; Odstrčil, 2023) simulations and OMNI solar wind observations. Both univariate and multivariate approaches were tested to assess the contribution of individual and combined features to the time-of-arrival prediction accuracy. The best-performing models improved time-of-arrival forecasts by roughly 35 and 47 min with univariate and multivariate approaches, respectively. Although this modest gain suggests that modifications to the background solar wind may not drastically alter CME arrival predic-

[40] https://www.issibern.ch/.
[41] https://ccmc.gsfc.nasa.gov/tools/DONKI/.

tions, it underscores the need to better characterize heliospheric conditions when coupling CME and solar wind models. This also suggests that there is a need for the CME arrival time team and the solar wind team to work together.

4.1.3. Capability assessment and preparation for utilisation of data from the vigil mission

As preparation for the ESA Vigil mission continues, development work is also ongoing to assess the expected improvement in space weather forecasting capabilities which can be expected when these data are available to use in conjunction with data from the Lagrange Point 1 (L1). The ongoing activity on ESA regarding the use of Lagrange Point 5 (L5) data in CME propagation models[42] is to analyse current capabilities and development needs to prepare for the utilisation of these new data out of the Sun-Earth-line when available. The "Use of L5 Data in CME Propagation Models" study is taking a systematic and quantitative approach to assessing the anticipated benefits that the combination of L5 with L1 and near-Earth data will have on operational space weather capabilities. This study is performed considering data from different locations (L1 and the equivalent angular location similar to L5 from STEREO) under 7 use cases (CME onset; CME arrival at L1; use of magnetograph data for modelling background solar wind; use of L1 and L5 to improve CME ensemble modelling; Data assimilation using L1 and L5 in situ and remote data; solar wind feature at L1; SEP propagation analysis). This is performed with the multi analysis of a set of 15 CMEs observed with a combination of 5 experiments, observing CMEs with different instruments and best and degraded cadences and resolutions. The assessment summary showed that extending the coronagraph view from L1 with equivalents from the L5 vantage point resulted in improved CME fits and forecasts. Furthermore, for CME forecasts with the ELlipse Evolution model based on HI (ELEvoHI), which used HI data for both CME initial conditions and CME tracking, the benefit of using higher cadence and higher resolution data was shown. It was also pointed out that using EUV image data from L5 might improve the representation of CME initial conditions. Another key point was on the use of L5 magnetograph data in addition to L1 magnetograph data results showed that the L5 magnetograms could improve predictions at all solar cycle phases, plus the use of assimilated in situ data from L5 reduced the error.

The following discussion provides examples of recent publications considered for the literature review as part of the aforementioned ESA study. In terms of CMEs propagation modelling considering multiple vantage points, most of these show the benefits of space weather missions on vantage points as L5, considering STEREO data as basis for these studies (mostly the coronagraph and the HI) in an equivalent angular location; as in Amerstorfer et al. (2021), they evaluated different setups of the ELEvoHI model, which uses STEREO HI data, with a number of inputs by performing hindcasts for a CME set that occurred when STEREO was near L4/5. The uncertainties in CME propagation kinematics observed with STEREO HI are studied in Barnard et al. (2021), considering CME propagation geometric models, in regards to the performance of the modelled CME arrival time errors and comparison among them, concluding that certain errors are minimised with data addition from the vantage point of L5. Another case study of modelling a CME observed with science data from STEREO-COR and STEREO HI at the similar location around L5, and other observational data from L1 with a number of models is shown in Rodriguez et al. (2020). In Bauer et al. (2021) the effects on low resolution data from STEREO HI compared to the science data (high resolution) was investigated, leading to more accurate forecasts with science data of the ambient solar wind, Turner et al. (2023) demonstrated that assimilating in-situ solar wind observations from L5 and L1 improves solar wind speed hindcasts.

4.1.4. Solar flares

The nature of solar flares means the modeling and hence validation approaches applied to solar flare forecasts are necessarily somewhat different than for various other space weather phenomena. Solar flares bathe the heliosphere in radiation that propagates at the speed of light, emanating over 4 pi from the coronal source. Although immediately responsible primarily for ionospheric disturbances, solar flares also provide a "reference point" in time and space for other energetic phenomena such as CME and SEP production.

Evaluation and validation of solar flare prediction capabilities are generally performed on a statistical rather than case-by-case basis. An overview of the history, approaches, and challenges for forecasting itself and forecast evaluation can be found in Leka and Barnes (2018, Chapter 3). Early validation methodologies were outlined in conjunction with, for example, work at the US NOAA Space Environment Center (SEC, now SWPC) in the context of self-evaluation of flare forecasts, and for evaluating new methods (see Sawyer et al., 1986). The probabilistic forecasts provided by NOAA for solar flares are (very generally) viewed as the "standard" against which to compare. The historical record of these forecasts comprises a number of "event definitions" that are widely adopted by new facilities (for better or for worse, see below).

Research interest in flare forecasting became more widespread with (1) the recognition of solar magnetic fields as possible flare drivers and (2) the availability of statistically significant samples of appropriate data sources. After the research literature began to be populated with wildly differing claims and evaluation methods, a community-

[42] For more information, please refer to https://swe.ssa.esa.int/use-of-l5-data-in-cme-propagation-models.

based series of workshops on "Forecasting the All-Clear" culminated in the first of a series of inter-method comparisons that invoked multiple quantitative skill scores (Barnes et al., 2016). Key take –aways from this study were that (1) even when methods said they were calculating the same quantity (e.g., "total magnetic flux of an active region"), subtle differences in data-handling details could in fact impact the resulting forecast skill scores, (2) sample size was critical, and (3) similar results as measured by one skill score could differ significantly when a different skill score was used. Finally, no method was found to perform significantly better than climatology.

Efforts to improve solar flare forecasting have, in the years since the launch of the SDO, often taken advantage of both the large-sample datasets provided by this facility and some of the pre-computed descriptive *meta*-data that accompany the HMI vector magnetic field data products. Examples of data products from HMI include the Space-weather HMI Active Region Patches (SHARP, Bobra et al., 2014), which comprise a selection of the highest-performing parameters investigated with regards to flare forecasting by Leka & Barnes (2018). Generally speaking, research on improving flare forecasting has focused on SDO data, often invoking different ML algorithms of varying complexity, and relied on "internal" evaluation.

Of note in this context, the European Union based Flare Likelihood And Region Eruption Forecasting (FLARECAST) project (Georgoulis et al., 2021; Campi et al., 2019; Florios et al., 2018) performed a multi-institution multi-pronged approach to improve flare forecasting, and is the largest coordinated project for this type to date. FLARECAST was instrumental in (1) investigating new characterizations of the boundary data, (2) performing head-to-head tests of multiple statistical methods, including numerous ML-based approaches, and (3) assembling subject-matter experts together with end users in order to orient flare forecasting toward the most useful framework possible.

The research and operations communities came together in 2017 for a directed, head-to-head comparison of operationally-running forecasting methods (Leka et al., 2019a,b; Park et al., 2020). The effort was limited to operational flare forecasting (those from government or similar entities), plus a few methods that run autonomously and had high readiness levels, in order to keep the effort focused and tenable. Of note, it was widely agreed that due to timing with regards to solar cycle and SDO data availability, these studies presented deep dives into methodology of forecast comparisons; the sample size that was available is too small for the methods' performances to be used otherwise.

One key take-away from this effort, discussed at length in Leka et al. (2019a), re-enforced the result from Barnes et al. (2016) work: a single evaluation metric is insufficient to gauge performance. The Leka et al. (2019b) paper focused on methods to discern whether differences in performance could be traced to different method particulars, including data, statistical methods, and (importantly) whether there was a "human-in-the-loop". The latter indeed proved advantageous.

In Park et al. (2020), the focus was on key failure modes for present operational forecasts. Two were highlighted: limb flares (where all methods failed due to lack of data), and the "first/last flare challenge", where methods routinely failed to predict the first flare produced by an active region (leading to a "miss") and also routinely failed to predict the first "quiet" (leading to a "false alarm"). Of note, for this latter failure mode, Park et al. (2020) developed a new visual metric by which to directly evaluate a method's performance in this context.

With the growing number of new flare forecasting methods appearing in the research literature (far too many to cite, but most based on SDO data and a majority employing ML techniques), it is time for a similar direct-comparison effort, this time focusing on research-based developments. Indeed, the need for improved solar flare forecasting was included in the most recent NASA Decadal Survey.

To address the challenge called out in the Decadal Survey, and to improve flare forecasting a few specific tasks must be undertaken by the research and operational communities. First, we need to understand how well new research methods can perform as compared to present operational facilities. To do this well is not a small feat; all presently-available flare lists likely contain inconsistencies, especially those that extend to pre-SDO epochs. To establish such a list requires, to begin with, an internationally recognized and curated historical flare event list that is (1) broadly accepted by the research and operational communities, (2) recognizes "historical" as being up-to-yesterday, meaning almost-current in content, and thus (3) supported for development and on-going curation.

Second, it would be illustrative to establish a similar comparison between present operational methods and presently-developing research-level methods as was undertaken in 2017. Based on the NOAA standards for operational flare forecasting, such an exercise could result in a welcoming, community-available validation tool by which new methods could be evaluated in a quantitative manner, and compared to established methods such as have been amassed through the CCMC near-real-time forecast collections.

Third and finally, the communities – both research and operational – need to critically re-evaluate the framework of solar flare forecasting. It is becoming recognized that probabilistic forecasting for solar flares based on the SWPC-type standard event definitions (Table 4.1) may not provide the most actionable information. While specific user needs can be straightforwardly addressed through customized cost/benefit analysis of error types, the physical processes involved do not always scale directly from the peak Soft X-ray flux; flare fluence, hardness, location, or other spectral emission may be more informative for action related to (for example) SEP production, ionospheric

Table 4.1
Standard event definitions from NOAA space weather prediction center (and others).

Flare Class Thresholds	Validity Periods	Latency Periods	Issuance Frequency	Forecast Scope
GOES X-Ray Sensor (XRS) Peak Fluxes of C1.0+, M1.0+, X1.0+	24hr	0hr, 24hr, 48hr	1x/day at 00:00UT (publicly available)	Full-Disk Probabilistic (archive nominally available)NOAA active regions-based (archive nominally available via synoptic drawings)

impact, or CME velocity. Forecasts which focus on extremely short validity periods take on the character of identifying solar flare "precursors" that indicate an imminent (≪1 hr, often ≪5 min) event (e.g., Kusano et al., 2020, Hudson, 2020). Both of these novel frameworks (forecasting for different flare event-types or phenomena, and forecasting for extremely-short-term event probabilities) may have very limited customer bases, but they also present the challenge of possibly needing new evaluation strategies.

4.1.5. Solar energetic particles

Eruptions on the solar surface can accelerate charged particles to near relativistic speeds which can then spread throughout the inner heliosphere to the Earth, Mars, and beyond. These solar energetic particles (SEPs) are a radiation hazard for astronauts (Hu et al., 2009), satellite electronics (Baker, 2000), and, for the most intense events, for passengers onboard airplanes traveling along polar routes (Anderson et al., 2014).

Inspired by scientific interest as well as potential real world impacts, many types of SEP prediction models have been developed in the research and operations communities (Whitman et al., 2023). Some models are built from first principles to reproduce the physics in the solar corona, solar eruptions, and the inner heliosphere to increase our understanding of the key physical processes that accelerate SEPs. Other models aim to produce statistically-derived forecasts via empirical formulations or ML methods, i.e., computationally lightweight approaches that have the potential to generate forecasts quickly in real-time. With the proliferation of SEP models, there arose a need to validate model performance in a quantitative and consistent way.

Starting in 2018 and 2019, an SEP modeling community defined a small set of ten "SHINE challenge" SEP events and invited the community to submit forecasts. This challenge list, dubbed the SEPVAL (SEP Model Validation) 2023 Challenge, was later expanded to 33 SEP events and a nearly equal number of 30 non-event (quiet) periods to enable a more complete validation of each model. An approximately equal sample of SEP events and non-event periods were selected by the NASA Johnson Space Center Space Radiation Analysis Group (SRAG), focusing on the types of SEP events and parent eruptions (flares and CMEs) that are relevant to SRAG operations. Nearly all of the selected periods were associated with strong flares and fast CMEs, except three SEP events for which the flares were not visible from Earth because the source regions were beyond the Sun's western limb. Prior to the SEPVAL effort, many published validation studies focused on SEP events only without testing for quiet periods. A specific aim of SEPVAL was to include SEP events and time periods following solar eruptions that did not result in significant proton flux enhancements at Earth so that the validation would reflect correct predictions, misses, and false alarms. The SEPVAL 2023 Challenge lists are available on Zenodo[43] and through the SEPVAL Challenge webpage.[44]

The SEPVAL challenge is interested in evaluating SEP model performance in a real-time-like workflow, but this is limited by the fact that the available input observations do not always reflect what was actually accessible at the time of each event, since historical data catalogs often lack sufficient version control. We instead focused on providing the best measurements possible. Moon-to-Mars Space Weather Analysis Office (M2M SWAO, see section 9.3) reviewed each CME and produced high-quality 3-dimensional CME parameters with the same tools used to support SRAG operations. The updated CME parameters are saved in the "SEPVAL_CME_CATALOG"' in the CCMC DONKI[45] and included in the SEPVAL event lists. Models also had the option to use 2-dimensional plane-of-sky CME parameters published in the human-derived CDAW CME Catalog[46] or automatically-derived CACTus catalog.[47] In this way, SEPVAL allows cross-comparisons between models using their most appropriate workflows for a standardized set of challenge events.

A comprehensive Rules of Participation[48] was circulated to all SEPVAL participants. One focus of the SEPVAL challenge was to evaluate models as they would perform in an operational setting, so participants were asked to produce forecasts in a "real-time-like" mode without changing parameters, tuning, or recalibrating their models from event to event. Each model developer was asked to indicate which event periods in the challenge list were used in training their model. Specific energy

[43] https://doi.org/10.5281/zenodo.15020585.
[44] https://ccmc.gsfc.nasa.gov/community-workshops/ccmc-sepval-2023/.
[45] https://kauai.ccmc.gsfc.nasa.gov/DONKI/search/.
[46] https://cdaw.gsfc.nasa.gov/CME_list/.
[47] https://www.sidc.be/cactus/.
[48] https://drive.google.com/file/d/1qC9YxkV9pCnNgRFLFuFuwI-ZuL7zyyaTY/view.

channels and thresholds used by SRAG in operations were communicated and recommended (but not required) for forecasts:

- >10 MeV protons exceed 10 pfu (relevant to astronauts during Extravehicular Activities (EVAs))
- >100 MeV protons exceed 1 pfu (relevant to astronauts in a spacecraft)

ISWAT H3-01 Action Team leads assumed the responsibility for leading the development of a generalized validation framework for SEP models, called Solar Particles in the Heliosphere validation INfrastructure for SpWx (SPHINX). The SPHINX validation infrastructure has the capability to prepare observations (e.g., proton measurements), organize forecasts (from a challenge or SEP Scoreboards), perform the validation, and display the results in a variety of formats, including HTML reports and an interactive web interface called VIVID (see Fig. 4.1). From the beginning, SPHINX was developed to have a flexible, generalized approach, with the ability to use observations from a wide variety of satellites as "ground truth" and an appropriate validation workflow for any type of SEP prediction model.

The community challenges played a critical role in the development of SPHINX as a generalized tool. As of March 2025, 21 SEP models have participated in the SEP validation project, covering a wide range of model approaches and forecasted quantities. No two models are exactly alike, requiring SPHINX to be robust to the range of forecasts. Workflows were developed with the goal to validate *the intent* of each model. In other words, many quantities are validated and many metrics are produced by SPHINX with the goal that the user can choose the results that best represent their model's aim. For example, some models may predict the maximum flux during an SEP event (which may be the result of an Energetic Storm Particle event due to the passage of a CME at Earth) while other models predict the initial peak at the beginning of an SEP event (onset peak) that is more closely associated with the acceleration physics low in the solar corona. In SPHINX, metrics are calculated for both the max flux and the onset peak, allowing the user to choose the relevant results.

Throughout the community validation challenges, ongoing communication between the participants and the team leads was key. Team leads worked closely with model developers to understand each model's approach and correctly represent forecasts in the CCMC JSON format, often going through multiple iterations. After a round of validation was completed, the results were circulated to the model developers for their review and approval before being shared publicly. The leads hosted virtual question and answer sessions to clarify how the validation was performed and to review the information in the output files. Discussion sessions at SHINE, ISWAT, mini-ISWAT, and European Space Weather Weeks provided venues to review and critique the validation approach being developed in SPHINX, facilitate communication between researchers and end-users, identify metrics of interest, and better understand caveats like the impact of imbalance of a validation set on the metrics.

Through these efforts over the past 6+ years, the SEP Validation ISWAT Team H3-01 produced the mature SPHINX validation infrastructure and a first estimate of the state-of-the-art of SEP model performance across the research community based on the benchmark SEPVAL Challenge set. As with all performance evaluations, there are caveats in interpreting the results. SPHINX and validation results for SEPVAL and the SEP Scoreboards (see §6.3) are described in a NASA Technical Report (Whitman et al., 2026).

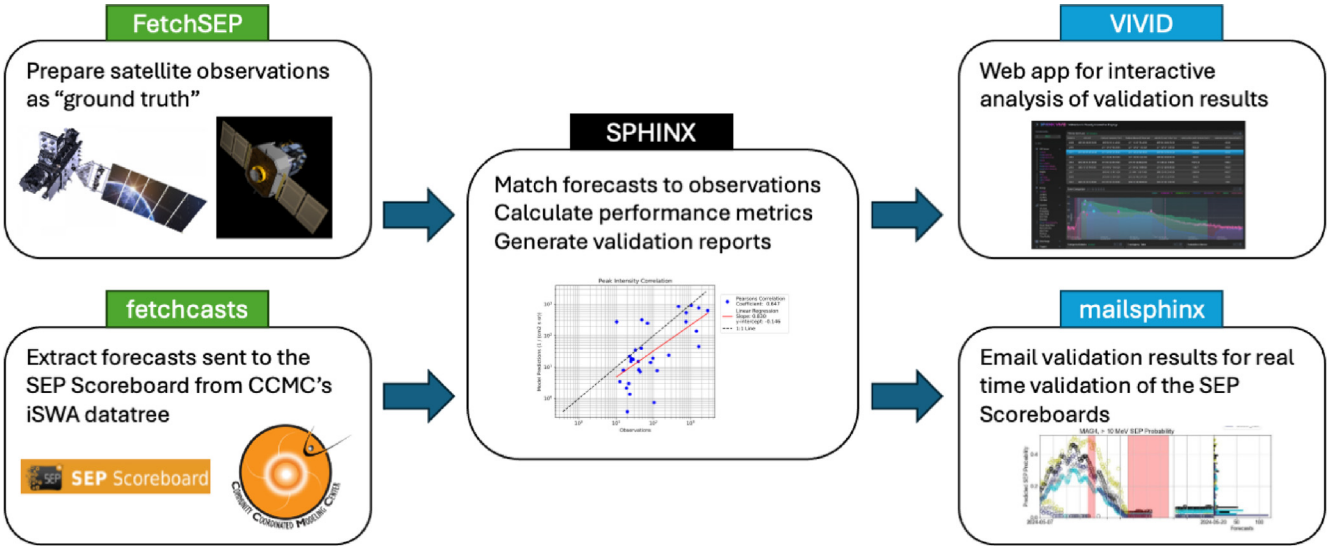

Fig. 4.1. Information flow between components of SPHINX Validation Framework, that support both historic validation and pre-event forecasts called Scoreboards. For discussion of the SEP Scoreboards component for real-time validation, see §6.3.

4.1.6. Thermosphere neutral density

Liemohn et al. (2021) and Chou et al., 2023). outlined standardized approaches for assessing magnetosphere and ionosphere-thermosphere model performance, ensuring consistency and reliability in comparing simulation results with observational data. As Pearson correlation coefficient and RMSE are commonly used in previous validations (e.g., Bruinsma et al., 2018; Kalafatoglu Eyiguler et al., 2019; Shim et al., 2011, 2012, 2014, 2017, 2018, 2023; Tsagouri et al., 2018), metrics are organized into two groups: fit performance metrics (point-to-point) and event detection metrics. Other metrics are introduced including accuracy, bias, precision, association and extremes. Timing error is also considered (Shim et al., 2018, 2023).

The metrics for model-data comparison are also divided into storm times and quiet times (Bruinsma et al., 2021; Bruinsma and Laurens, 2024). To accurately characterize the storm-time response of the thermosphere and ionosphere, the observations and model output need to be debiased to minimize the effect of non-storm-related model differences by normalizing the model and data with pre-storm or monthly mean parameters (Bruinsma et al., 2021). To evaluate baseline (quiet time) model differences, no debiasing is required. The simplest metrics in the thermosphere model validation are the mean observation to model densities and their STD in log scale following Sutton (2018) and Bruinsma and Laurens (2024). The correlation coefficients R of observed and modeled densities are also computed along with the coefficient of determination R^2 which represents the fraction of observed variance reproduced by the model.

The assessment of thermospheric models has been initiated at CCMC within the COSPAR ISWAT framework from 2022. In the first stage, a comprehensive assessment of thermospheric models under geomagnetic storm conditions, defined by an Ap index $\geq$ 80, is conducted. Using models and tools hosted at CCMC, this unbiased assessment examined both empirical and physics-based models during storm periods. After applying a debiasing procedure based on the pre-storm phase, modeled neutral density outputs are compared against high-fidelity observational datasets from the Delft University of Technology[49] derived from CHAMP, GOCE, GRACE, GRACE-FO, and SWARM-A satellites. Key point-to-point performance metrics, including mean density ratios, STD, and correlation coefficients, are used to construct thermospheric model scorecards. These scorecards serve as a valuable resource for identifying the most suitable model for specific applications. The ultimate goal of this effort is to establish a near-real-time thermosphere model assessment capability at CCMC in the near future. An updated thermosphere-ionosphere interactive visualization and validation platform has also been developed at CCMC to facilitate model-data comparisons and provide users with access to scores based on consistent and standardized metrics. The full list of time periods is available online.[50] The modeled neutral density dataset used in the validation study can be accessed online.[51] The thermosphere-ionosphere interactive visualization and validation platform is available at the CCMC.[52]

4.1.7. Ionospheric TEC, foF2 and hmF2

The ionosphere is a dynamic region of the Earth's upper atmosphere extending from approximately 60 km to over 1000 km altitude. It contains a great number of charged particles and exhibits high variability and complexity due to its coupling with both terrestrial weather from below and space weather from above. As a result, the ionosphere demonstrates enigmatic variation with time of day, season, solar activity geographic and geomagnetic location. In addition to regular diurnal cycles, the ionosphere is subject to irregular variations caused by geomagnetic storms, solar flares, irregularities and traveling ionospheric disturbances. The ionosphere is critically important for human activities as it greatly affects the propagation of terrestrial HF radio waves as well as *trans*-ionospheric satellite signals, such as GNSS. Accurately modeling and forecasting the ionosphere is therefore crucial to mitigate its impacts on communication and navigation systems.

Operators of radio communication systems often use the ionosphere to extend the range of communication. The critical frequency relates to the maximum usable frequency, which is the highest frequency that can be reliably used for radio wave communication. For signals transmitted at a frequency below the maximum usable frequency, the radio waves can be reflected by the ionosphere, enabling long-distance propagation. In vertical sounding, foF2 represents the critical frequency of the F2 layer, which represents the highest frequency that can be reflected by the ionosphere which can be derived from the peak density NmF2. hmF2 is the peak density height of the F2 layer that affects the ray path of radio wave signals through the ionosphere. For HF signals, such as those from GNSS satellites, the radio waves pass through the ionosphere but are subject to group delay and phase advance, which are dependent on the TEC along the signal path. Therefore, several ESEQs such as foF2, hmF2, TEC, TEC gradients, ionosphere absorption, and scintillation indices, are useful for evaluating model performance and characterizing the ionosphere's influence on radio wave propagation and GNSS positioning. The ionosphere exhibits different patterns that vary with geomagnetic latitude, time of the day, and the combination of solar and geomagnetic activities. The quality of GNSS positioning is critically affected by the bulk ionisation in the ionosphere (quantified by TEC) as well as by propagation effects originated by ionospheric

[49] https://thermosphere.tudelft.nl/.

[50] https://kauai.ccmc.gsfc.nasa.gov/CMR/TimeInterval/viewAllTI.

[51] https://github.com/j0a8c2k1/CCMC-Thermospheric-Validation

[52] https://ccmc.gsfc.nasa.gov/ionvp/.

irregularities, such as phase fluctuations and scintillation (Chou et al., 2024; Forte et al., 2026).

The CCMC initiated the Coupling, Energetic, and Dynamics of Atmospheric Regions (CEDAR) electrodynamic thermosphere ionosphere challenge in 2009 to evaluate the modeled parameters, including NmF2, hmF2, TEC and vertical ion drift, using Incoherent Scatter Radar (ISR) observations (Shim et al., 2011, 2012). Several metrics, including RMSE, prediction efficiency, ratio of the maximum change in amplitudes, and ratio of the maximum amplitudes (i.e., yield), are utilized to assess the models' performance during quiet and storm time conditions. Metrics based on ratio are used to quantify the model capability to produce peak values or short-term variations during a certain period. Shim et al. (2017, 2018, 2023) evaluated the TEC and foF2 during storm time. They proposed additional metrics, such as RMSE normalized by the mean absolute value of the observed TEC, time difference between the modeled and observed peak time, and correlation coefficient, to evaluate the capability of ionospheric models in simulating the storm effects. Tsagouri et al. (2018) assessed climatological modeling of foF2 and hmF2 using ionosonde data. The assessment was based on the calculation of model scores representative for the association between modeled and observed values, as well as the biases and the accuracy of the predictions, including: the correlation coefficient (R), and the coefficient of determination (R2), for the association; the ME and the STD standard deviation (STD) of the differences between modeled and observed estimates for biases; the RMSE and the MRE)for testing the accuracy. The MRE was suggested to effectively support the comparative analysis of the models' performance over different locations, seasons, and local time. Chou et al., 2023 validated the modeled TEC during the 2013 and 2021 geomagnetic storms using the madrigal TEC observations. Multiple metrics are used to quantitatively assess the models' accuracy, precision, association, bias, and capability in capturing the TEC changes in response to the storms. The skill score based on the metric scores is further proposed to evaluate the overall performance of ionospheric models against the reference model (International Reference Ionosphere 2016; IRI-2016, Bilitza et al., 2017).

Recently, CCMC has initiated a campaign to assess the models' performance during geomagnetic storms across solar cycles 23 to 25. The ISWAT G2B-05 team validated modeled foF2 and hmF2 using the ground-based ionosonde data across ~ 200 storm events.[53] The CAMEL online visualization tool visualizes the ionosonde validation results for the historic storm events.[54] In addition to CAMEL ionosonde validation, CCMC has conducted several ionospheric validation projects by proposing innovative technologies, methodologies and metrics.

4.1.7.1. foF2 and hmF2 validation using GNSS RO observing system simulation experiments (OSSEs). CCMC conducts OSSEs to validate foF2 and hmF2 using the FORMOSAT-7/COSMIC2 RO data. Rather than directly comparing the RO-derived foF2 and hmF2 with model outputs, line-of-sight geometries between LEO satellites and GNSS are inserted into ionospheric models to derive the synthetic RO TEC. Abel inversion is then applied to derive synthetic RO electron density profiles, from which foF2 and hmF2 are extracted (e.g., Chou et al., 2017). This approach mitigates biases introduced by the RO retrieval algorithm and enhances model-data comparisons.

4.1.7.2. TEC validation using ground-based GNSS networks. Global ground-based GNSS networks consist of thousands of continuously operating stations maintained by various organizations for applications including geodesy, navigation, tectonics, ionospheric research, and precise positioning. These dense GNSS networks offer continuous, high-resolution two-dimensional TEC observations. Ionospheric models hosted by CCMC and other institutions are validated by leveraging the dense GNSS networks, providing a comprehensive model validation on a global scale and deeper insights into model performance (e.g., Chou et al., 2023).

4.1.7.3. Ray-tracing and amateur radio validation. Ray-tracing simulation of HF radio wave paths are compared with amateur radio data. In this approach, HF radio wave propagation paths are simulated using ray-tracing algorithms that incorporate electron density profiles derived from ionospheric models (Coleman 1998, 2011, 2017). These simulated paths are then compared against actual signal paths recorded by amateur radio operators worldwide. By analyzing discrepancies between simulated and observed paths, it can evaluate how well different models reproduce ionospheric foF2 and hmF2. This method not only enhances model validation in terms of geophysical parameters but also provides insight into the operational applicability of the models for HF communication and radio propagation.

4.1.8. Near-earth space radiation and plasma environment

The near-Earth radiation and plasma environment consists of diverse particle populations of different origins that often evolve dynamically over time and space and span many orders of magnitude in energy. Such an environment poses challenges from both science and space weather-impact perspectives. It brings about a slew of deleterious effects that could affect spacecraft electronics, materials, and could pose a significant health challenge for life in space. The impacts include surface charging, internal charging, total dose, single event effects, radiation effects for aviation, and radiation impacts on crewed missions. Therefore, this region of space is of great interest to various groups of end users, such as those involved in satellite design, launch, operations, and anomaly resolution; flight

[53] https://kauai.ccmc.gsfc.nasa.gov/CMR/TimeInterval/viewAllTI.
[54] https://ccmc.gsfc.nasa.gov/camel/IonosondePlasmaDensity/home.

crew and passengers (aviation); emergency management personnel; astronauts and space travelers; and stakeholders and policymakers (Zheng et al., 2019; Zheng et al., 2026).

To speed up transitioning science progress to useful information (and eventually better tools and products) for end users and to help bridge the space environment science community and engineering and user community, efforts have been made in defining ESEQs that are directly related to corresponding impact-driven anomalies or translatable into impacts (see Table 11 of Zheng et al. 2026, this issue). Such ESEQs are especially relevant to the Near-Earth Space Radiation and Plasma Environment due to their close ties to end-user needs/concerns. In addition, these ESEQs have to be directly measurable. Model validation based on these ESEQs is the first step in transitioning any model into operations. More details on relevant ESEQs can be found in §3.1.

The integral flux of the greater than 10 keV electrons has been found to correlate well with surface charging anomalies. Therefore, it has been chosen as the ESEQ for surface charging related model validation. The initial model validation work (Yu et al., 2019) on surface charging used the 10–50 keV electron flux as the quantity for comparison. It should be mentioned that the event in Yu et al. (2019) was chosen due to the fact that there was an occurrence of surface charging (an example of impact-driven model validation effort). Different types of metrics were employed to evaluate model performance, including cross-correlation, prediction efficiency, root-mean-square error, prediction yield, and the symmetric signed percentage bias (Morley et al., 2018a), but all were applied to the logarithmic value of fluxes. The validation work of aviation radiation models by Meier et al. (2018) used the ambient dose equivalent rate dH*(10)/dt and the absorbed dose rate in silicon *dDSi*/dt as Impact Quantities (ImpQs) and the simple relative deviation as the metrics. The five dosimetric quantities: absorbed dose in silicon, absorbed dose in tissue, dose equivalent, ambient dose equivalent, and effective dose, serve as the physical quantities used in the validation of the Nowcast of Aerospace Ionizing Radiation System (NAIRAS) performance as well (Phoenix et al., 2024). The recent validation effort relevant to the internal charging environment under the COSPAR ISWAT G3 Cluster activities uses 0.9 MeV and 1.8 MeV electron flux values as ESEQs because they are the measured quantities of Van Allen Probes and they are closely related to internal charging effects (Zheng, 2025). The results are available on CCMC's CAMEL[55] online platform.

The commonly used metrics for flux-based quantities are prediction efficiency, root-mean-square error, prediction yield, and log ratio-based metrics such as median symmetric accuracy (MSA) and the symmetric signed percentage bias (SSPB) advocated in Morley et al. (2018a). However, the use of MSA and SSPB may not be advantageous under some circumstances as cautioned in Simms et al. (2023). Threshold-based binary event analysis metrics have been used by Ganushkina et al. (2019). Other metrics include the normalized difference (e.g., Subbotin and Shprits, 2009, Drozdov et al., 2017; Wang et al., 2020) and statistical metrics as mentioned in Zheng et al. (2019) and Liemohn et al. (2025). As discussed in the above sections, it is important to choose more than one metric to test out multiple aspects of a model's performance.

The ISWAT G3 Cluster[56] has expanded its validation efforts beyond the environment related to internal charging to include surface charging. Validating models within the 'Near-Earth Space Radiation and Plasma Environment' faces common challenges as the others, particularly in acquiring well-vetted, high-quality data for both boundary/initial conditions and model-data comparisons. Uncertainty quantification presents another significant challenge. While our internal charging validation guidelines establish a framework for fair "apples-to-apples" comparisons—such as using consistent datasets for boundary conditions and excluding validation-related measurements—certain model-specific variations persist. These include differences in wave types and diffusion coefficients. At this stage, we believe the community benefits more from broader participation of models in the validation process rather than viewing these variations as problematic. The diversity of approaches contributes valuable insights that might otherwise be lost through overly rigid standardization. Another non-uniformity when comparing performance of different models is the non-trivial aspect of extracting fluxes along a satellite trajectory. Unlike a simple interpolation, such a flux extraction procedure requires a magnetic field model which could introduce another source for differences and uncertainties.

4.2. Highlights of validation studies from roadmap special issue 1 "science and applications"

In the solar and heliospheric domain:

- Reiss et al. (2023) worked to unify the validation of coronal and heliospheric models with agreed-upon time periods, metrics, and metadata standards.
- Tiburzi et al. (2023) validated heliospheric models through pulsar observations, comparing the IPS 3-D reconstruction technique with pulsar dispersion measures, and found good comparative results.
- Shaifullah et al. (2023) compared pulsar observations to EUHFORIA, an MHD heliosphere modeling framework, and found a strong correlation for solar wind density along a given line of sight.

[55] https://ccmc.gsfc.nasa.gov/iswat/G3-04/#validation-campaign-in-the-ccmc-camel-framework.

[56] https://iswat-cospar.org/g3.

- Iwai et al. (2023) combined observations of IPS from two different instruments with the SUSANOO-CME model and demonstrated that the resulting MHD reconstruction of a combination of merged CMEs (found previously to be impossible using this model without the incorporation of IPS data) was more accurate with the combined IPS data than with data from a single instrument.

In the ionosphere-thermosphere domain:

- Tsagouri and Belehaki (2023) report on the validation test of the Solar Wind driven autoregression model for Ionospheric Forecast (SWIF), which uses IMF observations at L1 as a proxy for upcoming ionospheric storm-time disturbances. They evaluate the efficiency of the SWIF model in providing alerts before upcoming ionospheric storm effects and discuss challenges in forecasting ionospheric responses from L1 input.
- Abbas and Ameen (2023) evaluated the International Reference Ionosphere extended to the Plasmasphere (IRI-Plas 2020) model using different ionospheric and solar proxy indices to predict the critical frequency of the F2-layer over the Pakistan and Japan longitudinal sectors. Their analysis showed that the model performed best with the solar proxy Mg-II among all other solar and ionospheric indices.
- Bruinsma et al. (2023) described the main thermospheric density datasets of this century and compared them to quantify the differences. They found large discrepancies among the datasets, sometimes in the tens of percent, at all altitudes.

4.3. Statistical approaches to validation

Event-based comparisons between modeling and observations for short time intervals driven by a particular set of initial and boundary conditions suffer from several issues. First of all, there is significant uncertainty associated with the initial and boundary conditions. For instance, global geospace models typically use upstream boundary conditions provided by the NASA OMNI[57] dataset, which has been obtained by combining plasma and magnetic field data from multiple spacecraft orbiting at the Lagrangian L1 point and propagating it to the nose of Earth's bow shock (King and Papitashvili, 2005). However, some plasma measured at L1 might not reach the Earth or its properties might change significantly between L1 and the magnetopause (Borovsky, 2018; Walsh et al., 2019). Additionally, parts of geospace have high Reynolds numbers (El Alaoui et al., 2013; Wrench et al., 2024) and can exhibit chaotic behavior such that accurately modeling the timing and location of all dynamic structures is a major challenge. To mitigate these issues, *statistical* data-model comparisons have become popular in the community, aided by the strong increase in computer power that has become available in recent years. Different approaches have been followed and are discussed below.

One flavor of statistical data-model comparison involves analyzing data of several simulation runs corresponding to different time intervals and events or studying long runs where the system is driven by a large sample of solar wind conditions. For instance, Ridley et al. (2016) used 662 simulations obtained by several global magnetohydrodynamic (MHD) models with different simulation settings produced at the CCMC over the years to analyze 2503 satellite trajectories and statistically compared the simulated magnetic field with that obtained by spacecraft observations at geosynchronous orbit. They produced probability distributions of the error in each magnetic field component and used them to develop a 1–5 star rating system for the models. The rating was based on the model result compared to the median and standard deviation of the error from all model results of the particular data set. They concluded that models performed better for lower levels of geomagnetic activity and that coupling global MHD models with ring current models for the inner magnetosphere produced statistically better results. Further, they showed that with enough events, models could be statistically differentiated from each other when assessing their performance. Kubyshkina et al. (2019) studied an 11-day period (comprising a 5-day interval including the March 2015 storm and a 6-day interval including the June 2015 storm) that contained different solar wind regimes and different levels of geomagnetic activity. They performed a statistical comparison of the magnetic field obtained by observations, by empirical and adapted models, and the physics-based SWMF/Geospace (global MHD model BATSRUS coupled with Ridley ionosphere electrodynamics model and Rice Convection Model for the ring current, Tóth et al., 2005, 2012). Although the performance of the different models depended on the level of geomagnetic activity and on different magnetospheric regions, one of the important results of the study was that the SWMF performance was similar to that of the empirical models under moderate to highly disturbed conditions of the June 2015 event. Al Shidi et al. (2022) performed 122 simulated storms in the period 2010–2019 and analyzed statistically the simulated ground magnetic field perturbations versus observations at several locations world-wide. They concluded that regional predictions at mid-latitudes are quite accurate, with a median Heidke Skill Score of ~ 0.5, while regional high-latitude predictions are more challenging. Al Shidi et al. (2024) statistically studied the time-shifting errors associated with methods that propagate solar wind properties measured at L1 to the bow shock nose. They performed 123 simulations of storm events with the SWFT in geospace configuration and studied the probability distribution of errors between modeled and observed geomagnetic indices (SYM-H – high resolution counterpart of Dst, auroral indices AE, AU, AL, and cross polar

[57] https://omniweb.gsfc.nasa.gov/ow.html.

cap potential) as a function of different solar wind driving and propagation parameters. They found that the magnitude of the errors depends on the propagation parameters (for instance, the accuracy of the SYM-H prediction depends on the spacecraft perpendicular distance from the Sun-Earth line).

Statistical data-model comparisons have also been performed via ensemble modeling (Murray, 2018; Morley, 2020), whereby model inputs or model parameters are slightly perturbed or multiple models are combined together to produce a distribution of future states of a dynamical system. This approach has been used to improve model forecasting but also to provide uncertainty quantification. Cash et al. (2015) performed 39 simulation runs with the Wang-Sheeley-Arge (WSA)-Enlil operational model to assess the effect of input parameters on the arrival time of an unusually fast and strong CME, identifying needed modifications to improve the model forecasting ability. Chen et al. (2018) performed 21 simulations using the Rice Convection Model-Equilibrium to study the effect of the electric field boundary conditions and identified the need to improve the particle loss terms in the model. Morley et al. (2018b) used the SWMF to perform 40 simulations of the April 2010 geomagnetic storm with different solar wind inputs and assessed the uncertainty in the model predictions of SYM-H and ground magnetic perturbations at selected locations. In the context of radiation belt modeling, Camporeale et al. (2016) used a two-dimensional (pitch angle and energy) quasi-linear Fokker-Planck model to study the effects of uncertainties in the input parameters associated with the Kp index, the latitudinal extent of the chorus waves that drive electron acceleration and the ambient electron density, highlighting the critical role played by the time-dependent electron density in the model. Thompson et al. (2020) and Watt et al. (2021) performed ensemble modeling to conclude that the distribution and time variability of the diffusion coefficients employed to model wave-particle interactions are critical to radiation belt modeling. Hua et al. (2023) studied relativistic electron acceleration by chorus waves for the 9 October 2012 storm. They performed ~ 14600 two-dimensional quasi-linear Fokker-Planck simulations varying four input parameters: lower-band chorus wave amplitude and peak frequency, and background magnetic field and ambient electron density. The distribution of input variables was sampled through a superposed epoch analysis of 11 storms that were similar to the chosen storm. They concluded that uncertainties in the input parameter significantly affected electron acceleration, with cases where the simulation output differed from observations by four orders of magnitude. Chand et al. (2025) ran Global Ionosphere Thermosphere Model (GITM, Ridley et al., 2006) 512 times during solar min and solar max conditions to explore how uncertainties in chemical reaction rates affect the nitric oxide density in the thermosphere and the resulting temperature structure.

Several authors have also performed what we will refer to as *behavioral* comparisons between modeling and observations, focusing on quantities that are less susceptible to the precise timing and locations of magnetospheric structures and dynamics or performing a comparison against data-driven empirical models. For instance, White et al. (2001) compared autocorrelation functions of plasmasheet density and magnetic field components obtained from MHD simulations against the same quantities derived from ISEE-2 data. El Alaoui et al. (2013) successfully compared power spectral densities and probability distribution functions of magnetotail turbulence obtained from MHD simulations and THEMIS data. Gordeev et al. (2015) compared several MHD models available at CCMC against empirical relations obtained from observations for several macroscopic parameters of the Earth's magnetosphere (including the subsolar magnetopause distance, the plasma sheet thermal pressure, and the cross polar cap potential drop), concluding that no model consistently emerged as the best in all metrics. Haiducek et al. (2020) performed a 1-month long MHD simulation which featured over 100 substorms to show a distribution of substorm waiting times comparable to that of observations and a weak (but statistically significant) skill of the model in predicting substorms.

Finally, Delzanno and Borovsky (2022) advocated for a new type of validation approach based on systems science, i.e., the science that studies the holistic behavior of systems through their interconnected subsystems. The idea is to apply systems science tools (such as correlation analysis, information theory, etc.) to models and compare them side by side with the application of the same tools to actual observations, aiming to answer the question 'Does the simulation behave in the same manner as the real system behaves?'. As such, these ideas involve a behavioral comparison between modeling and observations that can also be performed statistically. Delzanno et al. (2025) analyzed a month-long simulation dataset obtained by the Multiscale Atmosphere-Geospace Environment (MAGE) model (Zhang et al., 2019; Sorathia et al., 2020). They applied canonical correlation analysis (CCA) to construct solar wind driver and Earth's geomagnetic response canonical functions that maximize the linear Pearson correlation between each other. They found that the MAGE model was able to reproduce quite well the behavior of the magnetosphere and the intercorrelation between the input variables used in the investigation [more so for quieter time intervals, consistent with the result of Ridley et al. (2016). They also found that high resolution SuperMAG (Global Ground-Based Magnetometer Initiative, Gjerloev, 2009) counterparts of auroral indices (Bergin et al., 2020) should be used to quantify future model improvements. This result is also consistent with findings from Haiducek et al. (2017). It is worth emphasizing that the biggest advantage of CCA in this context is that it validates models at a system level, both from the perspective of the solar wind driving and of the Earth's geomagnetic response.

5. Modeling physical phenomena: how to validate?

In this section we discuss the extension of validation from time series parameters to validation of modeling of physical phenomena and what metrics may be of use in such a validation. One expression of this might be thought of as the N-dimensional equivalent of the event-based metrics for the one-dimensional time series discussed in Section 3.3. As a worked example one might try to isolate CME "objects" in 3D model output, and equivalents in 2D HI imagery, (or 3D output derived from IPS), attempt to compare their spatiotemporal association, and validate this statistically over many simulation time steps, and many events. In this discussion, the term "metric" is used in the general sense of an objective, quantitative measure to validate physical phenomena and not necessarily in the sense of a mathematical metric satisfying positivity, symmetry, and the triangle inequality. We use the example of coronal hole (CH) segmentation (delineation) as a concrete example here, considering methods of binary comparison between modeled open field regions and observationally derived CHs. The discussion is likely applicable to validation of a range of physical phenomena, however, and is (and should be) extended to non-binary comparisons in many cases.

We can consider the modeled open field regions and observationally derived CHs as two binary segmentations, a partition of the 2D space into regions based on binary class membership, e.g., whether a pixel belongs to an open or closed flux region in models or a dark or light region in EUV images. Here, the foreground (FG) is considered membership in the class of interest (e.g., a CH) and the background (BG) is considered the remaining regions. Comparison of two segmentations can thus rely on base measurements related to cardinality of pixel-level agreement on class membership, usually characterized by the contingency table entries of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The definition of the contingency table generally assumes the existence of a ground truth established through direct measurement of the underlying state. In many applications, however, including CHs, there may be no directly measurable ground truth. In such cases, one of the outputs (model or observation) can be used as the reference "truth." It is important to acknowledge in such cases that various metrics are describing comparison between two predictions that may both be unrepresentative of the true underlying state. In this case, the contingency table entries are TP (pixels that are FG in both segmentations), TN (pixels that are BG in both segmentations), FP (pixels that are BG in the reference segmentation but FG in the other), and FN (pixels that are FG in the reference segmentation but BG in the other).

Selection of what metrics will adequately describe a comparison between the two segmentations is task specific and requires careful consideration of what characteristics are important for the specific application domain, c.f., Fig. 5.1. Metrics are designed to measure certain characteristics (e.g., areas or boundaries), but may also have biases or sensitivities to other characteristics. Understanding what a metric is quantifying and the limitations of that quantification is important to correctly interpret validation of physical phenomena. Furthermore, the same metric may have multiple different names (e.g., true positive rate and recall) or may have multiple different definitions with the same name (Taha and Hanbury, 2015). We therefore advocate for a careful and deliberate study of appropriate metrics for validation of physical phenomena, similar to those that have been undertaken in other domains (e.g., Taha & Hanbury (2015) in medical image segmentation and Gilleland et al. (2009) in terrestrial weather forecasting). To that end, the S2-08 action team within COSPAR

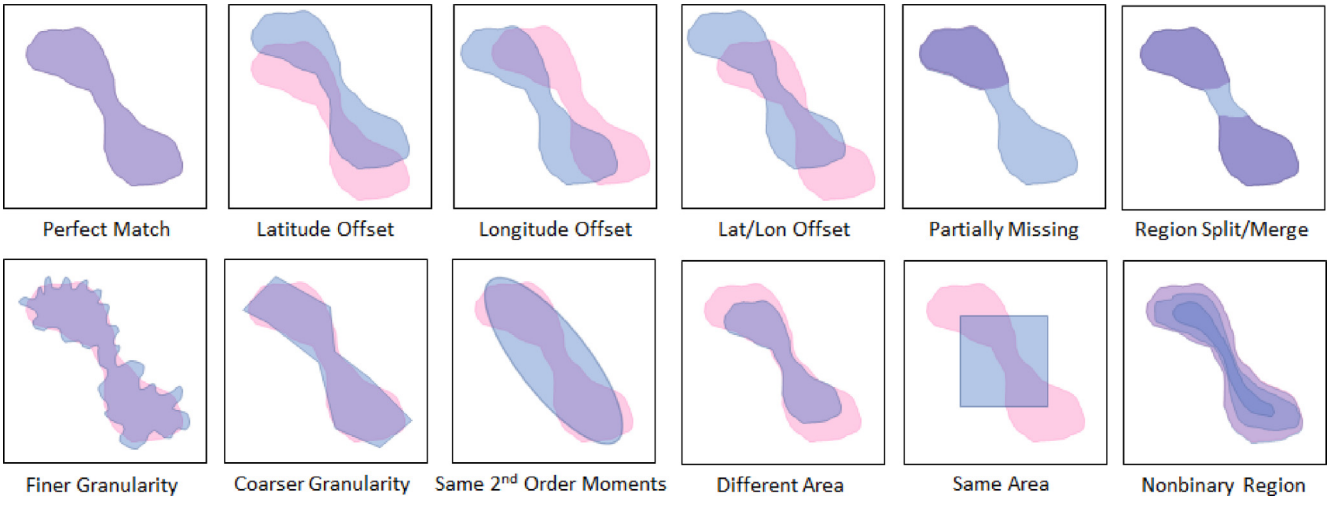

Fig. 5.1. Illustrative examples of comparing two 2D segmentations. In this figure one segmentation is illustrated in blue and the other in pink; areas of overlap will thus appear purple and background areas are white. Differences between segmentations may have different impacts or interpretations depending on the application area, e.g., whether preservation of shape or area is more important.

ISWAT was formed in February 2025 to begin this process for the specific case of validating CH segmentations. As briefly introduced below, research in medical image segmentation, terrestrial weather forecasting, and image quality can be brought to bear on the problem of validation of CHs.

Medical image segmentation has encountered and proposed solutions to many of the same issues faced in validation of physical phenomena in solar data, including different practical implications for different errors (false positive versus false negative), imbalanced data, large data volume, and even a lack of ground truth. Common metrics for binary segmentation were categorized by Taha and Hanbury (2015) as based on spatial overlap, area, pair-counting (considering all pairs of disjoint objects between segmentations), information theory (how much information one segmentation reveals about the other), probability (based on distributions of the pixels in each segmentation), or spatial distance. Notably, all 20 common metrics studied in Taha and Hanbury (2015) can be defined in terms of the contingency table entries, can be extended to non-binary class membership, and can be defined for 2D or 3D data. Additionally, Taha and Hanbury (2015) studied the correlation between metrics and intrinsic properties of the metrics to develop a set of guidelines for selecting segmentation metrics.

Terrestrial weather forecasting has proposed solutions to similar issues faced in space weather forecasting, including the "double penalty" associated with forecasts spatially offset from observations (Gilleland et al., 2009), the importance of quantifying skill over chance forecasts, and the presence of substructures of interest (e.g., storm cells) in forecasts. As accurate identification of CHs is important for precision of highspeed solar wind estimates, accuracies of forecasts using CH segmentations may be of interest. In this scenario, the terrestrial weather community has extensively studied a problem analogous to CH validation, namely the validation of gridded weather forecasts (Gilleland et al., 2009). Gilleland et al. (2009) categorize metrics as filtering methods (to characterize the extent to which forecasts and observations match at different spatial scales) and displacement methods (to characterize spatial displacements necessary for forecasts and observations to match).

The image processing community has active research in quantification of similarity between images in applications such as image enhancement, compression, and image generation. Full-reference image quality metrics are metrics designed to quantify similarity (or difference) between a given image and a reference image. The mean relative error, (or the related measure of peak signal to noise ratio, PSNR) is an easy to compute full-reference image quality metric but is not well-aligned with human perception (Wang et al., 2004 and references therein) and suffers from sensitivity to small alignment errors, i.e., the "double penalty." The Structural Similarity Index Measure (SSIM) is a widely used metric formulated to better quantify similarity in structure between two images (Wang et al., 2004). Recently, the Learned Perceptual Image Patch Similarity (LPIPS) metric, derived from features of deep convolutional neural networks used in image classification, has become a common image quality metric shown to better align with human visual perception (Zhang et al., 2018). While these metrics are largely designed to align quantitative metrics with human opinion of similarity, they may be worth considering for validation of solar phenomena due to their ubiquity and especially in cases where current validation strongly relies on qualitative comparisons.

A collaboration within the community will be necessary to establish what characteristics are important for CH validation, which metrics best quantify those characteristics, and whether other metrics are needed. We advocate for a study modeled after Taha and Hanbury (2015) with consideration of the forecast metrics in Gilleland et al. (2009) and common image quality metrics such as SSIM (Wang et al., 2004) and LPIPS (Zhang et al., 2018).

There is a mutually-beneficial relationship between modeling, theory, and observations, each serving a critical role in understanding complex physical phenomena. Good examples of this are the so-called crescent-shaped particle distribution functions which show the local particle velocity configuration in phase space. These distribution functions not only serve as validations of the theory and modeling, but also can be reliable indicators of important energy conversion processes in the magnetosphere such as magnetic reconnection. They are usually observed when electrons and ions decouple from the magnetic field in regions called electron and ion diffusion regions. A typical distribution consists of two parts: a symmetrical core distribution in the center and a crescent-shaped element for larger velocities, formed by particles undergoing a magnetic field reversal (Fig. 5.2, panel a). Although expected early on by Lee et al. (2004) using test particle simulations, they were rigorously predicted using particle-in-cell simulations and analytical theory by Hesse et al. (2014). The authors showed that such distribution functions should be expected at the electron diffusion regions in asymmetric reconnection, places where the magnetic fields reconnect on the day-side magnetopause. The Magnetosphere Multi-Scale (MMS) mission was designed to look for and resolve such predictions near magnetic reconnection regions. MMS verified the existence of crescent-shaped distributions at the magnetopause, just as the theory and simulations predicted (Burch and Phan, 2016). The changes to crescent signatures based on spatial location were then explored in PIC simulations by Bessho et al. (2016) and the minimum necessary conditions for their development were studied by Lapenta et al. (2017). This specific example demonstrates how numerical modeling and analytical theory often serve as essential tools, anticipating phenomena that only later are validated by observations.

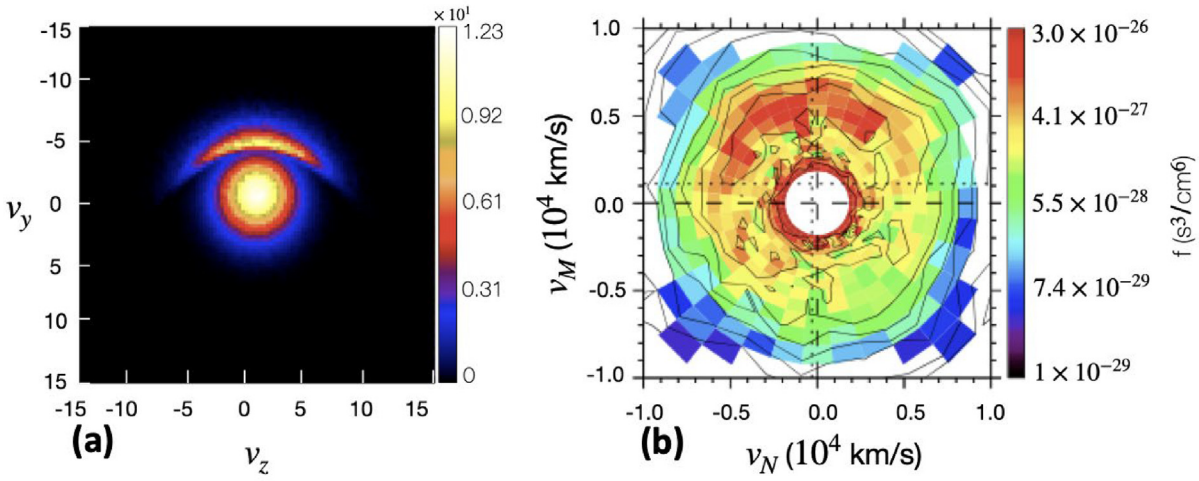

Fig. 5.2. Comparison of crescent-shaped distribution functions between numerical modeling and observations. Panel (a) shows the reduced electron distribution at the flow stagnation point in a particle-in-cell simulation (adapted from Hesse et al., 2014). Panel (b) shows the electron distribution function observed by the MMS at the magnetopause, displayed in LMN boundary-normal coordinates (adapted from Burch and Phan, 2016).

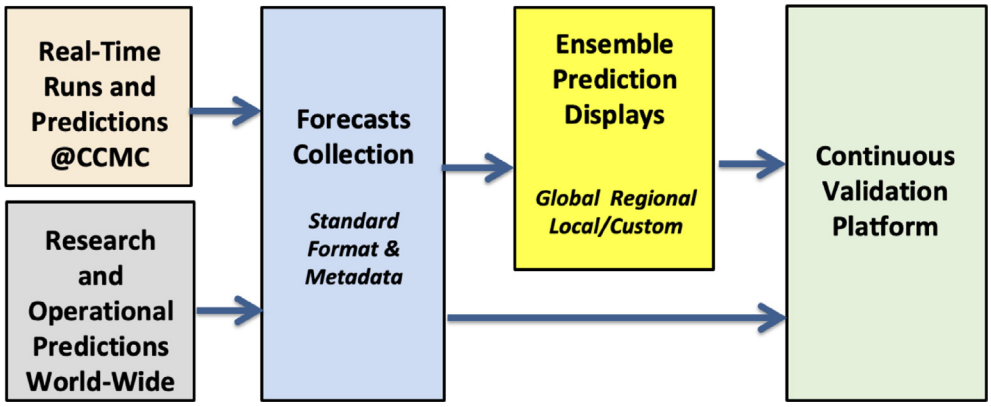

Fig. 6.1. Conceptual design for community pre-event ensemble prediction and continuous validation system.

6. Pre-event community ensemble predictions

Traditionally, space weather models are assessed based on historical events. While valuable for model development and benchmarking new capabilities, such an approach does not reflect the realistic demands of real-time forecasting. The idea for community-wide pre-event ensemble prediction was put forward more than 10 years ago. Here "ensemble" refers to a group of predictions from diverse sources (multi-model ensemble) rather than a statistical ensemble of forecasts (see Appendix B). A conceptual design of the project originally named "Scoreboard" is illustrated in Fig. 6.1. The philosophy behind the project is to build a community product that collects and displays an ensemble of forecasts derived from different research and operational space weather models world-wide before an event occurs. The scoreboards also enable world-wide community involvement in real-time predictions, foster community validation projects, and ultimately help researchers improve their forecasts. All modelers and experts world-wide are welcomed to participate in these community efforts. Databases of forecasts in standard format enable systematic and unbiased evaluation of the predictive abilities of state-of-the-art models developed by research groups worldwide using different modelling approaches and analysis tools. A key objective of the Scoreboard effort is to bring the space weather community into the position to identify new emerging capabilities and models with high potential for operational usage and to demonstrate capabilities to operational agencies and users. Participation in the Scoreboard activities provides an objective measure of model maturity and readiness for forecasting. Participating models inherently increase their Readiness Level (RL; see §7) by demonstrating skill in a realistic forecasting scenario.

With a sense of urgency to start collecting pre-event predictions as soon as possible, the main focus of the Scoreboard project was on collecting forecasts in a standard format and building ensemble forecast tables and displays with the view that the collected forecasts are available for

comprehensive validation by the community. A statistically significant number of forecasts would need to be collected before a community validation or automated cumulative validation platform calculating skill scores would be meaningful. The CME Arrival Time Scoreboard (see §6.1 for more details) was the first community ensemble forecast project to be deployed in 2013. The CME Arrival Time Scoreboard included a sortable table of forecasts with a "score" (difference between predicted and actual arrival time) per event. This design inspired the name "Scoreboard", that was also applied to Flare and SEP community ensemble displays (see §6.2 and §6.3). The first comprehensive CME Scoreboard validation was performed by community members Riley et al. (2018), and recently updated by Kay et al. (2024). Therefore, the Scoreboards have focused on collection, display and demonstrating the operational potential of different methods, and allowing community validation studies. While observational data is also visible and a per-event comparison can be made, the Scoreboards have remained separate from skill scoring validation platforms, which the "Scoreboard" name implies. The lack of cumulative skill scores in the Scoreboards has triggered confusion and some community voices to rename the project or add scores. However, both the standardized collection/display and validation are equally important and valuable, with one fully dependent on the other. As we have not reached a naming consensus, in this paper we will continue referring to the project as a Scoreboard. These concepts are now being brought together for the first time in the form of the SEP Scoreboard, SPHINX Validation Framework, and VIVID Validation Display (see §4.1.5 and Fig. 4.1). Following this example, it is now a priority for every Scoreboard to also have a companion validation platform, which is a clear community need.

By tying together the space weather community and fostering worldwide engagement in real-time predictions, these Scoreboards have contributed to improving forecasting capabilities of CMEs, Flares, and SEPs. While the CME Arrival Time Scoreboard is developed and led by the CCMC, Flare and SEP Scoreboards are examples for community-led initiatives, supported through the CCMC's infrastructure and coordination.

In what follows, we provide an overview of the status of existing Scoreboards and present the vision of the newly formed Scoreboard teams. Beyond these ongoing projects, several additional Scoreboards are currently in development, including the Solar Indices, IMF Bz, Solar Wind, and Geomagnetic Indices Scoreboards. All Scoreboards forecast submissions have a standard format and forecasts are publicly available via an API, which can be used according to the CCMC and Scoreboard[58] Rules of the Road.[59]

6.1. CME arrival time scoreboard

The CME Arrival Time Scoreboard[60] captures manually entered forecasts of CME arrival time, error bar, confidence in arrival (hit likelihood), and geomagnetic storms parameters (i.e., Kp). The CME scoreboard has inspired many retrospective publications of "problem" forecasts and workshop session discussions, and the forecasts have been validated by Riley et al. (2018) and Kay et al. (2024). Since March 2013, there have been over 160 registered users and over 4400 arrival-time predictions using 46 active unique prediction methods for over 700 CMEs.

Riley et al. (2018) found a mean absolute error of 13 h, and a standard deviation of 15 h for a subset of models that had the most forecast submissions. Kay et al. (2024) followed up by using the accumulated larger sample size and found similar results, yielding a mean absolute error of 13.2 h and standard deviation of 17.4 h, with models only showing slight improvement over time. An API for the CME Scoreboard was developed in 2024.[61] CCMC, the UK Met Office and the CME Arrival Time working team (Verbeke et al. 2019) have developed a draft JSON schema that will soon allow for automated submissions and the gathering of more forecast metadata. Additional future plans include (1) an automated validation framework and display based on the validations code used in Kay et al. (2024), (2) capturing CME all-clear predictions, (3) capturing missed CME predictions.

6.2. Flare scoreboard

The Flare Scoreboard[62] contains both research and operational probabilistic flare forecasts for the next 24 h. Forecasts are received in a fully automated manner and displayed as a full disk forecast probability time series and table. A new standard JSON forecast submission format replaces the original Extensible Markup Language (XML) and text formats. Additionally, there is an interactive solar disk display where the user can click on an active region to view the specific forecasts for the probability of that region flaring. The initial version of the Flare Scoreboard application is integrated into the CCMC ISWA system, however the Flare Scoreboard front-end display is being renovated to match the SEP Scoreboard (see §6.3), based on user feedback, in addition to adding more flare models.

6.3. SEP scoreboard

The SEP Scoreboard[63] was initiated by The Royal Belgian Institute for Space Aeronomy (BIRA-IASB) and the

[58] https://ccmc.gsfc.nasa.gov/scoreboards/.
[59] https://ccmc.gsfc.nasa.gov/static/files/CCMC-RulesOfTheRoad.pdf.
[60] https://kauai.ccmc.gsfc.nasa.gov/CMEscoreboard/.
[61] https://ccmc.gsfc.nasa.gov/scoreboards/cme/#cme-scoreboard-web-service-calls-api.
[62] https://ccmc.gsfc.nasa.gov/scoreboards/flare/.
[63] https://ccmc.gsfc.nasa.gov/scoreboards/sep/.

UK Met Office together with CCMC in 2016. In 2018, Johnson Space Center's Space Radiation Analysis Group (SRAG) became involved as an expert user of the SEP Scoreboard as part of the Integrated Solar Energetic Proton Alert/Warning System (ISEP) project, a collaboration between NASA SRAG, CCMC, and M2M SWAO. As part of ISEP, over 10 SEP models have been implemented from the research community and are now available on the SEP Scoreboard display in real-time. The SEP Scoreboard automatically displays and ingests forecasts of SEP onset, duration, peak flux, probability, all-clear, and overall profile. ISEP is actively evaluating the SEP Scoreboard models and iterating with model developers on improvements. The SEP Scoreboard is currently focused on real-time forecasts including all-clear forecasts, and displays diverse information, creating a unique challenge for model output gathering, display, and validation. The CCMC is collaborating with SRAG to construct the SEP Scoreboard displays such that they will be useful for future human deep space missions to Moon and Mars.

In order to reinforce the idea of a community standard, the SEPVAL challenge (see §4.1.5) adopted the same JSON schema required for the ingestion of forecasts into the SEP Scoreboard.[64] In this way, models participating in the SEPVAL challenge are brought a step closer to contributing to the SEP Scoreboard while the developed validation infrastructure has the capability to validate forecasts submitted to SEPVAL as well as those made in real-time to the SEP Scoreboard.

6.4. Solar indices scoreboard

Solar X-ray and EUV radiation is absorbed in the Earth's atmosphere, driving both ionisation and heating of the upper neutral atmosphere. Increased ionisation levels at various ionospheric altitudes can lead to local instability mechanisms from which irregularities and, hence, propagation disturbances capable of degrading GNSS positioning quality can arise; an increase in heating will result in enhanced satellite drag during times of high solar activity. Many ionosphere-thermosphere models are capable of using measured spectral information to drive the models, as well as modeled EUV and FUV spectra based on solar indices (e.g., F10.7 solar radio flux at 10.7 cm, Mg II core-to-wing ratio and the solar sunspot number, SSN).

The ultimate goal of the ISWAT action team S2-02[65] is to create a real-time comparison scoreboard of publicly available model predictions of solar indices, along with ensemble community model predictions. The continuous comparison of different models and ensembles for selected solar indices and EUV bands (e.g., corresponding to GOES observations), provided by such scoreboards, will facilitate a basis for transitioning research-class models towards more operational products to meet user specific applications.

The Solar Indices Scoreboard will be hosted at the CCMC. For the first prototype of the scoreboard it is planned to include established solar indices forecasts provided by the Air Force Research Laboratory Solar Indices Forecasting Tool (SIFT; Henney et al., 2015; Kniezewski et al., 2024) model currently available online[66] and real-time forecasts from SWPC. An example comparison between SIFT F10.7 1, 3, and 7 day forecasts with error bars and real-time observations is highlighted in Fig. 6.2. Future implementation of other indices (e.g., S10, F30) as well as EUV spectral bands are being developed and additional national and international participation are presently being negotiated.

6.5. Planning for IMF Bz scoreboard

The southward component of the interplanetary magnetic field (IMF Bz) Scoreboard represents a crucial advancement in space weather prediction, addressing one of the most challenging aspects of forecasting—the accurate prediction of the Bz magnetic field component – its intensity, but also the cumulative duration of southward excursions. Other solar wind drivers are also important, but the Bz component plays a dominant role in determining geomagnetic storm intensity and potential impacts on Earth's technological infrastructure (e.g., Dungey, 1961; Gonzalez et al., 1999; Wang et al., 2003; Richardson and Cane, 2011; Vourlidas et al., 2019). Recent research efforts have extended earlier approaches and explored a variety of techniques, including physics-based models, empirical methods, and ML approaches, to predict this elusive parameter. For examples of pattern-matching using L1 observations see e.g., Riley et al. (2017), Owens et al. (2017) for the general case, with the CME-specific case in Chen et al. (1997), and recent ML extensions in e.g., Reiss et al. (2021b) and Riley et al. (2023). Attempts to extend the forecast lead time based on remote-sensing of the sun or near-sun domain include Savani et al. (2015, 2017); Kay et al. (2017), Möstl et al. (2018); examples exploring achieving the same using "sub-L1" in-situ observations closer to the sun than L1 include Kubicka et al. (2016); Good et al. (2019); Davies et al. (2024); Lugaz et al. (2025); Weiler el al. (2025).

The IMF Bz Scoreboard,[67] led by Predictive Science, University of Reading, Applied Physics Laboratory John Hopkins University, and CCMC, is in early planning phases. There are a dozen IMF Bz prediction models, some MHD models, others driven by data, and pattern recognition models. IMF Bz predictions from some models already being streamed to ISWA (e.g., Heliosphere Tomog-

[64] https://ccmc.gsfc.nasa.gov/scoreboards/sep/example-sepsb-json/.
[65] https://iswat-cospar.org/S2-02.
[66] https://gong.nso.edu/adapt/sift/.
[67] https://ccmc.gsfc.nasa.gov/scoreboards/imf-bz/.

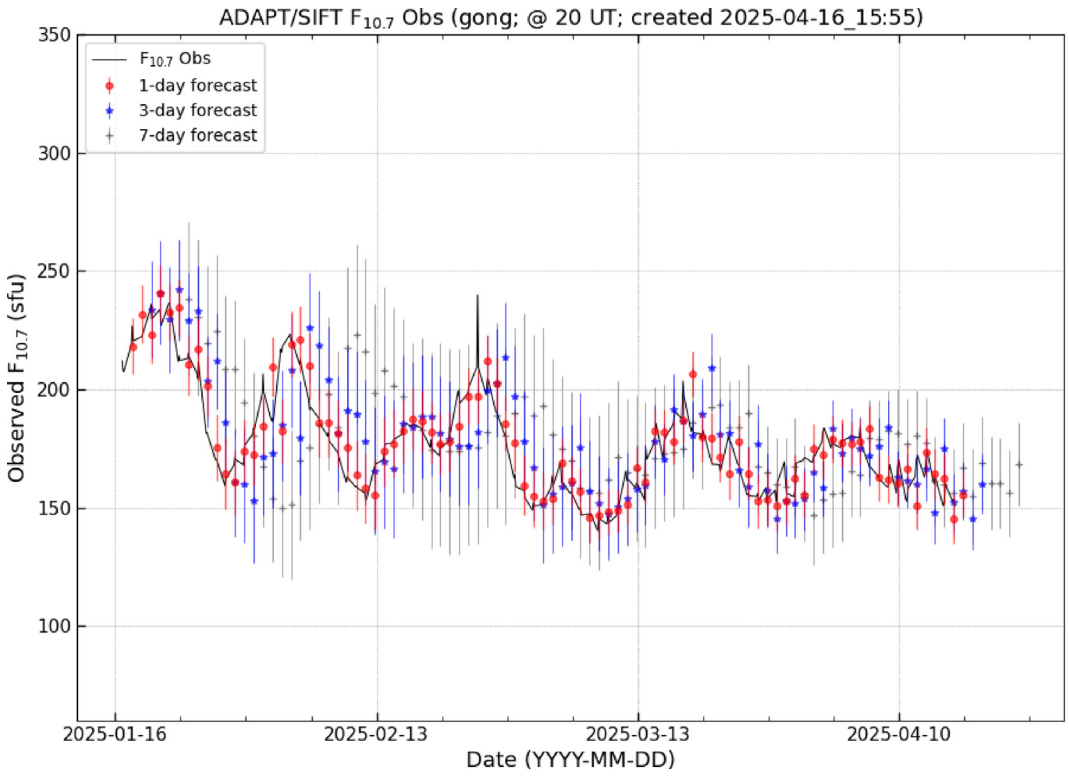

Fig. 6.2. Example solar index real-time comparison between model forecasts and observed values of F10.7, where the forecasts and uncertainties (error bars) are for 1, 3, and 7 days from the SIFT model.

raphy forecast[68]). The team is currently brainstorming the best way to display multiple real-time IMF Bz forecasts.

For an ideal Parker spiral heliospheric magnetic field structure, the field lies in the azimuthal plane, meaning Bz = 0 in radial-tangential-normal coordinates. Indeed, on long time-averages (> 1 day), a forecast of Bz = 0 is very accurate. Bz deviates from zero due to small-scale turbulent or stream interaction motions, or as the result of large-scale CMEs. This means that predicting the timing and phase of Bz is difficult due to both the stochastic component of Bz, and the requirement to first correctly predict the solar wind flow that will advect large-scale Bz structures to Earth. For this reason, point-to-point metrics – such as MSE, MAE, correlation, etc – can misrepresent the usefulness of a forecast. There are situations where the exact timing of a strong Bz < 0 turning is less important than the knowledge that such an event will occur over some time horizon (e.g., the next day) and/or an estimate of the magnitude. For this reason, it is often preferable to use threshold-based time-window approaches to validate Bz forecasts.

The IMF Bz Scoreboard will incorporate both event-based and continuous metrics to comprehensively assess forecasting performance, drawing from the methodologies developed for the SEP Scoreboard. Event-based metrics will evaluate model accuracy in predicting key Bz features during specific space weather events, such as the timing, intensity, and duration of southward Bz intervals that pose the greatest risk for geomagnetic disturbances. Continuous metrics will track model performance over extended periods, capturing the evolution of the Bz component in real-time to assess trends, fluctuations, and background conditions.

6.6. Planning for solar wind scoreboard

The Solar Wind Scoreboard, hosted at the CCMC, will serve the space weather and space science community as a hub for comparing real-time solar wind predictions at Earth, Mars, and other locations of interest. The Scoreboard will allow for viewing the ensemble of community-contributed model results and comparing their accuracy during extreme space weather events. Collaborating with the Ambient Solar Wind Validation Team (H1-01) in the COSPAR ISWAT initiative (section 4.1.1) will guide the Scoreboard architecture, including metadata standards, automated prediction submissions, and front-end development. An important aspect is to enable the space weather community to test the most advanced solar wind models in real-time scenarios that reflect operational use and to share lessons learned for a more efficient Research-to-Operation transition.

[68] https://iswa.ccmc.gsfc.nasa.gov/app/?layout=HelioTomoIPS

6.7. Planning for geomagnetic indices scoreboard

The Geomagnetic Indices Scoreboard project is led by the ISWAT G1-06 Action Team. The objective of this activity is to create a community-wide ensemble forecasting of Geomagnetic indices (Kp, Dst, Ap, Hpo) that are used as inputs for ionosphere-thermosphere forecasting models and as a base for interpretations of space weather conditions (see Table 3.2).

The Geomagnetic Indices Scoreboard aims to provide a platform for open validation and comparisons of different forecasting techniques for geomagnetic indices. The G1-06 team initiated a list of models participating in the ensemble, agreed on metadata for each model and forecasting method registration process, and agreed on a template (JSON file) for forecast submissions. The CCMC will develop and maintain software for the project building upon experience from the SEP Scoreboard. Several models (including the SWPC operational Kp index model and Heliospheric Tomography model) are already streaming results to the CCMC with outputs being displayed on CCMC ISWA.[69] Ongoing planning activities include: designing options for Scoreboard front-end display; discussing possible interfaces with Solar Wind and IMF Bz Scoreboards; and designing continuous validation system for model-model and model-data comparisons at different forecast lead times.

6.8. A vision for sun-to-impact scoreboards

There is a vision to link events within all Scoreboards and to create comprehensive space weather prediction chains. Progress with solar wind, geomagnetic and solar indices Scoreboards will enable running ITM and inner magnetosphere models in forecasting mode. Solar Flare Scoreboard could be used as an input for a possible ionospheric absorption Scoreboard. Designing front end displays for global and regional predictions for neutral density, charging, and absorption Scoreboards are next necessary steps. Involvement of space weather service providers and users is critical for this action. Suggestions for global neutral density forecasting products discussed at the ISWAT 2025 Working Meeting include global mean density at different altitudes (200 km, 400 km, 600 km). The major blocker now is the absence of in-situ neutral density observations for continuous validation.

6.9. Planning for long-term prediction scoreboard

Continuous observations of the Sun over the past few centuries have identified variations of solar activity on different time scales, the most prominent one being the 11-year sunspot cycle. Although the last decade has seen a lot of progress in developing predictive capabilities, there is still a significant divergence in forecasts (Pevtsov et al., 2026, this issue). In addition, different communities and end-users have different needs in terms of cadence, lead time, and accuracy of forecasted quantities. Usually, solar cycle forecasts are used in the context of mission design, e.g., forecasting the radiation environment for Moon and Mars human explorations to estimate received dose or the atmospheric drag to estimate fuel capacity for satellite re-entry. The needs for long-term predictions, months to years in advance, entail different challenges with respect to short-term space weather forecasting: large uncertainties and open science questions on the physical mechanisms driving the Sun and heliosphere evolution on time scales longer than solar rotation; paucity of long-term data beyond sunspot numbers; limited statistics for monthly and yearly averaged quantities available for statistical and ML analysis; and slow pace of new data for validating predictions. For these reasons, a new ISWAT team (S1-05[70]) was formed in October 2024 and started planning an effort to standardize historical and real-time validation of various solar cycle related quantities, including, but not limited to, sunspot numbers, F10.7, open magnetic flux, and galactic cosmic rays. The experience gained by existing validation activities on solar wind and other time-series-based indices will be leveraged when developing the long-term prediction Scoreboard.

6.10. The value of scoreboards: lessons from terrestrial weather

When considering the ISWAT Scoreboards endeavour, it is instructive to consider how effective equivalent endeavours have been in terrestrial weather. The closest analogs are the dashboards of the various WMO verification (validation) Lead Centres,[71] which compare multiple models; for the sake of brevity we relegate discussion of these to the paper by Henley et al. (2026) in this issue.

Here we concentrate on a simpler single-model analog. The canonical example of effective model forecast model monitoring in terrestrial weather is Fig. 6.3. This illustrates the long-term tracking by European Centre for Medium-Range Weather Forecasts (ECMWF) of their model's performance, as assessed by the anomaly correlation coefficient metric of the 500 hPa geopotential height ESEQ (described in ECMWF, 2025; see also Simmons and Hollingsworth, 2002, and §9.8.4 in Wilks, 2020; for further standardized terrestrial equivalents of ESEQs used by the Lead Centres above, see the WMO Integrated Processing and Prediction System manual, Appendices 2.2.34–2.2.36 WMO, 2023).

The near-monotonic improvement in model forecast skill for this metric over time, across all lead times, showcases the famous "quiet revolution" in numerical weather

[69] https://iswa.ccmc.gsfc.nasa.gov/app/?layout=indices_scoreboard.

[70] https://iswat-cospar.org/s1-05.

[71] See links to deterministic, ensemble and long-term WMO Lead Centre dashboards at https://community.wmo.int/en/forecast-verifications.

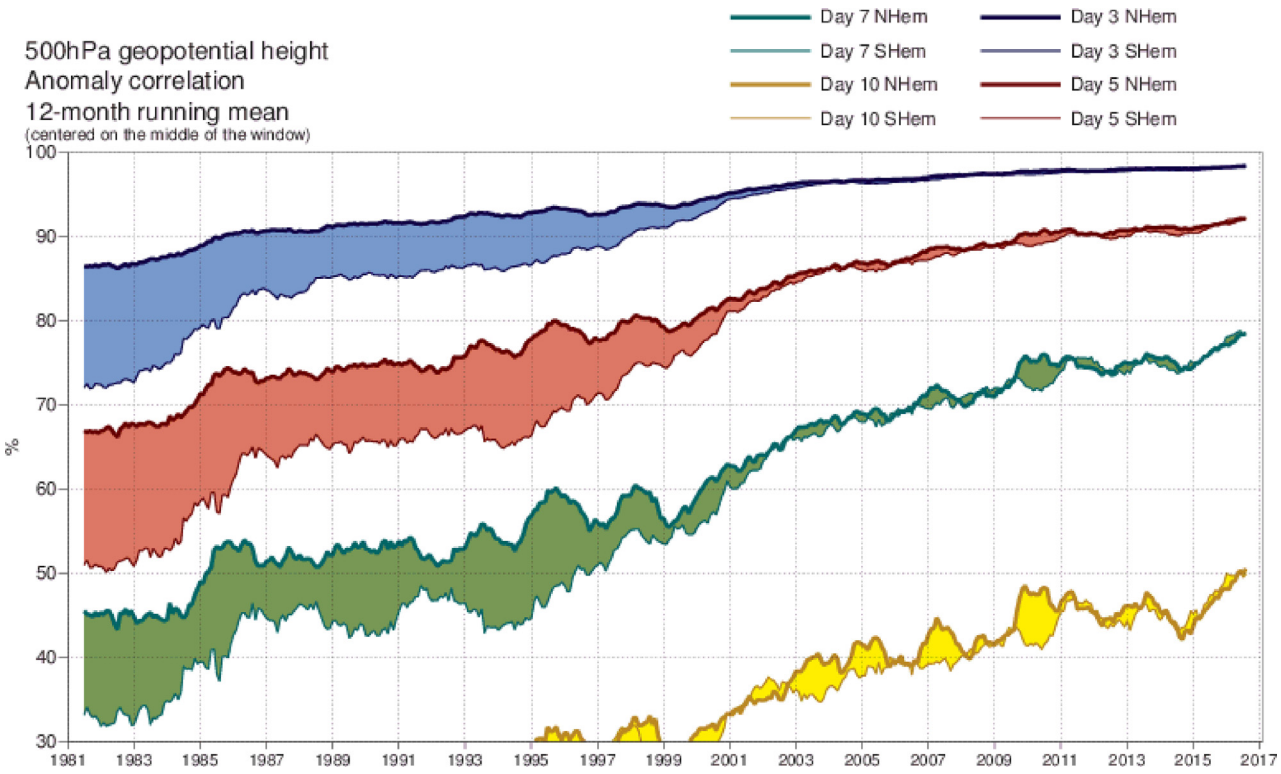

Fig. 6.3. Long-term tracking of forecast performance against a single metric, for ECMWF's terrestrial weather model. Larger y-axis values represent better forecast performance; different colours represent forecasts with different lead times (blue: 3 days ahead, red: 5 days, green: 7 days, yellow: 10 days). Thick (thin) lines represent forecasts for the northern (southern) hemisphere, coloured regions between these lines highlight inter-hemispheric performance differences for a given lead-time. Source: ECMWF (2017); note Current versions[1] of this plot follow Bauer et al. (2015) and use a non-linear y-axis adapted to inspect performance as values approach 100%. ([1]Available at https://charts.ecmwf.int/products/plwww_m_hr_ccaf_adrian_ts).

prediction. As discussed in the seminal work by Bauer et al. (2015), this improvement is attributed to a virtuous circle effect arising from the interplay of improvements in observations, improvements in modelling, and improvements in data assimilation combining models and observations to ensure model forecasts reflected reality better. Not only have the forecasts gained a day of lead time per decade, but the forecast skill between the hemispheres have equalised. This occurred in the late 1990 s, thanks to the onset of global coverage of satellite data – like other observations from many nations freely shared internationally via the WMO – and the assimilation of these satellite observations. Before this, forecast skill in the southern hemisphere was worse, as models had fewer observations there to constrain them via assimilation, and combat inherent model "tendencies" (e.g., Rodwell and Palmer, 2007; or Hsu et al., 2014 for a thermospheric/ionospheric example). The relative paucity in southern hemispheric observations then arises from limiting factors on the then primarily ground-based observing networks there: a smaller land mass and poorer economies.

A subtle but key point for the space weather enterprise to consider is that the main value of this diagram lies in the long-term tracking of a metric. When considering validation, it is easy to get caught up in detail: "are these ESEQs and metrics the right ones"? "what metadata do I need to capture about the model/observations/processing chains"? Such concerns are valid, and for robust analysis, the ESEQs and metrics need to be good enough, and sufficient metadata captured. However, such concerns should not delay simply starting to measure forecast performance, as the bar for good-enough long-term tracking is quite low: there needs to be sufficient trust that the performance shown comes from true forecasts, and the ESEQs and metrics chosen merely need to capture gross characteristics. The tracking shown in Fig. 6.3 encompasses huge changes[72] in the metadata, the models, and the observing systems – none of which are reflected in the plot, and none of which are needed, as the core message comes across without: previous investments in the observing system, models, and data assimilation have yielded return-on-investment in the form of improved forecasts, helping make a solid case for further investments (see Henley

[72] As an instructive historical perspective on this considerable change, and how the terrestrial community has progressed despite this via a pragmatic focus on the good-enough, the CF conventions started in 2001 (Gregory, 2003; https://cfconventions.org/faq.html#how-long-has-cf-been-around-is-it-mature), its COARDS predecessor in 1995 (Hassell et al., 2017; COARDS, 1995). Meanwhile the 1st edition of the WMO manual (WMO, 2023) standardising terrestrial ESEQs was published in 1992.

Fig. 7.1. Illustration of the R2O2R process, as described by Nelson et al., 2022, and guided by the NOAA RLs.

et al., 2026 for discussion on socioeconomic studies, also important for making the case).

The terrestrial community developed, agreed and standardized good-enough core ESEQs and metrics – this provided sufficient structure and stability to enable the key task of comparing models with reality, and tracking models' forecast skill over time. The space weather enterprise just needs to start tracking equivalent good-enough ESEQs and metrics; those being developed via ISWAT's Scoreboard endeavour are a good start.

7. Tracking frameworks

Tracking frameworks with specific applications in mind can help a user, or potential user, decide which product best fits their use case. They can also evaluate and monitor progress of a model during R2O development efforts, and act as prompts to check for aspects important to operations, such as robustness, running costs, and maintainability. This is the high-level base motivation for all tracking frameworks. Tracking frameworks can help identify which research products, tools, and models are ready to transition into operations. Defining such a framework is important for enabling the community to track the progress of the transition process and measure the usability of a research project in application.

Beyond this high-level motivation, different frameworks have been developed to address specific community needs. For instance, the motivation of Technology Readiness Levels (TRLs) is to have technology fly and operate in the space environment, while the motivation of Application Usability Levels (AULs) and NOAA Readiness Levels (RLs) is to ensure clear communication between developers/researchers and users about the use and current development stage of a product.

Here we will quickly examine the different frameworks and tools used to track a product's developmental progress, consigning detailed descriptions of these to Appendix C. We also extend the Klenzing et al. (2023) summary of use of the AUL framework across heliophysics with a summary of their implementation in the SWIMMR R2O program in the UK, and lessons learned.

7.1. ESA technology readiness levels (TRLs)

Technology readiness level (TRL) frameworks are in use in various organisations and countries worldwide, including NASA, ESA, and Canada.[73] The European Space

[73] E.g. https://ised-isde.canada.ca/site/innovation-canada/en/technology-readiness-levels.

Agency (ESA) TRLs are motivated to track the progress of hardware and flight software development for use in space (Mankins, 2009; Mankins, 1995; Olechowski et al., 2015). TRL definitions[74] are summarized in Appendix C.1 (Table C1).

The TRLs are used in many aspects of the research/space flight process. First, they provide a method to estimate the maturity of a technology for space flight. This allows scientists and other users to determine the instrumentation for use in a project's proposal and mission formulation stages. Subsequently, this helps funding agencies know 1) that a mission will likely be able to achieve mission success given the instrument requirements and 2) reduce risk by knowing how much technology development is necessary for the mission to fly and work successfully in space. As the TRLs do not require identification of a user or user-defined requirements, instrument development can occur and effectively be put on the shelf, waiting for an opportunity to arise. This framework has been used extensively within the aerospace instrument development community, and continues to be used in European Union programs.

However, the TRLs provided limited success with being used for flight and ground software, as the original metrics are designed for other uses. Specifically, when discussing and tracking ground-based products not attached to any individual space mission or space data, the TRL framework poses issues. If, as a program, you want to use one framework but have a diversity of projects, by definition, only those that make it to space will be able to reach a TRL of 9. In addition, having an institutionalized structure where some projects cannot attain the highest level imposes an implicit bias that those projects are less developed and less mature. Nevertheless, the TRL framework's success as an overall approach has informed other complementary R2O frameworks discussed here, designed for transitioning software (e.g., models) to operational service environments.

7.2. ESA SWE network product maturity levels (PMLs)

The ESA SWE Service Network has generated and adopted for internal usage the Product Maturity Scale (PML). The PML scale is intended to capture the current maturity level of a given development element with respect to its ultimate use as part of an operational system. The main scope of utilising the PML is to help identify the best strategy to develop newly proposed capabilities to a pre-operational status that meets users' needs and to track increasing maturity as this process proceeds. PML definitions are summarised in Appendix C.2 (Table C2).

7.3. NOAA readiness levels (RLs)

The Space Weather R2O2R Framework (Nelson et al., 2022) established an interagency process aiming to formalize an effective R2O2R process toward advancing the current state of space weather forecasting in the U.S. The Framework's partnering agencies (NOAA, NASA, DOD, and NSF) identified the NOAA Readiness Levels (RLs), therein, as the metric by which the maturity of evolving space weather forecasting capabilities will be measured – aiming to establish a common language and definitions essential to success (Fig. 7.1). Formally defined within the NOAA Administrative Order 216-105B,[75] these RLs are similar to the TRLs discussed in §7.1. It bears noting that in application of the RLs, a capability only achieves a given maturity level once it has completed all elements associated with that maturity level.

The RLs, like the TRLs, have nine levels, generally described via four phases; 1–2 Research, 3–5 Development, 6–8 Demonstration, and 9 Deployment. As a result, they can help determine the operational value of different applied research capabilities and can aid in guiding development work toward an identified user need. The formal definitions of the 9 RLs are provided in Appendix C.3 (Table C3), as are plain-language definitions which may be useful to model developers.

These formal definitions are intentionally somewhat general, affording the flexibility to appropriately classify all relevant research and development efforts across NOAA. Expanded descriptions of the RLs, their associated activities, and other relevant definitions pertaining to the R2O2R process may be found in the accompanying NAO 216-105B Procedural Handbook.[76] Note that the varied environments referenced within the NAO and The Handbook describe environments wherein the main efforts for each RL may be most effectively completed. Additionally, the rather ambiguous phrase "relevant environment" refers to relevancy of the environment with respect to a given maturity level and the associated activities. Thus, at RL5 the implied relevance is with respect to the Proving Ground (PG) validation. In other words, the PG validation cannot be completed within the developer's environment (this is intended to represent an objective validation) nor on the targeted end-use environment (e.g., operational environment) as it has not yet sufficiently demonstrated its potential to warrant the commitment of such resources.

7.4. NASA application readiness levels (ARLs)

The Application Readiness Levels were developed within NASA's Earth Science Applied Science Program (e.g., Pulkkinen et al., 2017). This framework has been used

[74] https://sci.esa.int/sci-ft/50124-technology-readiness-level/ and https://trlcalculator.esa.int/.

[75] https://www.noaa.gov/organization/administration/nao-216-105b-policy-on-research-and-development-transitions.

[76] https://www.noaa.gov/sites/default/files/legacy/document/2020/Mar/Handbook_NAO216-105B_03-21-17.pdf.

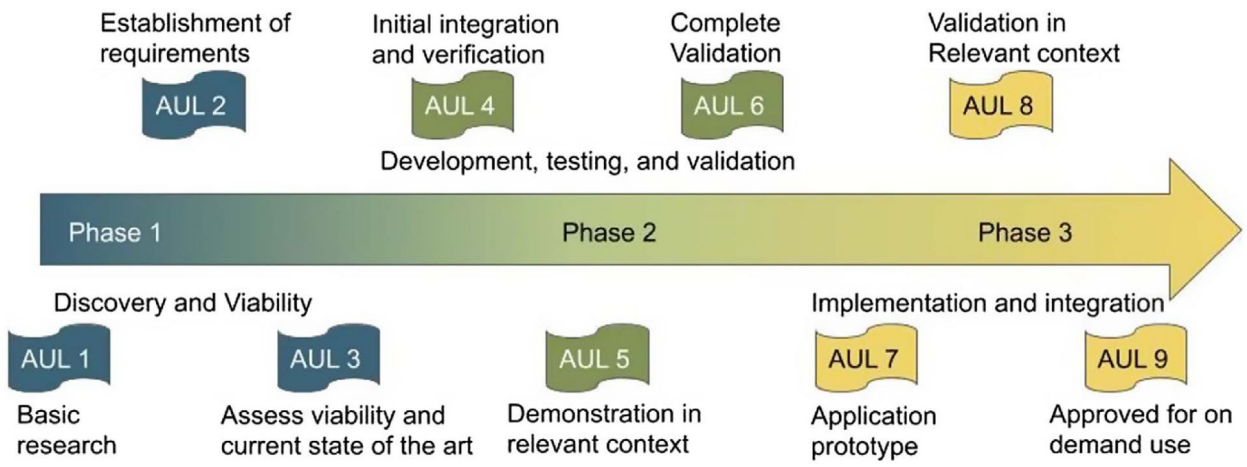

Fig. 7.2. The Application Usability Levels (AULs).

successfully to track the progress of research projects towards use or readiness for use with an applied partner such as an industry partner, NOAA, or other operational use.

Additionally, the ARLs have been used to measure the health of the Applied Science Program and the progress/-successes of this funding program. Using the same framework across all projects allows one to see where roadblocks may exist, such as a need for validation data. This can drive other programmatic decisions, such as which missions may be needed, data types, and resolution requirements.

The ARLs also spell out specific milestones which must be finalized to achieve the level. If a milestone cannot be met, even if milestones in subsequent ARLs are achieved, the level associated with the product is of the lowest level with all previous milestones that have been achieved. Thus, if a product has achieved all milestones in ARLs 1–7, is used in an operational context, but has not completed the user documentation, the product is evaluated at an ARL 7. A table with details of the ARL levels is provided in Appendix C.4 (Table C4).

While this may seem counterintuitive to allow a product to be evaluated at a lower level while operating in an operational setting, this information is key. This allows a product to continue development while addressing other roadblocks. For example, it may be difficult to get the necessary data to validate a forecasting model fully in the space weather context. However, that model may still provide actionable information and be useful to the end user.

7.5. ISWAT application usability levels (AULs)

The Application Usability Levels were developed as a community exercise, under the IFSWCA (now ISWAT Team O1-01), with guidance based on experience with ARLs from the Applied Science program (Halford et al. 2019). They suggested the name change to reflect that the product should be not only ready for use at the end of the project, but usable, and ideally in use by the time it reaches the implementation and integration phase. Additional lessons learned from the other frameworks were taken into account. For instance, from the TRLs, the lesson learned about the need for flexibility to address and account for a diverse set of projects was taken into account.

The motivation for the AULs comes from the understanding that a tracking framework can ensure clear communication between developers/researchers and users about a product's use and current development stage. Like the others, the AULs have three phases and 9 levels. AULs 1–3 are the Discovery and viability phases. Phase 2, the Development, testing, and validation phase, encompasses levels 4–6, and phase 3 (AULs 7–9) is the implementation and integration phase. The sequencing is illustrated in Fig. 7.2; further details are provided in Appendix C.5 (Table C5). Like the ARLs, the AULs have milestones that need to be completed to achieve the level. One noticeable difference between the AUL milestones and those in the ARLs is the focus on communication between the developer or researcher and the product user. Additionally, validation, verification, and feasibility are explicitly called for throughout the framework.

7.6. AUL implementation in SWIMMR for transition to operations

The Space Weather Innovation, Measurement, Modelling and Risk (SWIMMR) programme[77] was a successful £20 million, four-year initiative in the UK, completed in 2024. Its aim was to drive research and innovation in space weather to enhance the UK's capability to monitor and forecast space weather, thereby mitigating its risks and impacts. The programme also sought to foster closer ties between UK academic partners and the Met Office during and beyond its four-year duration and to expedite the R2O

[77] https://www.ralspace.stfc.ac.uk/Pages/SWIMMR.aspx.

transition process. Numerous UK establishments were involved.

The Met Office facilitated the development, deployment, and management of models by establishing an R2O platform. This platform allowed partners to build and deploy models using Docker containers in a pre-operational Amazon Web Services (AWS) environment. The platform enabled monitoring and logging and had an interface for accessing files and modelling system output. It was developed for six modelling systems addressing: Satellite Risk, Aviation Risk, Ground Effects, Ionosphere Effects, Thermosphere Effects, and Sun to L1 forecasting. Two SWIMMR 'pilot partners' were selected for early engagement and onboarding of systems to test the new platform, which also benefited partners by granting them early access to the Information Technology (IT) infrastructure.

To manage and support partners on the R2O journey, the Met Office required an R2O tracking process. AULs were adopted for this purpose due to the familiarity and development in the community, as well as their flexibility. While TRLs are designed for transitioning hardware to operations, AULs can be used to track for instance, the progress of software packages and data products.

Some adaptation was made to the AUL framework to suit the SWIMMR projects. At the start of the SWIMMR programme, projects were assigned as completing Phase I (AUL 3) because the Met Office was the recipient of the systems and because justification of research, state-of-the-art and operationalisation had already been provided by partners at the proposal stage. Therefore, the expectation was that projects would be tracked from AUL 3 to AUL 8, with further progress to AUL 9 to be fulfilled post-SWIMMR. The AUL Milestone text was enhanced to clarify, for example, that at AUL 5, the system demonstration was on the partner's computing platform, and at AUL 6, a prototype system was successfully built on the Met Office platform. An AUL Checklist, based on that provided in Appendix B of Halford et al. (2019), was provided to partners to summarise the documentation required at each level and to help assess and ensure compliance with each AUL level. A set of document templates was also supplied to partners to provide a framework for the necessary reporting of progress.

One partner successfully completed the necessary requirements and documentation to achieve AUL 8. There were some challenges in engagement with the AUL framework, potentially due to several factors. First, the decision to use the AUL framework was made after the programme had begun, which may have introduced additional documentation and unforeseen milestones for partners. Second, the staggered approach to onboarding, with pilot-partners first, delayed some partners' access to the IT infrastructure, possibly leading them to prioritise catching-up over compliance with AUL reporting. Third, partners may have been unfamiliar with the AUL framework. However, as the engagement issues became apparent, the Met Office took proactive steps to simplify the process.

The SWIMMR Programme has been a success, and efforts are currently underway to complete the operationalisation process. As the first programme of its kind for UK space weather, it encountered challenges, but these were resolved quickly and effectively. For future programme and project calls, it will be essential to clearly outline R2O milestones and ensure that partners are familiar with the AUL framework.

The AUL framework facilitated communication, and frequent progress meetings between partners and the Met Office were crucial in resolving problems and assessing progress. Moving forward, lessons learned will be used to enhance the AUL framework in the Met Office, leading to several benefits: a clear path to follow for R2O, clearer communication on progress and goals, clearer verification requirements for partners; and a better understanding of the documentation required from partners.

Such a framework may also help partners evidence the impacts of their work: the UK's Research Excellence Framework[78] requires academic institutions to periodically demonstrate the excellence of their research; this is often assessed via publications, but can also consider impacts.[79] Progress against a tracking framework may give partners a useful indicator to evidence the impact of their work in R2O projects; initial exploration of this insight has already occurred during the RISER project (see §3.2.2).

Overall, there are promising opportunities identified to further enhance the effectiveness of the AUL framework. The SWIMMR Programme has laid a strong foundation for future advancements in space weather research and operationalisation.

8. Examples of R2O2R pipelines

8.1. Modelers-CCMC-SWPC-users open R2O2R pipelines

A conceptual design for Modelers-CCMC-SWPC-Users R2O2R pipeline is illustrated in Fig. 8.1. The blue and black arrows show models and data flow from research to operations or from operations to research. Thick orange arrows correspond to demonstration of new capabilities and feedback from users. The red arrow shows operational services provided by SWPC.

New models are delivered to CCMC through the Modelers-CCMC collaborative environment at AWS cloud (frequently referred to as CCMC AWS Cloud) or at NASA High-End Computing Capability (HECC). For each model and computational task, the collaborative environment components are selected based on cost efficiency and performance. Seamless data flow between hybrid infrastructure components at AWS, HECC and CCMC on-premise systems (including validation platforms) is enabled by

[78] https://2029.ref.ac.uk/about/what-is-the-ref/.

[79] https://2029.ref.ac.uk/guidance/section-1-overview/#section-key-changes-for-ref-2029.

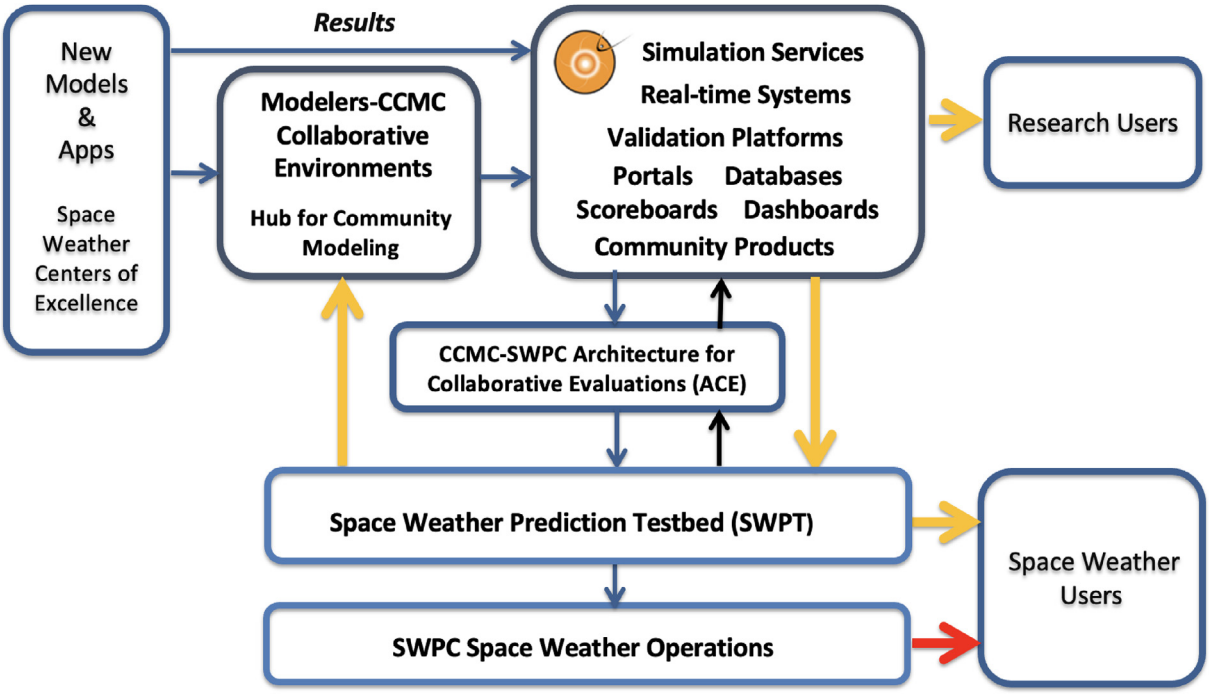

Fig. 8.1. Modelers-CCMC-SWPC-Users R2O2R Pipeline Scheme. Blue / black arrows show the flow of models and data. Thick orange arrows correspond to "Demonstration potential of new capabilities" and "Feedback from Users to Developers and CCMC". The red arrow corresponds to operational services provided by SWPC to end users. CCMC Space Weather portals and databases include Integrated Space Weather Analysis (ISWA) system and Database of Notifications, Knowledge, Information (DONKI). Community Products are designed by the community in collaboration with SWPC.

Apache Airflow.[80] Modelers are able to build, install, and run their models in the collaborative shared environment. CCMC points of contact (scientists, software developers, and system engineers) are available to assist wherever necessary. NASA's LWS Strategic Capabilities Program is an example of a successful implementation of this approach where model developers have been working with CCMC points of contact and utilizing the CCMC shared collaborative environment from early stages of projects. The same approach is now being implemented for deliverables from NASA Space Weather Centers of Excellence. This collaborative environment can be used as a hub for further model improvement and collaborative development of real-time community modeling systems focused on specific forecasting problems or user needs (e.g., solar energetic particles, neutral density and satellite drag, ionosphere variability).

After an on-boarding process and initial evaluations of robustness and performance are complete, the models are deployed in CCMC production systems and made available to the community through simulation services, real-time systems with dedicated queues, validation platforms, space weather portals, dashboards, and Scoreboards. Model results and timelines generated outside the CCMC can be streamed to CCMC space weather portals and databases (including ISWA, DONKI), Scoreboards discussed in section 6, and validation platforms discussed in sections 3.6 and 4.1. Community Products are designed by the community in collaboration with SWPC. Promising Community Products can be demonstrated to users at the Space Weather Prediction Testbed (SWPT) directly through real-time systems and space weather portals at the CCMC. Feedback from users through the SWPC Testbed is used for further improvements.

The CCMC-SWPC Architecture for Collaborative Evaluations (ACE) is a collaborative, cloud-based environment that is designed to mirror an operational environment at SWPC. Community Products that demonstrate potential to improve current operational capabilities are selected for implementation at ACE. When evaluations at ACE are complete, models/products are transitioned to SWPT for further evaluation, particularly in terms of user-impact. Such capabilities demonstrating sufficient improvements over existing operational utilities are then transitioned to operations.

8.2. ISEP – NASA In-House R2O2R pipeline

The Integrated Solar Energetic Proton (ISEP) Alert/Warning System project is an effective NASA in-house R2O2R pipeline for space radiation environment prediction in support of human missions beyond LEO. ISEP goals are to identify, transition, and evaluate new models (R2O); develop SEP Scoreboard software tailored for SRAG; and implement these capabilities within CCMC

[80] https://airflow.apache.org/.

as an experimental quasi-operational prototype. Through ISEP, the SEP Scoreboard now provides forecasts from over 10 models from the research community. One element of the partnership is the transitioning of a subset of models/software that involve human-in-the-loop from CCMC to M2M Space Weather Analysis Office (SWAO), who provide expert secondary space environment analysis for SRAG console operators, while SWPC provide primary forecasting support to SRAG. Central to this project was the development of the SEP Scoreboard web application by CCMC. ISEP also coordinates with SWPC to transition any promising models.

8.3. NASA space weather centers of excellence pipelines

The NASA Space Weather Centers of Excellence program address crucial space weather goals that require multi-institutional, interdisciplinary collaboration. Each Center is focused on a specific space weather problem, progress on which advances basic understanding and useful space-weather tool development for increased resilience of technological systems and support of space exploration.

- The Center for All-Clear Solar Energetic Particle Forecasts (CLEAR) aims to build an integrated prediction framework for the "quiet" solar wind, solar eruptive events, and SEPs specifically targeting magnetic connectivity and "all-clear forecasts" of low-radiation periods.
- The Space Weather Research and Technology Applications (SPARTA) Center focuses on the effects of ionospheric scintillation on communication, navigation, and timing systems and on providing a global scintillation forecast capability.
- The Space Weather Operational Readiness Development (SWORD) Center of Excellence aims to improve predictions of the impact of space weather events on LEO satellites and in cislunar space, specifically targeting improved predictions of neutral density at LEO during geomagnetic storms.

While The Centers also address fundamental research questions, their primary objective is to advance operational capabilities by developing their target capabilities to at least a level where their usefulness and performance have been demonstrated in an experimental quasi-operational setting, which is available at the CCMC. CCMC could play a role in that process by building upon the long-running collaboration between the SWMF employed by the CLEAR and SWORD centers, for example. In practice, the target is to transition models developed in those projects to CCMC for validation and community use and incorporation into CCMC realtime systems, and to demonstrate new capabilities to users through the SWPC Testbed as a step preceding transitioning to the final space-weather operational customer (SRAG and/or SWPC). There are also plans for SPARTA to follow the same R2O2R pipeline with the US Air Force as an operational user.

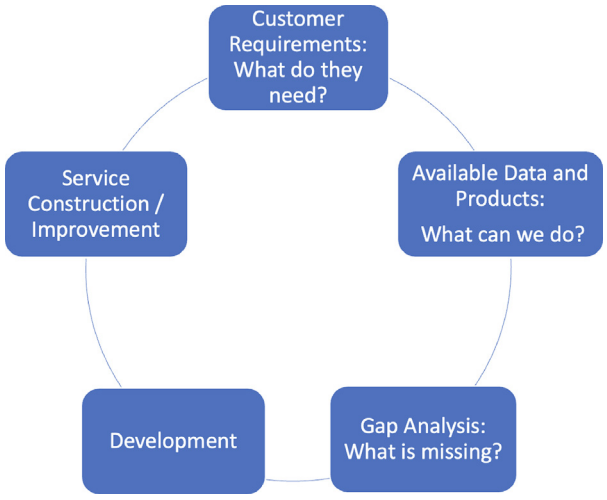

Fig. 8.2. Process of continuous improvement for ESA SWE Services.

8.4. The ESA space weather service network's pre-operational R2O2R pipeline

The ESA SWE Service Network provides a robust platform for continuous R2O2R processes where a wide variety of models, products, tools and techniques are developed and tested with end users in the loop.

All ESA SWE developments carried out under the Space Safety Programme focus on meeting end user needs developed through a process of extensive user consultation. The programme's requirement baseline is the SWE Customer Requirements for the ESA Space Safety Programme,[81] which evolved from the Space Situational Awareness – Space Weather Customer Requirements Document.[82] Through a process of continuous improvement illustrated in Fig. 8.2, new capabilities are developed, validated, and tested with end users actively engaged in testing. Feedback from end user testing is incorporated into future development planning for the network.

ESA works closely with the scientific community through expert groups such as the ESA Space Weather Working Team and the Space Weather Expert Service Centres (ESCs).[83] Each of ESCs comprises a distributed set of expert groups contributing particular data, products and/or expertise. ESCs carry out horizon scanning activities, aiming to identify new research results showing potential for future service improvement.These cases are captured in ESC development planning and service development roadmaps, which aim to capture new and maturing capabilities using an agreed set of maturity scales.

A core example of the SWE Service Network R2O2R processes can be observed in the development and integration in pre-operational environments of new applications, data streams, and their updates. The development activities

[81] https://swe.ssa.esa.int/documents/d/guest/esa_s2p_swe_crd.
[82] https://swe.ssa.esa.int/documents/d/guest/ssa-swe-crd-1001.
[83] https://swe.ssa.esa.int/expert-centres.

often originate from elements, including models of space weather phenomena and tools for estimating engineering impacts, developed within the research domain as they reach sufficient maturity for demonstration and testing with end users in the loop.

The determination of the maturity and expected usage of a new/updated service component is fundamental for setting the appropriate development and integration strategy. The usability is assessed by conducting a thorough requirement analysis in the context of real-life use cases. This initial process is done in conformity with the ESA SWE requirements baseline, and includes inputs from user engagement often collected via structured project workshops where the attendees represent the targeted user communities. This helps refine the development approach and ensure full understanding of the context in which the space weather information will be used. At the next stage, plans for development and validation are agreed. Development and validation activities that envisage major software development follow the ECSS protocols.

Validation efforts are considered and reviewed against the ESA SWE validation guidelines[13] and the ESA SWE requirements baseline. When a product development activity includes a validation campaign, the guidelines form the basis of the campaign strategy. The validation result is used to assess the performance of the final product. Often validation results uncover room for improvement of the underlying algorithms or data streams and trigger development corrective actions.

Product verification within the ESA SWE Network is ensured in the preparation phases of product integration in the ESA SWE Service portal after development and validation phases. As part of the verification process, all products formally undergo a live acceptance test procedure. This testing verifies the correct functioning of the products and their compliance with the ESA SWE requirements baseline.

The product is then integrated into the portal once the live acceptance test is deemed successful and all possible corrective actions or desirable enhancements are performed. The ESA SWE Portal is the pre-operational environment where the product is made available via user-tailored services built on grouping complementary products that target user requirements for a given service into tailored service interfaces. The General Data Services is an additional service domain that targets a broad audience including the scientific community.

Once integrated, service elements are subject to user test campaigns to demonstrate whether the final products are meeting users' expectations. Regular performance monitoring is carried out to ensure that the services remain available with the target >95% availability for the SWE Service Network's current pre-operational service provision mandate. Feedback on the existing products is constantly collected and evaluated. This enables tailoring developments to the needs of end users and provides valuable feedback for future developments.

Expert Service Centres carry out and document coordinated validation campaigns involving more than one product. These exercises act as a regular validation check on products and allow identification of those products that perform better in different space weather conditions. The Ionospheric Expert Service Centre has carried out several campaigns tackling TEC and Rate Of change of TEC values (ROT) and indices (ROTI). The validation campaign focussing on TEC included comparison among several TEC map products (global and regional) provided via the portal and against TEC map products external to the portal (e.g., TEC maps from Madrigal Database and the International GNSS Service IGS). The validation has been carried out by focusing on selected locations in Europe, by analyzing the complete available dataset, and finally by comparing maps. For each approach different validation metrics have been adopted.

The Solar Weather Expert Service Centre carried out an assessment of validation practices and caveats of recent solar-flare forecasting studies and reported on community activities carried out to date and the skill scores used. The results of this internal assessment were fed into the development of the SWE Service Network Validation Guidelines[49].

Within the SWE Service Network, some Solar Weather Expert Service Centre flare forecast products such as Royal Observatory of Belgium (ROB) Solar Influences Data Analysis Center[84] (SIDC) Solar flare forecast[85] have also published their validation or/and verification efforts in scientific journals (Devos et al., 2014), UK Met Office Solar flare forecast[86] (Murray et al., 2017), FLARECAST (see references at 4.1.4 Solar Flares) A-Effort[87] (based on Georgoulis and Rust, 2007).

Other products that have been the subject of dedicated validation efforts as part of their development include the e-Callisto radio-spectrograph network[88] in 2021. In the case of observational data validation, the definition of the ground truth can be any of the similar observational data. In this sense, the cross correlation among instruments and the data processing become part of this process. In the case of the eCallisto validation, it consisted of the evaluation of the data quality, and in the cross-comparison of the availability, quality, and categorisation of the stations.

Other products rely on the cross-correlation among instruments, as the case of Proba2/Lyra on EUV and GOES/XRS for X-ray flares. The product Proba2/Lyra data and SIDC Solar EUV flare detection shows this association: in the first one, the Lyra data is rescaled to GOES values to offer a proxy scale, and the second one lists the flares observed by Lyra with EUV flux scaled to its GOES equivalent, based on SolarDemon (Kraaikamp and

[84] https://sidc.be/SILSO/ssngraphics.
[85] https://swe.ssa.esa.int/sidc-S109b-federated.
[86] https://swe.ssa.esa.int/ukmo-S109c-federated.
[87] https://swe.ssa.esa.int/rcaam-federated.
[88] https://swe.ssa.esa.int/ecallisto-federated.

Verbeeck, 2015). The setting of thresholds for detection algorithms and the statistics derived from Proba2/Lyra LYRAEFF flare list and GOES/XRS are shown in Ryan et al. (2016).

In addition, some products provided via the Heliospheric Weather Expert Service Centre have been the subject of dedicated validation efforts which have been reported in scientific journals. The Drag Based Ensemble Model (DBEM) has been evaluated on their different versions (Čalogović et al., 2021). Furthermore, DBEM has been used for cross-comparison among the models and ensemble parameter comparison on an event in Dumbović et al. (2021). The Empirical Solar Wind Forecast model (ESWF) has been cross-compared in its successive version performances by Milošić et al. (2023).

The ESA Space Weather Office and Space Weather Service Network engage with end users in a variety of ways. In terms of new capability development and testing, this is most often in the context of end user test campaigns and workshops. All ESA space weather product and toolkit development activities include an element of user engagement in both the definition and testing phases for new capabilities and the feedback. This enables tailoring developments to the needs of end users and provides valuable feedback for future developments.

Validation and performance assessment are also a key part of the regular SWE Service Network community workshops, which bring together the SWE Service Network community of developers and providers along with end users to discuss progress and needs for further development.

8.5. SWIMMR – Met Office R2O pipeline

As part of the SWIMMR program discussed in §7.6, a cloud-based R2O platform has been designed to facilitate the transfer of models to the Met Office.

The key concept behind the R2O platform is to provide a collaboration framework and a pipeline to deploy research codes into a pre-operational environment for formal evaluation to AUL8 (see §7.6). The platform consists of an API to facilitate access to datasets required by the models, AWS infrastructure to deploy and run the models, and a portal webpage to access model logs, model outputs and model visualisations. The portal is required as direct access to the Met Office cloud infrastructure is restricted due to security considerations.

The design of the AWS infrastructure was based on the observation that previous R2O efforts have been hampered by the requirement to resolve conflicts resulting from running the models on a variety of platforms with different programming environments. To resolve this issue, the underlying architecture of the Met Office R2O platform was built to accommodate models using a containerised approach. Containers bundle together the operating system, the software packages (programming languages, libraries, etc.) and the model code. This approach also allows for some flexibility, where subcomponents of the models can be separated into different containers which are then orchestrated, the process through which the execution of the various containers and their interactions is coordinated. Developing and running models in containers resolves most of the issues arising from working in different development environments, but some issues can still arise if the containers are run on different processors, which requires different builds of the programming languages for example. In practice, our collaborators develop the models on their own platforms in a containerised environment, commit their code to GitHub and create pull requests to integrate their code with the R2O platform. Successful approval of these pull requests on GitHub triggers an automated deployment pipeline based on GitHub actions, through which the containers are built, deployed onto the cloud on AWS EC2 compute instances and executed.

Modern software engineering practices are based on the concept of having several execution environments for software, with usually a development environment for developing and testing new features and a production environment where the changes tested in the development are deployed upon acceptance testing. This approach is similar to the Met Office practice of testing weather models on a parallel suite running on the HPC (High-Performance Computing) for a length of time, in order to identify potential issues, before transitioning this parallel suite to the operational suite. The original design of the SWIMMR platform was predicated on a single execution environment for the code, and updates to the code would supersede the previous version. This created issues with code robustness, as some breaking changes were introduced and overwrote the latest stable versions of the code. The latest version of the R2O platform allows for the simultaneous running of several versions of the code: the beta release, the latest stable version of the code which is closest to transfer to operational implementation, and other versions of the code based on feature branches currently undergoing user-acceptance testing.

The Met Office R2O platform not only facilitates the R2O workflow by facilitating the transfer of code from collaborators in the research community to the Met Office but also provides a collaborative environment through which Met Office scientists and research software engineers can work more effectively with academic partners to improve operational readiness of the codes and to develop new features. The joint development platform allows to share issues with the models directly with the developers, and to identify pathways for future improvements. The ambition is for the R2O platform to enable operations-to-research use-cases too, to close the loop, and hence become a R2O2R platform.

Although this platform was designed for the SWIMMR project and improved as part of the follow-on SWIMMR acceleration project, the ambition is to expand the use of

the platform for other R2O activities. It is anticipated that this approach will considerably shorten the timescale on which research models transition to operations.

Approaches to R2O have been implemented in several organisations across the enterprise, and a key aspect to facilitate future efforts is to share lessons learned. One key lesson from SWIMMR for cloud development projects is the need to capture the full lifecycle of the models to accurately forecast the costs (e.g., data ingestion, data transfer, API, compute instances and data storage). The design of the system needs to be informed by comprehensively capturing all potential use cases. For example, the requirement to be able to run case studies and hindcast experiments was not part of the initial requirements and modifying the system after the design phase is completed is more complex than integrating the requirement from the onset.

Limited partner experience with the vagaries of near-real time observations emerged as another lesson. Many partners had experience running their models against quality-controlled science-grade data (e.g., in-situ observations from OMNI, or science-grade coronagraph imagery), and encountered issues when running their models against the near-real-time datasets, with issues such as datagaps or quality issues (e.g., Loto'aniu et al., 2022). Work was required to make models robust to these issues (e.g., Smith et al., 2022; Beggan et al., 2025). The Met Office APIs only provided near-real time observation data streams; a useful future extension will be to ensure these can run in "canned data" mode, providing examples of known near-real-time data issues, which model developers can test against.

Another emergent complication was platforms: the Met Office R2O platform is mainly cloud based, but some models' computational needs were too large, or had data ingestion needs maladapted to the cloud. As such, some models had to run on HPC platforms. This brought challenges of operating cross-platform, access to queues suited for real-time use, and coordinating multiple project timelines. As a shared key resource, the Met Office operational HPC (allowing real-time runs) has coordination overhead and its own project timeline, and is not designed with external collaborator access in mind. Meanwhile, it is easier to arrange external access to the research HPC, but this does not have the valuable guaranteed near-real-time queues: frustrating for those using the research HPC in the original SWIMMR project, as this immediately complicated their ability to demonstrate real-time potential! Follow-on work has required significant effort to ensure partners can access the near-real-time queues needed, and to aid engagement with the more formal processes around the operational HPC.

Finally, it is crucial to provide training to all research collaborators in the use of the platform and to provide tools and processes for joint project management. Careful consideration of roles is important: as the models approach operational status, institutional security requirements tighten, as do expectations of systems being robust enough for forecasters to rely on outputs being available for use. As such, in the follow-on phase, we are placing more guardrails around our partners, and giving them less free rein to change the increasingly-mature systems unsupervised, requiring Met Office staff reviews of proposed changes, and regular coordination sessions. Nevertheless, we recognise balance is needed – they are the model experts, and their development experience needs to remain acceptable. As such we need to avoid our processes or resourcing constraints needlessly inhibiting their progress. We are still learning how best to do this.

Description of CME propagation from the Sun towards earth depends not only on the properties of injected CMEs, but also on the distribution of the time-dependent ambient solar wind, which exerts forces on a CME and may change its speed, shape, and direction (Shen et al., 2022).

9. Space weather capabilities assessment in real-time operations

We briefly review existing efforts to validate space weather capabilities (models and products) used in real time operations. Many of the examples discussed represent near-real time validation, we note though that offline historical validation remains an important foundation – and is briefly covered too.

9.1. Space weather prediction center operational validation

The most fundamental type of real time (RT) validation is to present forecasters, scientists, and other space weather customers with predictions alongside observations for visually inspecting the soundness/reliability of the model inputs and performance. As a notable example of this, SWPC maintains a public source for RT validation[89] of the SWMF/Geospace model (Tóth et al., 2005, 2012). The web site displays RT plots of the model inputs (solar wind measurements at L1) and output predictions for Kp and Dst, displayed as time series together with RT measurements. An interactive display lets users monitor the performance of the continuous operational model runs (updates every minute) over time intervals from 3 h to 7 days prior to the current time, as well as short-term predictions made possible because the solar wind driver is provided by upstream L1 observations.

Another interactive public web site displays in situ solar wind measurements at L1, currently ACE and DSCOVR, soon to be replaced by Space Weather Follow-On at Lagrange Point 1 (SWFO-L1), together with solar wind predictions from the WSA-Enlil model. Such comparisons will be enhanced with the planned operational deployment

[89] https://www.spaceweather.gov/products/geospace-geomagnetic-activity-plot.

of the WSA dashboard developed at NASA Goddard Space Flight Center (GSFC) and hosted by CCMC,[90] which will allow SWPC forecasters to interactively assess not only the WSA predictions for solar wind properties at L1 (speed, magnetic polarity), but also the degree to which WSA-derived coronal holes match EUV observations from SDO/AIA.

SWPC is also currently implementing more sophisticated metrics in line with standards developed by the ISWAT ambient solar wind validation team (Reiss et al., 2023). These will be made available to forecasters to compare RT observations at L1 with daily ambient solar wind simulations. Efforts are also underway to use RT solar wind validation metrics to improve CME arrival time predictions by optimizing the WSA parameters (Meadors et al., 2020) and by selecting optimal ensemble members from the United States Air Force (USAF)/United States Space Force (USSF) ADAPT model (Hickmann et al., 2015) as boundary conditions for WSA-Enlil simulations.

Daily runs of the ionospheric outputs from the WAM-IPE model at SWPC are validated by comparing the simulated TEC with results from the GloTEC (Global TEC) data-assimilative model (Fuller-Rowell et al., 2022). GloTEC assimilates GNSS slant TEC measurements from ground-based receivers and radio-occultation observations from the COSMIC-2/FORMOSAT-7 constellation. A similar validation of neutral density is now being considered that would compare WAM-IPE model results to the USAF/USSF empirical High Accuracy Satellite Drag Model (HASDM; Storz et al. 2005), which assimilates measurements of drag effects on a set of LEO calibration satellites.

Validation of SWPC's nowcasting geoelectric models, a collaboration with United States Geological Survey[91] and Natural Resources Canada, has been performed using GIC measurements provided by power distribution operators. So far, two case studies have performed retrospective evaluation of GICs in power systems during geomagnetic storms: 1) one corresponding to the Southern California Edison System (Balch et al., 2023) during the September 2017 and April 2022 storms, and 2) another corresponding to the Tennessee Valley Authority (TVA; Balch et al., 2024) system for the March 2023 storm. These studies have shown that geoelectric fields from SWPC's models produced GIC estimates with high correlation to measurements. Continuous and (near) real-time validation of the geoelectric models could be possible by using observations from the North American Electric Reliability Corporation (NERC), which collects GIC measurements from electric power providers on an event case, and the Electric Power Research Institute (EPRI) which facilitates the real-time sharing of GIC measurements between subscribed power operators.

In the context of SEP prediction models, threshold-based validation metrics such as contingency tables are of most interest to SWPC (Bain et al., 2021). Though such metrics are often used for historical validation, they could also be useful in an RT or near-RT context. For example, if a model predicts an imminent threshold-based event, forecasters could monitor SEP flux levels to see if they are indeed approaching the predicted thresholds. This continues to be an area of interest as SWPC evaluates new SEP prediction models for potential operational deployment.

9.2. Met office space weather operational validation

To date, the focus of most operational validation at the Met Office has been in validating products issued by forecasters in the Met Office Space Weather Operations Centre (MOSWOC). Murray et al. (2017) outline the solar region summary product and the derivation of subjectively-adjusted full-disk M and X-class flare probabilities from objectively-derived equivalents (Bloomfield et al., 2012; Gallagher et al., 2002) for individual active regions and their full-disk combination. Such gross subjective adjustments are done by the forecasters to account for factors known to be relevant (e.g., via previous stratified validation efforts – see §3.6.3), but not yet integrated into the underlying objective model. The authors discuss how these subjectively-adjusted probabilities feed into other products, notably the probabilistic "Radio blackouts – X-ray flares" table in the more widely-distributed technical forecast product (designed for expert users) and its simpler counterpart, the "plain language" forecast product (designed for users with less expertise). This human-in-the-loop" approach is shown to improve performance over the objective reference model it is built upon (indeed, this and other human-in-the-loop methods for flare forecasting were found to be top-performers in the intercomparison results reported in Leka et al. (2019b). See §4.1.4). The monotonic fall-off in skill of these subjectively-adjusted probabilistic forecasts as lead time increases to four days ahead was also characterised.

Similarly, Sharpe and Murray (2017) further this flare validation analysis, and extend it to the "Geomagnetic storms" table also provided to users in the technical and plain language forecast products (see example in Murray et al., 2017). This also contains subjective probabilities – at the time of the study these were grounded on objective forecasts from the British Geological Survey[91] developed for real-time operational use[93] (Thomson, 2000), using linear-prediction (Thomson et al., 2001) and neural-network (Thomson, 1996) approaches. MOSWOC made subjective adjustments to these primarily to capture expected modulations from impending CMEs (inherently not captured by the objective methods, which forecast

[90] https://ccmc.gsfc.nasa.gov/wsa-dashboard/.
[91] https://www.usgs.gov/.
[91] https://geomag.bgs.ac.uk/data_service/space_weather/forecast.html.
[93] https://geomag.bgs.ac.uk/research/space_weather/sw_partners.html.

based on data from the past few days or last solar rotation), and occasionally if STEREO in-situ data suggested evolution in the ambient solar wind structures not well-characterised by 27-day recurrence (see §3.4). In recent years, these probabilities are simply based on previous subjective forecasts, appropriately time-shifted, i.e. are no longer grounded on the objective forecasts. This is largely due to this making it easier for MOSWOC forecasters, to retain adjustments made between subsequent shifts, and also allows a simpler technical implementation. Further stratified validation work is needed to characterise the impact of this subjective-only approach in times dominated by ambient (recurring) conditions, and in times of transient-dominated conditions.

Both Murray et al. (2017) and Sharpe and Murray (2017) show displays from the Met Office internal "Warnings Verification System". This near-real time verification (validation) dashboard, originally developed to validate terrestrial forecasts for mariners (Sharpe, 2013) has been extended to encompass space weather. Its main focus is the subjective probabilistic tables in the technical and plain language forecast products – the flare and geomagnetic storm tables discussed above, but also the "Solar radiation storms (high energy protons)", and "High energy electrons (≥ 2 MeV)" tables (these latter two have not yet been discussed in the literature). These dashboards allow forecasters to get an impression of how recent subjective forecasts have performed, on a rolling basis, over the last few days (to see recent detail), and as a ranked probability skill score (relative to climatology) over the recent months (to evaluate longer-term skill). In principle MOSWOC forecasters can use these to draw conclusions, and integrate these into their decision-making, e.g., if the dashboards suggest a sequence of recent forecasts had probabilities which were too low, given what eventuated, a forecaster might adjust their forecast probabilities upwards, if these "prevailing" conditions seem still to be persisting.

To date, only a very limited internal effort has been made to validate the "Geomagnetic activity – Earthbound Coronal Mass Ejections" table (Pope, 2016, covering only 8 months in 2014). This table appears in the technical and plain-language forecast products when an Earth-bound CME is predicted, encompasses brief comments on the nature of the CME, estimated by forecaster analysis of the CME in coronagraph imagery (Millward et al., 2013; Millward et al., 2024), as well as the crucial subjectively-adjusted forecast arrival time at Earth. The latter is derived from introducing the resulting "conefile" parameters for any CMEs present into the WSA-Enlil-Cone modelling system (Odstrčil, 2003; Odstrčil, 2023), and having MOSWOC forecasters review the resulting objective output (modelled images of CME propagation through the heliosphere, and time series of arrival times at planets including Earth) and making subjective adjustments to the forecast arrival times as needed. The vital validation of the Met Office's CME arrival time forecasts over longer time periods has to date solely been done externally, relying on MOSWOC forecasters manually uploading a subset of their subjective forecasts to the CCMC CME Scoreboard, and subsequent valuable community efforts to validate all scoreboard entries (Riley et al., 2018; Kay et al., 2024).

As described in Riley et al. (2018), MOSWOC forecasters' subjective adjustments to the modelled forecast CME arrival times aim to compensate for effects such as biases in the ambient solar wind, where differences between the modeled and observed solar wind the CMEs propagate in may degrade the model's ability to accurately represent the effective drag force acting on the CMEs (e.g., Cargill, 2004, Dumbović et al., 2018, Amerstorfer et al., 2021, §4.1.2), and hence degrade the accuracy of forecast CME arrival times. From fruitful conversations with the model developers and other domain experts, the Met Office's current implementation of the coupled WSA-Enlil-Cone (Odstrčil, 2003; Odstrčil, 2023) heliospheric modelling system may not be optimised to correct such biases; equally there are some inherent assumptions in the models which may be at issue. To improve the skill of CME arrival time forecasts, which currently do not show terrestrial-like year-on-year improvements (Riley et al., 2018; Kay et al., 2024), further robust validation is required to disentangle the multiple sources of error which can arise at all stages, which may compound or compensate, to allow a "divide-and-conquer" approach, which has served terrestrial weather well (see discussion in §6.10).

Validation efforts must consider studies such as

Synthetic studies of limitations on CME fitting due to the inherent observational degeneracies and subjective judgements involved (Verbeke et al., 2023), and equivalent operational studies (S. Gonzi, private communication.[94]

Coping with the "through the lens of two models" complexities of coupled modelling (M.L. Mays, private communication). The WSA dashboard mentioned in §9.1 has already been useful in this regard.

Improving the observed "true" ICME arrival time at L1. ICME catalogues show inconsistencies (e.g., Nguyen et al., 2019; Kay and Davies, 2025), and are not all well-maintained, timely, or themselves validated enough to meet the needs of near-real-time validation.

It is also essential to be pragmatic, recognising limiting factors which apply, and adapt the validation approach accordingly. Doing robust CME validation requires technical work to capture all the relevant metadata across multiple systems, from the near-Sun CME identification and analysis, to heliospheric modelling of the sun-Earth transit, to capturing any near-Earth ICME signatures in the in-situ observations from satellites at L1. Overall metadata to associate all of these items is also needed. This is inherently complex, as it needs to account for cross-system connections not always having simple one-to-one relationships:

[94] See also discussion of UC1 in https://swe.ssa.esa.int/use-of-l5-data-in-cme-propagation-models.

for example a seemingly Earth-bound CME may not be observed at L1; multiple CMEs may merge on their passage to Earth; or even an ICME observed at L1 may not have a clear CME progenitor, if the real-time coronagraph imagery was degraded (e.g., by datagaps, potentially backfilled later), or lacked a clear signature, and thus was not analysed in a given forecast. This complexity has meant internal Met Office attempts to capitalise on the human-in-the-loop forecaster to make judgement calls and overcome complications. An initial attempt was implemented internally, but ultimately foundered, as what Met Office R2O scientists had deemed essential for validation was arguably inherently over-demanding. This placed unrealistic demands on the forecasters, who have to evaluate and forecast the entire sun-to-Earth chain during their 12 h shift, so are severely limited on the time they can dedicate to a given task. As such, little information was gathered, and the interface to gather CME arrival times was retired during a platform upgrade. The future outlook is brighter though: consultations from ISWAT team leads developing the CME arrival time scoreboard (§4.1.2) have been helpful, as they have allowed mutually-valuable discussions to clarify what metadata is essential, and what is optional. Useful overlaps with SWPC have also been identified (M. Marsh, personal communication). And internal work is planned to enable automated submission of forecasts to the CME scoreboard: not only should this reduce the burden on MOSWOC forecasters, freeing them up for higher-value tasks e.g., tailoring communications to users, but it should also allow more granular validation of CME forecasts, and how any errors in CME forecast arrival times evolve as a function of lead time, potentially providing actionable O2R insights back to model developers.

Other inherent limitations also apply to CME forecasting. Notably developing a validation framework which can cope with CME merging: heliospheric models may not provide unique per-CME identifiers, inhibiting automated attribution of (potentially-merged) modelled ICME signatures (e.g., at L1) back to the CMEs injected at the model's inner boundary. Similarly, at least in operations, CME tracking is critically under-observed, with no operational use (at least at the Met Office) of existing heliospheric imagery to monitor the region between the outer boundary of coronagraphs and in-situ observations at L1: for near-real time operations, this is currently a heliospheric equivalent of the "ignorosphere". The forthcoming Vigil mission should improve matters, provided operational systems and processes are adapted to capitalize on it. ESA-funded studies (§4.1.3) show promise; as does work to consider integration into operational tools (e.g., Wharton et al., 2019), or transformative processes such as "ensemble pruning" (e.g., Harrison et al., 2017), which carry the promise of CME forecast arrival times with errors which shrink as the CME nears Earth, trading off lead-time for accuracy, much as is the norm for terrestrial forecasts (e.g., for hurricanes, and their narrowing "cone of terror" as they approach landfall). We note that even simply having forecasters routinely monitor Vigil HI imagery may allow some subjective progress in skill!

Fundamentally, operational space weather systems require more validation effort than we have allocated to date. Operational models at the MOSWOC have mostly relied on prior one-off validation efforts done by SWPC – e.g., Akmaev et al. (2010) for the D-Region Absorption Prediction (D-RAP) model (Sauer and Wilkinson, 2008), and an equivalent for the Relativistic Electron Forecast Model (REFM) (Baker et al., 1990**).** MOSWOC has conducted further one-off internal validation at rare updates of key models (e.g., WSA-Enlil-Cone, Odstrčil, 2003; Odstrčil, 2023). To improve the performance of operational models over time, and emulate the year-on-year progress of our terrestrial colleagues, MOSWOC requires near-real-time validation, and an ability to disambiguate the complex cases such as coupled-modelling chains e.g., the CME case discussed above. MOSWOC is planning out the verification (validation) strategy – see Henley et al. (2024) – and hope to continue the productive collaborations across the enterprise.

Finally, as an aside, Thomson (1996) is an early example of applying ML to operational space weather forecasting, not captured in the (naturally non-exhaustive!) Grand Challenge review by Camporeale (2019). Likewise the REFM model discussed there is not strictly ML (as a linear prediction filter approach), but from an operational perspective of explainability and maintenance behaves equivalently, with a set of weights derived from observations. And more recently, MOSWOC is onboarding new ML models via SWIMMR (e.g., Beggan et al., 2025, §7.6, §8.4). The

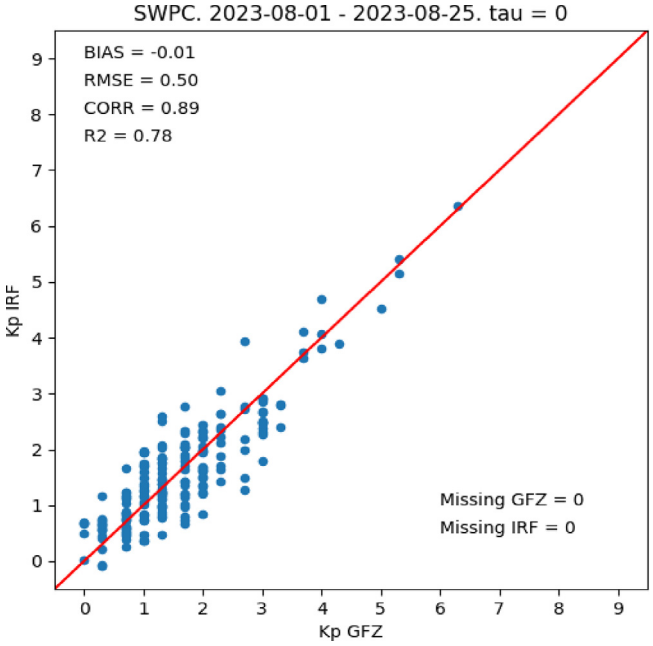

Fig. 9.1. Continuous validation outcomes from product forecast of Kp by IRF. The validation is performed with direct visual comparison in a scatterplot between Kp values predicted by the product and observed Kp values generated by GFZ. In the top left corner of the plot, a few metrics are also provided to the users.

space weather community may have some reservations on the merits of ML versus physics-based approaches, but the prime consideration for operations is usually forecast skill, as the above should testify. A key challenge for operational adoption of ML has been the explosion in the relevant literature, illustrated up to 2018 by Camporeale (2019), and the relative paucity of intercomparisons (§4) between similar ML approaches, or (better) approach-agnostic intercomparisons, as done by the flare community (§4.1.4). The vast body of literature accumulating poses a signal-to-noise challenge for operations at least: without like-for-like intercomparisons (ideally in near-real-time) to distill the lessons from the literature, it is hard to derive actionable insights: What models work? Why does a given model work better? How does a given model compare to incumbent operational approaches? Are any performance increases sufficient to outweigh other considerations (e.g., maintainability, robustness in operations, effort to adapt systems). Efforts from the ML community to involve themselves in community intercomparisons would be of great value for operations, and would help make a stronger case for operational uptake of ML.

9.3. Moon-to-mars space weather analysis office operational validation

Moon to Mars Space Weather Analysis Office (M2M SWAO) at NASA Goddard Space Flight Center (GSFC) carries out human-in-the-loop real-time analysis of space weather for NASA missions across the solar system, and performs real-time prototyping and validation of state-of-the art forecasting capabilities implemented at the CCMC or installed at the M2M SWAO environment. M2M SWAO also performs post-event and anomaly analyses. Formalized post-event analysis activities conducted by M2M SWAO have been standardized to include the following components: i) event analysis, ii) data availability, and iii) human-in-the-loop activities.

i) Event analysis focuses on the characteristics and associated phenomena of a specified event. For SEP events, this may include details about associated activity/triggers (e.g., flares or CMEs), the nature of the event onset (gradual vs rapid), and the duration of the event. Details of each event analysis are documented and updated in the M2M_Catalog within the DONKI

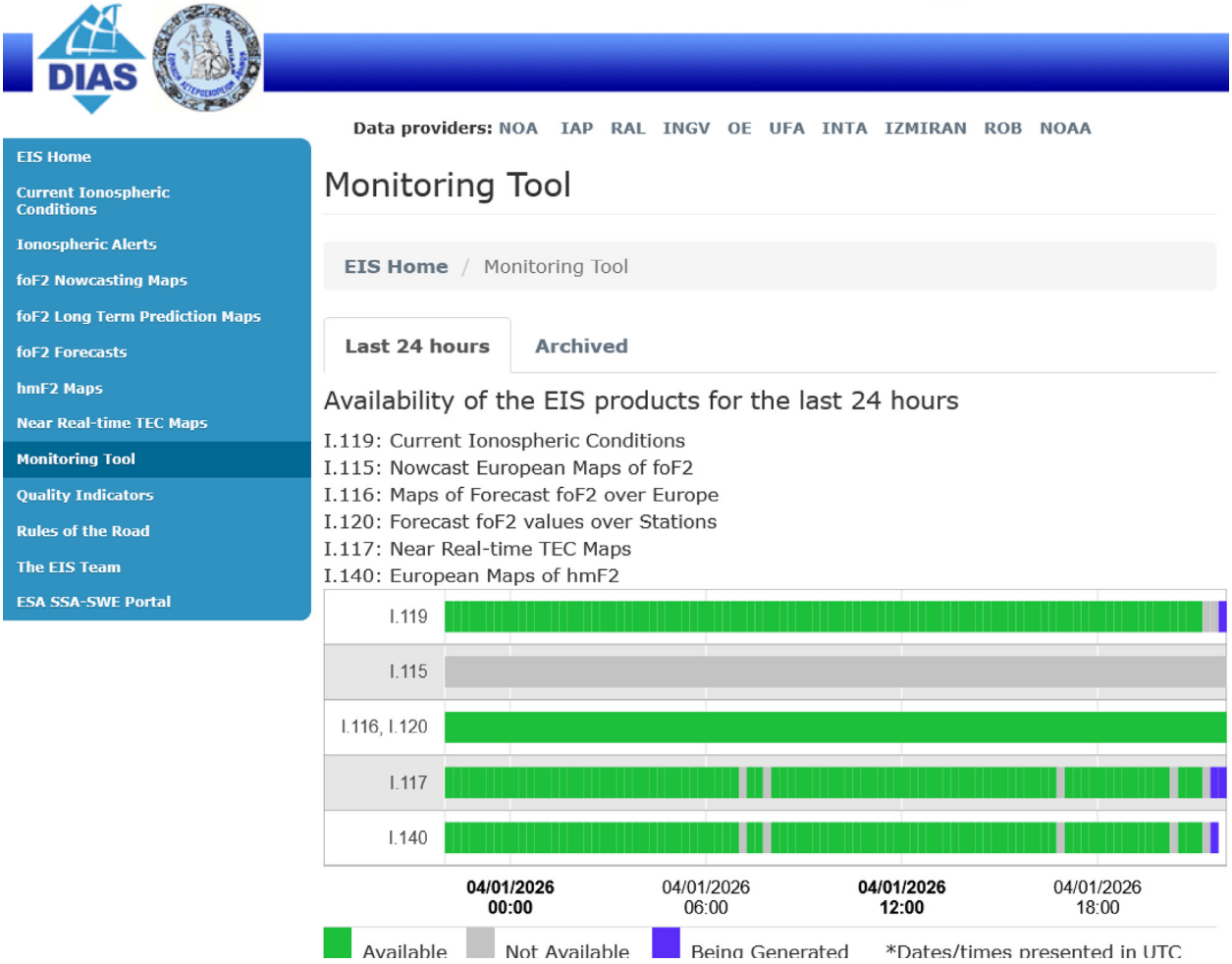

Fig. 9.2. EIS products by NOA integrated in the ESA SWE Service Portal provide a monitoring tool that allow the user to monitor the availability of the products and check whether they are currently being generated. . Adapted from https://swe.ssa.esa.int/web/guest/dias-federated

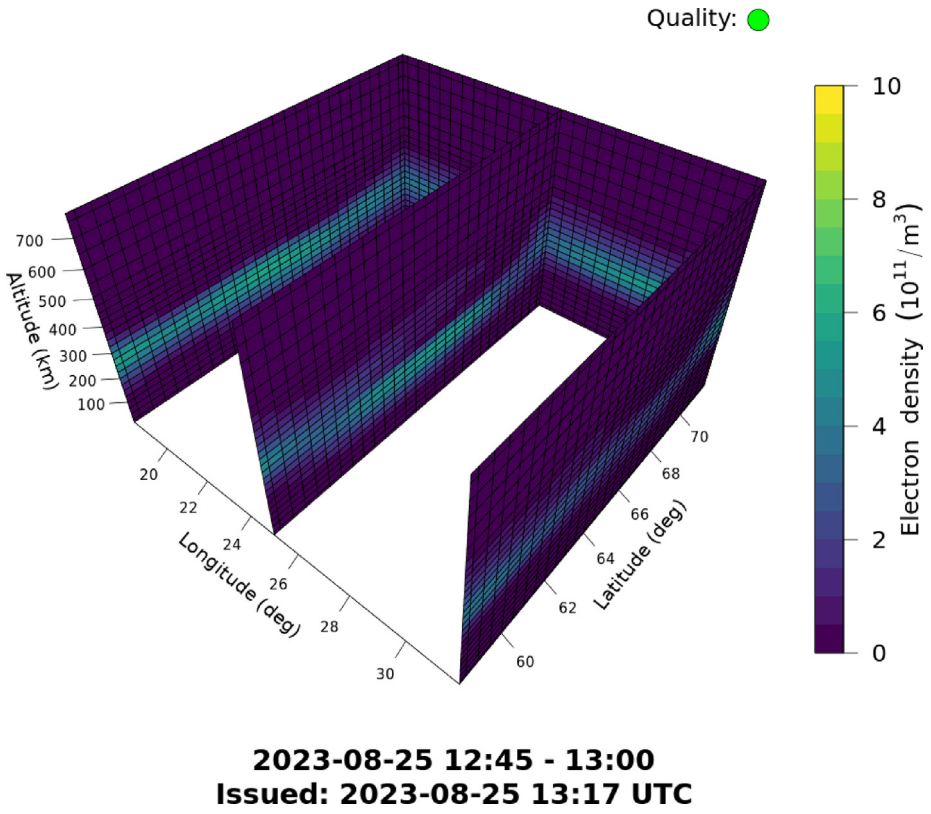

Fig. 9.3. Electron density reconstruction of the ionospheric region over Scandinavia (100 km – 800 km altitude range) for the 25th of August 2023 at 13:00 UTC, provided by FMI within the ESA SWE Service portal. The quality information indicator displayed at the top of the graph changes according to the quality of the inversion approach as follows: Green dot = nominal, Orange square = no a priori information available by ionosondes, Red triangle = data from less than 5 GNSS receiver stations available above 65° N or less than 50 receivers in total.

space weather database developed and maintained by CCMC.

ii) Data availability is an important factor which may impact model performance. Many models require input data and intermittent or missing input data may significantly reduce the effectiveness of a model to generate an accurate or timely forecast. Analysis of data availability for a specific event may provide clarity about the model's performance and further insight into the reliability of a model.

iii) Human-in-the-loop activities can be relevant for evaluating model performance as they may directly or indirectly be related to the configuration of a model to run in real time. For example, some models may require human-produced analysis (e.g., 3D kinematic CME parameters, flare source location). A thorough review of such human-in-the-loop activities may identify notable delays or uncertainties as factors contributing to a model's performance.

9.4. Examples of continuous validation within the ESA space weather service network

The ESA SWE Network encourages the implementation of continuous online validation systems. The continuous validation allows the developers to control the accuracy of the products' outcome. The main objective is to support the users in evaluating the reliability of the product they are using, and eventually guide them to those products that have the best performance in specific conditions or for specific use cases.

The ESA SWE portal includes several examples of continuous validation. The Swedish Institute for Space Physics (IRF) provides, via the portal, forecasting products accompanied by real-time comparison with truth data. Fig. 9.1 shows the continuous validation results of the predicted Kp index generated by the product Forecast of Kp (by IRF) against the actual Kp index values provided by the German institute GeoForschungsZentrum (GFZ). Forecast of Kp (by IRF) uses a scatter plot and reports skill scores. The plot is updated at each prediction, and it is initiated monthly.

The provision of products via the ESA SWE Service portal implies verification of product availability and timeliness. The verification is currently performed by human intervention on a regular basis during working hours. In some instances, the product provider implements internal backend processes aimed at collecting availability and timeliness information. As a rule, each product is expected to inform the user in case of unavailability due to lack of input data. For example, the products belonging to the European Ionosonde Service (EIS) provided within the ESA SWE Service portal by the Ionospheric Group of the National Observatory of Athens (NOA) provide an

overview of their availability. Fig. 9.2 shows the availability and the latency of the latest EIS products.

To enhance the user experience, several products accompany their outcome with quality indicators. In particular, the product TomoScand3D, provided in the ESA SWE Service portal by the Space and Earth Observation Centre of the Finnish Meteorological Institute (FMI), reports a color-coded quality indicator that is based on the diversity of input data used to generate a tomographic tridimensional electron density reconstruction of the ionosphere over Scandinavia. Fig. 9.3 shows an example of TomoScand3D graphical output with the quality indicator at the top of the plot.

10. Summary and recommendations

This paper has provided a comprehensive overview of model assessment work undertaken by the international space weather community in recent years to advance the state-of-the-art, including work undertaken under the aegis of COSPAR ISWAT, as well as work at organisations such as ESA, NASA CCMC, NOAA SWPC, and the Met Office.

Below, the key messages from all this work are distilled into recommendations useful for the space weather enterprise at large to consider, and especially for decision makers at institutions acting as key drivers for space weather research. The intended audience for these recommendations includes:

- researchers who perform the vital underpinning work to provide the models and methods used for operations, who validate these against the intended goals, who consider how best to measure this, and how to identify the missing gaps in the approaches to date
- organisations responsible for R2O programs (NASA, ESA, NOAA SWPC, Met Office, and others with similar functions)
- national space weather programs which fund R2O and O2R efforts, including the crucial late-stage R2O efforts before entering operations, to fully bridge the R2O "valleys of death" (National Research Council, 2000, 2003).
- international coordination bodies with an interest in space weather, notably WMO, ISES, COSPAR, CGMS (for more details see Ishii et al., 2026b, this issue, and a website for the International Space Weather Coordination Forum[95])

Recommendations are numbered for future references.

#1. Establish like-for-like validation as a standard practice across space weather research and operations.

Adopt like-for-like validation and regular intercomparisons as a standard operating procedure to enable researchers and forecasters to derive actionable insights. These insights include: Why does a given model perform better? How does one class of models compare to others? How do new models compare to operational models? Are new model capability increases sufficient to justify trade-offs in maintainability, robustness, and implementation efforts?
Prioritize the standardization of validation procedures, including agreement on events or time intervals, metrics, and essential physical quantities across space weather research and operations. Establishing like-for-like validation requires a coordination approach across all levels of the space weather enterprise. From the top down, research funding bodies and operations agencies funding R2O work must treat like-for-like validation as an essential deliverable. At the same time, bottom-up engagement through community-led efforts, as discussed in §4, are crucial for the community to agree on validation standards, support their adoption, and ensure buy-in.

#2. Validate both forecasts of the environment and impacts.

The concepts of Essential Space Environment Quantities (ESEQs) and Impact Quantities (ImpQs) provide a useful separation of concerns for the enterprise: some models can concentrate on forecasting the environment; other models can take the resulting forecasts of the environment as input to predict the resulting impacts on technological systems. This approach is used to good effect in terrestrial weather, and allows a clearer attribution on sources of uncertainty.
We recommend the space weather enterprise aims to validate both ESEQs and ImpQs, while being aware of the linkages. A validation effort grounded in ESEQs allows modelers to aim to forecast the environment, an easier target for validation, as it may be observed remotely or in-situ, without issues of proprietary data, or changing technologies being impacted differently over time. Meanwhile, having concurrent validation efforts targeting ImpQs helps keep the enterprise focused on user needs, what ESEQs are essential to predict well to fulfill these needs, and hence how to prioritise model development work – and validate the resulting changes accordingly!

#3. Ensure the quality and appropriate use of observational datasets in model validation

Ensure that observational datasets used for model validation are accurately selected and quality-controlled. Raise awareness for the limitations of observational data, such as noise and resolution differences. In addition, clarify if the purpose of validation is to enhance scientific understanding or assess operational performance as using low-quality data for scientific studies can result in incorrect conclusions. Integrate error estimation of observational data into validation workflows. In parallel, also manage the needs for near-real-time model validation use-cases, using near-real-time data which has not undergone such quality control. Significant R2O barriers have been encountered when researchers have developed models trained on science-grade observations not available in near-real time; in many cases embedding assumptions deep into the code, requiring significant effort to rewrite such prototypes to assess operational viability. To avoid similar pitfalls for near-real-time model validation, we recommend automated model validation tools allow two modes: validation against cleaned-up observational datasets, and validation against messy near-real-time observations.

Further effort should be undertaken to ensure that datasets prepared for validation are easily findable and accessible to the community.

#4. Develop and apply automated model validation tools to support like-for-like comparisons and track progress over time.

Advance the development of automated tools for validation to make community-agreed standards accessible and applicable. Prioritize automation to reduce labor demands, minimize inconsistencies, and ensure validation efforts keep pace with ongoing predictive capability developments. Integrate well-curated observational datasets in automated validation tools to enable like-for-like comparisons.Routinely applying these tools will enable the community to assess model capabilities and track progress through new developments over time. To ensure broad usage, automated validation tools need to be flexible, well-documented, and easy to use. Web-based systems with standardized input formats and automated reporting can streamline validation workflows, allowing both researchers and operators to consistently evaluate model performance and report progress.

#5 Support pre-event community ensemble predictions to identify models with potential for operational deployment.

Expand on efforts like the ISWAT/CCMC Scoreboard projects, discussed in §6, to assess the predictive capabilities of research models under realistic operational conditions. These efforts are of pivotal importance to demonstrate the capabilities of new science models to operational agencies and end users.

Develop Scoreboards in collaboration with operational centers to ensure front-end design and functionalities meet operational needs. This collaboration ensures that operational centers recognize the practical usefulness of Scoreboard projects and can shape them for optimal use as supplementary tools for forecasting and in space weather testbed systems.

Furthermore, models participating in Scoreboard activities should be compared to operational model outputs. Outputs from operational models need to be made readily available to enable a meaningful comparison. This allows for routine benchmarking against a broader set of research models.

#6 Integrate continuous validation into Scoreboard projects

Scoreboard projects should adopt continuous validation as part of their operating procedure. Collecting and displaying forecasts is an essential first step. However, continuous validation provides actionable insights for forecasters, and can also support O2R by allowing them to flag up issues to model developers. Furthermore, the front-end should be intuitive and effectively communicate the quality of recent forecasts by different models.

#7 Adopt common open validation libraries to support coordinated R2O and O2R efforts across space weather centers.

To avoid duplicating efforts, adopt shared validation libraries across all centers involved in R2O and O2R efforts, such as NASA CCMC, ESA SWE, NOAA SWPC, and the Met Office. Individually, each research and operational center faces significant resource challenges in developing comprehensive validation frameworks to validate space weather models from Sun-to-Earth. Shared infrastructure would enable centers to pool expertise and accelerate progress toward robust, community-supported validation. Collective collaboration on common validation libraries would streamline the assessment of space weather modeling capabilities. Identify any blockers to collaboration, and explore pathways for advancing.

#8 Establish a collaborative agreement on naming conventions

Agree on naming conventions, including unique IDs for parameters (starting with ESEQs), to harmonize multi-model and model-observation comparisons. Databases of parameters with metadata and lookup tables linking alternative names and identifiers for the same parameter would be helpful for interconnecting different repositories and to ensure that we are comparing the same quantities.

Ensure any such efforts to reach agreement involve both research and operational parts of the enterprise: we are in a symbiotic relationship, and both stand to benefit considerably from a widely-adopted standard. To set this up to succeed, it is vital to consult widely, to ensure the diverse needs involved are well-captured (e.g., some operational actors may need to fit with established terrestrial conventions and; some research actors may have considerable dependence on existing research conventions).

In terrestrial weather, both research and operations communities have gained significantly from the Climate and Forecasting (CF) conventions providing a unified metadata standard, common across all the terrestrial spheres. Astronomy has likewise benefited from some unification in the International Virtual Observatory Alliance (IVOA). We recommend space weather takes on this challenge too, to reap rewards of easier validation, model coupling, and data assimilation.

#9 Harmonize tracking frameworks to enable consistent progress evaluation

Centers running R2O programs, such as SWPC, CCMC, ESA VSWMC, and the Met Office, have developed tracking frameworks to monitor the progress of their R2O projects. They also play an important role in evaluating critical factors such as model robustness, computational requirements, and operational feasibility.

Tracking frameworks have largely developed in parallel, often informed by institutional predecessors. As noted in §7 and Appendix C, definitions in these frameworks share many traits, even at the level of detailed definitions. We recommend that the centers and R2O funding agencies initiate a dialogue to assess the merits of harmonizing their frameworks, identify any blockers, and explore strategies to move forward.As with other recommendations, the whole enterprise stands to benefit: models often move between different R2O programs as they get developed; having a common framework would allow model developers and those running R2O programs alike to assess and communicate the R2O status of models more easily, in a compatible manner, and hence facilitate overall progress.

#10 Support shared computational environments and efforts for collaborative development of experimental real-time space weather modeling systems

The value of shared computational infrastructure on a cloud or HPC in accelerating model transitioning has been demonstrated by the NASA CCMC, UK SWIMMR, model developers, and space weather service providing partners (e.g., NOAA SWPC, NASA SRAG, and UK Met Office). Collaborative environments should be recognised as an essential element of R2O2R pipelines. These collaborative environments can be used as hubs for further model improvement and collaborative development of real-time community modeling systems and community products focused on specific forecasting problems or user needs (e.g., solar energetic particles, neutral density and satellite drag, ionosphere variability). Outputs from such experimental real-time systems can be streamed to ensemble prediction Scoreboards and used as inputs for community products ready for demonstration to users through testbeds at operational agencies.

#11 Strengthen coordination between research and operations to build trust and make progress

In conclusion, we emphasize the importance of a close collaboration between the space weather research and forecasting communities. An effective transition from R2O, and vice versa from O2R, requires a shared vision for comprehensive validation of our space weather modeling assets. Building trust is best achieved through close partnerships. Collaborative initiatives like ISWAT offer fertile ground for creating a common language between scientists, model developers, and forecasters. Since its formation, the ISWAT initiative has hosted international teams to tackle key questions across space weather research domains, including action teams focusing on model validation. Operational centers have become increasingly engaged in these activities. Community-driven efforts like the Ensemble Predictions Scoreboard and Modeling Challenges have been significantly energized by ISWAT's collaborative environment. We encourage investment in such community initiatives that foster collaboration, build and strengthen relationships, and bring the space weather community closer together. We encourage the recently established Heliophysics Open Modeling Environment (HOME, Corti et al., 2026) to join the ISWAT as an overarching activity.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

M. M. Kuznetsova, M. Reiss, C. Corti, M.L. Mays, L. Rastätter, J. Wang, J. Yue, Y. Zheng, J. Jones, M. Levisohn, C. Wiegand, and E. Resnick acknowledge support of CCMC from NASA Heliophysics and the National Science Foundation (NSF).

K. Muglach acknowledges support from the NASA cooperative agreement 80NSSC21M0180 and support from the NASA Living With a Star program through its Heliospheric Science Support Office.

C. J. Henney is partially supported by Air Force Office of Scientific Research (AFOSR) task 25RVCOR001. The views expressed are those of the authors and do not reflect the official guidance or position of the United States Government, the Department of Defense, or the United States Air Force.

M. M. Bisi acknowledges support from the Natural Environment Research Council of UK Research and Innovation (Grant number NE/X019004/1 [RISER] and Grant number NE/W002981/1 [EISCAT_3D FINESSE]) and also via the RAL Space In House Research programme funded by the Science and Technology Facilities Council of UK Research and Innovation (award ST/M001083/1).

B. Forte acknowledges support from the UK Natural Environment Research Council of UK Research and Innovation (Grant number NE/R009082/1, Grant number NE/V002597/1, Grant number NE/X019004/1, and Grant number NE/W003074/1).

This material is based, in part, upon work supported by the NSF National Center for Atmospheric Research, which is a major facility sponsored by the U.S. National Science Foundation under Cooperative Agreement No. 1852977.

The authors acknowledge the constructive feedback provided by Robyn Fiori, anonymous reviewers and the editor, which has significantly enhanced both the clarity and quality of this work.

In Memoriam

We are deeply saddened by the loss of our colleague, Tara Jensen. We are very grateful for the dedication Tara showed to the space weather enterprise over the years, which has been crucial in shaping the ideas on future directions in space weather model validation expressed in this paper. Her passing is a tremendous loss to the scientific community, but her legacy and impact will live on.

We had to say goodbye to Mostafa El Alaoui who sadly passed away during preparation of this manuscript. Mostafa played a key role at the NASA CCMC, helping bring cutting-edge space weather models into operational community use. The space weather community remembers him for his scientific rigour, his kind and collaborative spirit, and the lasting impact of his contributions on our understanding and prediction of the space environment. Mostafa composed the time period table used as a base for validation activities discussed in this paper.

Appendix A. Acronyms

ACE: NASA's Advanced Composition Explorer
ACE: CCMC-SWPC Architecture for Collaborative Evaluations
ADAPT: Air Force Data Assimilative Photospheric Flux Transport — Model
AIA: Atmospheric Imaging Assembly — Instrument
API: Application Programming Interface
ARLs: Application Readiness Levels
AUL: Application Usability Levels
AWS: Amazon Web Services
ASWO: Austrian Space Weather Office — Institution

BIRA-IASB: Royal Belgian Institute for Space Aeronomy — Institution
CAMEL: Comprehensive Assessment of Models and Events using Library Tools — Software
CCA: Canonical Correlation Analysis — Statistical method
CCMC: Community Coordinated Modeling Center
CEDAR: Coupling, Energetics, and Dynamics of Atmospheric Regions — NSF program
CF Conventions: Climate and Forecast Metadata Conventions
CGMS: Coordination Group for Meteorological Satellites
CH: Coronal Hole — Phenomenon
CHAIN: Canadian High Arctic Ionospheric Network
CME: Coronal Mass Ejection — Phenomenon
CME TOA: Coronal Mass Ejection Time of Arrival
CMR: CCMC Metadata Registry — Software/Database
COSPAR: Committee on Space Research
DONKI: CCMC Database Of Notifications, Knowledge, Information — Software/Database
D-RAP: D-Region Absorption Prediction model
ECMWF: European Centre for Medium-Range Weather Forecasts
ECSS: European Cooperation for Space Standardization
FLARECAST: Flare Likelihood And Region Eruption Forecasting — European Union project
ESA: European Space Agency — Institution
ESEQs: Essential Space Environment Quantities
EUMETSAT: Europe's Meteorological Satellite Agency
EUV: Extreme Ultraviolet
FTP: File Transfer Protocol
GCS: Graduated Cylindrical Shell reconstruction technique
GEM: Geospace Environment Modeling — NSF program
GIC: Geomagnetically-Induced Currents — Phenomenon
GION: Global Ionosonde Observation/Operation Network
GIRO: Global Ionosphere Radio Observatory — Network
GITM: Global Ionosphere Thermosphere Model
GNSS: Global Navigation Satellite Systems
GOES: Geostationary Operational Environmental Satellite — Spacecraft
GOLD: Global-scale Observations of the Limb and Disk — Spacecraft
GONG: Global Oscillation Network Group
GPS: Global Positioning System
GSFC: NASA Goddard Space Flight Center — Institution
HAPI: Heliophysics Application Programming Interface
HECC: NASA High-End Computing Capability
HF: high frequency
HI: Heliospheric Imager — Instrument
HMI: Helioseismic and Magnetic Imager — Instrument
HPC: High-Performance Computing
ICMEs: Interplanetary Coronal Mass Ejections — Phenomenon
IFSWCA: International Forum on Space Weather Capabilities Assessment
IGS: International GNSS Service
ILWS: International Living With a Star — International Working Group
IMF: Interplanetary Magnetic Field
IMF Bz: Z-component of Interplanetary Magnetic Field
ImpQs: Impact Quantities
IPS: Interplanetary Scintillation (radio observations)
ISES: International Space Environment Services
ISSI: International Space Science Institute
ISWA: CCMC Integrated Space Weather Analysis system — Software/Portal
ISWAT: International Space Weather Action Teams
ISWCF: International Space Weather Coordination Forum
IT: Information Technology
ITM: Ionosphere–Thermosphere–Mesosphere
JSON: JavaScript Object Notation — File format
LEO: Low-Earth Orbit
LWS: Living With a Star – NASA program
M2M SWAO: Moon-to-Mars Space Weather Analysis Office
MAE: Mean Absolute Error — Metrics
MAGE: Multiscale Atmosphere-Geospace Environment — Modeling Framework
ME: Mean Error — Metrics
MET: NCAR Model Evaluation Tool – Software
METplus: NCAR Model Evaluation Tool Plus – Software
MHD: Magnetohydrodynamics
ML: Machine Learning
MRE: Mean Relative Error — Metrics
MSE: Mean Squared Error — Metrics
NAIRAS: Nowcast of Aerospace Ionizing Radiation System — Model
NASA: National Aeronautics and Space Administration
NMHS: National Meteorological and Hydrological Service
NOAA: National Oceanic and Atmospheric Administration
NOAA RLs: NOAA Readiness Levels
NSF: National Science Foundation
NSF NCAR: NSF National Center for Atmospheric Research — Institution
O2R: Operations to Research
OSCAR: WMO Observing Systems Capability Analysis and Review Tool
OSEs: Observing System Experiments
OSSEs: Observing System Simulation Experiments
POD: Precise Orbit Determination
PSP: Parker Solar Probe — Spacecraft
PSW: COSPAR Panel on Space Weather

R2O: Research to Operations
R2O2R: Research to Operations to Research
REFM: Relativistic Electron Forecast Model
RINEX: Receiver INdependent EXchange format
RISER: Radio Investigations for Space Environment Research — UK Project
RL: Readiness Levels
RMSE: Root Mean Squared Error — Metrics
RO: Radio Occultation remote sensing technique
ROB: Royal Observatory of Belgium
ROT: Rate of Total Electron Content
ROTI: Rate of Total Electron Content Index
RT: Real-Time
SDO: Solar Dynamics Observatory — Spacecraft
SEEs: Single Event Effects
SEP: Solar Energetic Particles — Phenomenon
SEPVAL: SEP Model Validation — Project
SFTP: Secure File Transfer Protocol
SIFT: Solar Indices Forecasting Tool
SHINE: Solar, Heliospheric, and Interplanetary Environment — NSF program
SoHO: Solar and Heliospheric Observatory — Spacecraft
SolO: Solar Orbiter — Spacecraft
SPASE: Space Physics Archive Search and Extract
SPHINX: Solar Particles in the Heliosphere validation INfrastructure for SpWx — Software
SQL: Structured Query Language
SRAG: NASA Space Radiation Analysis Group
SSN: Solar Sunspot Number
STD: Standard Deviation — Metrics
SuperMAG: Global Ground-Based Magnetometer Initiative
SWE: ESA Space Weather
SWMF: Space Weather Modeling Framework — Modeling Framework
SWIMMR: Space Weather Instrumentation, Measurement, Modelling and Risk – UK program
SWPC: NOAA Space Weather Prediction Center — Institution
SWxTREC: Space Weather Technology, Research and Education Center
SWORD: Space Weather Operational Readiness Development — NASA Center of Excellence
TEC: Total Electron Content
TRLs: Technology Readiness Levels
VSWMC: Virtual Space Weather Modelling Centre — ESA program, Modeling Framework
WCMP2: WMO Core Metadata Profile Version 2
WICCT: WMO-ISES-COSPAR Coordination Team
WIGOS: WMO Integrated Global Observing System
WIPPS: WMO Integrated Processing and Prediction System
WIPSS: Worldwide IPS Stations — Network
WIS20: WMO Information System Version 2
WMO: World Meteorological Organisation
WSA: Wang-Sheeley-Arge — Model

Appendix B. Disambiguating terminology differences

We follow the National Research Council (2006) in using the term "**space weather enterprise**" (occasionally abbreviated to **"enterprise**") to mean all relevant stakeholders from both public and private sectors – operational space weather centers, forecasters, intermediaries, product users, researchers, R2O incubator centers, observation providers, policy makers, etc.).

At base, the term "**operational**" simply points to an organization's or utility's function (i.e., the state of working or functioning as intended). With respect to space weather forecasting, this term reflects the rigor and reliability inherent in the mission of weather forecast provision. More specifically, for the USA, as defined by the NOAA Administrative Order 216-105B[82], Operations refers to the "Sustained, systematic, reliable, and robust mission activities with an institutional commitment to deliver specified products and services.". Similar philosophies apply for equivalent nationally-mandated service providers worldwide. While space weather forecasts and situational awareness messaging are critical to the functions of space weather users, our use of the terms "operations", "operational", and the like, generally refer to the operational space weather forecast service providers, e.g., SWPC in the USA, Natural Resources Canada, MOSWOC in the UK, Royal Netherlands Meteorological Institute (KNMI), Royal Observatory of Belgium (ROB), Finnish Meteorological Institute (FMI), Austrian Space Weather Office (ASWO), South African National Space Agency (SANSA), National Institute of Information and Communications Technology (NICT) in Japan, Bureau of Meteorology (BoM) in Australia – and many more.

The term "**terrestrial weather**" is often used as a loose shorthand for terrestrial meteorology, encompassing timescales from weather (short) through seasonal (medium) to climate (long) – it is intended to distinguish (where applicable) the space and terrestrial communities. Where relevant, specific terrestrial subdomains are identified.

We follow Halford et al. 2019 in using the engineering definitions of the "**verification**" and "**validation**" terms, typically used in NASA and ESA[96] publications, and by some of the space weather modelling community. That is, the term "**verification**" means that a model conforms to its requirements (e.g., timeliness, output format, code quality, etc.). In contrast, the term "**validation**" refers to the comparison between a forecast and observation of a specific physical quantity using metrics that assess the differences. The term "**quality assessment**" refers to a broader concept. It may include validation but can also involve reviewing model results based on the theoretical analysis of the expected behavior of a physical system. Quality assess-

[96] Precise ESA definitions used in the context of ESA SWE Service Network are in the Terminology Glossary section of the ESA SWE Service portal, https://swe.ssa.esa.int/web/guest/glossary, and ultimately from the ECSS Glossary of terms (ECSS, 2023).[75].

ment, therefore, provides deeper insight into why a model may perform poorly and helps identify reasons for low validation scores (MacNeice et al., 2018).

To avoid reader confusion, we note however that the definitions of **"verification"** and **"validation"** terms are reversed in terrestrial weather – a convention since 1884 (JWGFVR, 2025). Many publications cited in this paper use the terrestrial convention – notably those used to illustrate terrestrial weather approaches, or from operational centers straddling space and terrestrial weather (e.g., NOAA SWPC, Met Office, etc.) or the WMO. Where the terrestrial weather convention is unavoidably used in text, the space weather convention is indicated in brackets thereafter – i.e. 'verification (validation)'.

"Ensemble" can have multiple meanings: most often this means multiple runs of a model with different initial conditions (e.g., as is common in terrestrial weather), to assess the "weather of the day" uncertainty. But this can have other meanings too. For instance, a "lagged ensemble" of forecasts issued by a given model at different lead times allows assessing the impact of evolving conditions (relevant e.g., for the CME forecasting case); a "perturbed physics" ensemble allows investigation of choices for different physics or parameterizations within a model; a "multi-model" ensemble can further explore the systemic uncertainty by evaluating wholly different models, or different implementations of similar models. In most of this paper, unless otherwise indicated, we use **"ensemble"** to mean "multi-model ensemble", assessing the various models present across the enterprise.

"Users" is a similarly thorny term. This often means "end-users", who use space weather services as one of many inputs into their decision-making processes (e.g., "can I service my grid today? Is there a risk of space weather-associated disruptions? What is the cost of delaying this?"). It is often also used to mean forecaster users in operational centers, evaluating model outputs and using these as inputs to derive products and services they then issue to end-users. Terrestrial weather has the useful term "intermediaries" for this case. We note some expert end users (for example engineers in the satellite sector with good awareness of the space weather environment, and model or observational limitations) are sometimes willing to receive operational model output directly, typically if the uncertainties are well-characterised. In this paper, we focus mainly on the chain where model output is presented to forecasters; as such **"users"** typically means "forecaster users".

Appendix C. Tracking Frameworks: Tables of definitions

These various tracking frameworks developed largely independently by various actors in the enterprise to meet their needs to assess R2O progress unsurprisingly share many similarities.

For convenient comparison of the various frameworks by the enterprise – those running R2O incubation efforts, and those being assessed by such frameworks – we have grouped the main frameworks in this paper, providing a top-level review in §7. In this appendix, we provide tables outlining the detailed levels within each framework. We leave attempting to map any equivalences between frameworks and their levels to enthusiastic readers, or a future effort.

C.1. ESA technology readiness levels (TRLs)

Table C1. Technology Readiness Level (TRL) definitions used by ESA (discussed in §7.1).

TRL	Level Description
1	Basic principle observed and reported
2	Technology concept or application formulated
3	The concept or application is proven through analysis and experimentation
4	Basic prototype validated in a laboratory environment
5	Basic prototypes validated in the relevant environment
6	System or subsystem model or prototype demonstrated in a relevant environment
7	System prototype demonstrated in a relevant environment
8	The actual system completed and qualified for flight through test and demonstration
9	The actual system has been proven through successful operations in space

C.2. ESA SWE network product maturity levels (PML)

Table C2. ESA SWE Network Product Maturity Level (PML) definitions (discussed in §7.2).

#	Name	Short Description
1	Basic Research	New ideas with potential interest for Space Weather situational awareness purposes.
2	Controlled environment	Preliminary testing to understand the validity of the product on a test case or in a controlled environment.
3	Prototype/ demonstrator	The product demonstrates its performance within a user-friendly interface.
4	Relevant environment	The service product is running in the relevant environment.
5	Monitored	The product is running stably with monitoring of the performance and availability.
6	In use	The product is running stably with monitoring and has recognised users (community).
7	Available for decision making	The product is running in a well-described environment according to agreed targets and is successfully used by recognised users (community) for decision-making purposes.

C.3. NOAA Readiness Levels (RLs)

Table C3. NOAA Readiness Level (RL) definitions (discussed in §7.3).

RL	NAO 216-105B Definition
1	Basic research, experimental or theoretical work undertaken primarily to acquire new knowledge of the underlying foundations of phenomena and observable facts, without any particular application or use in view. Basic research can be oriented or directed towards some broad fields of general interest, with the explicit goal of a range of future applications.
2	Applied research original investigation undertaken to acquire new knowledge. It is, however, directed primarily towards a specific, practical aim or objective. Applied research is undertaken either to determine possible uses for basic research findings or to determine new methods or ways of achieving specific and predetermined objectives.
3	Proof-of-concept for system, process, product, service, or tool; this can be considered an early phase of experimental development; feasibility studies may be included.
4	Successful evaluation of a system, subsystem, process, product, service, or tool in a laboratory or other experimental environment; this can be considered an intermediate development.
5	Successful evaluation of the system, subsystem, process, product, service, or tool in relevant environment through testing and prototyping; this can be considered the final stage of development before the demonstration begins.
6	Demonstration of a prototype system, subsystem, process, product, service, or tool in a relevant or test environment (potentially demonstrated).
7	The prototype system, process, product, service, or tool demonstrated in operational or other relevant environment (functionality demonstrated in near-real world environment; subsystem components fully integrated into system).
8	Finalized system, process, product, service, or tool tested and shown to operate or function as expected within user's environment; user training and documentation completed; operator or user approval given.
9	System, process, product, service, or tool deployed and used routinely.

C.3.1. Plain-language description of the NOAA Readiness Levels

In considering the application of the RLs to the specific domain of space weather, and as introduced by the 2022 space weather R2O2R Framework, plain-language definitions of RLs 4–8 (analogous to those provided in The Handbook) are provided below with the intent of clarifying the main effort associated with each level and the associated environment wherein the effort is undertaken.

- **RL 4: Validation of prototype capability in development environment** (e.g., laboratory or other experimental environment). Such validation results contributing to the establishment of RL 4 should sufficiently evince the potential for the emerging capability to advance over the current state-of-the-art.
- **RL 5: Validation of maturing capability in Proving Ground (PG) environment** (i.e., environment relevant to targeted end-use environment). Such objective validation aims to thoroughly characterize the expected performance of the capability in real world application and should include its integration with any necessary supporting elements. Such supporting resources are afforded via the PGs. Examples of PGs include the CCMC-SWPC Architecture for Collaborative Evaluation (ACE) at CCMC Cloud of NASA HPC, NASA's Moon-to-Mars Space Weather Analysis Office, and the Space Weather Prediction Testbed – SWPT.
- **RL 6: Demonstration of advanced capability in Testbed environment** (e.g., test environment, relevant end-use environment). Demonstration via the Space Weather Prediction Testbed toward the attainment of RL 6 is expected to focus on the determination and optimization of the ultimate impact of an advanced capability on the end-user. This occurs through broad engagement of the entire stakeholder chain (e.g., developers, scientists, forecasters, social scientists, decision makers, end-users) and affords iteration of critical components (e.g., visualization, products, etc.) and the concept of operations (conops).
- **RL 7: Demonstration of full-system prototype in operationally relevant environment** (e.g., near-real world environment, actual targeted operational environment, or other sufficiently operationally relevant environment). This preliminary full-system demonstration aims to probe the sufficiency of the realtime support elements (e.g., error handling, failure modes, monitoring systems, etc.) and to implement revisions toward attainment of established robustness requirements.
- **RL 8: Finalized full-system testing in targeted end-use environment** (i.e., operational environment). Successful completion of this final full system testing is expected to characterize performance on the end-user's targeted system and certifies reliability, robustness, and resilience, therein.

To reiterate, the preceding descriptions of RLs 4–8 are provided strictly toward illustrating the relationship between the main effort associated with the attainment of a given RL and the environment wherein the effort is undertaken. These descriptions should not be taken to represent the full scope of requirements comprising the readiness levels discussed.

C.4. NASA application readiness levels (ARLs)

Table C4. Application Readiness Level (ARL) definitions (discussed in §7.4).

Phase	ARL	Level description	Milestones
Discovery and Feasibility	1	Basic Research (Baseline Ideas)	Developed ideas for how specific research results could enhance decision-making Baseline support research identified and documented (i.e., results on the theory, models, remotely sensed products, and other current or planned measurements needed to support the application idea) – whether done by the PI's Team or not.
	2	Application Concept (Invention)	Independent application components formulated and created Decision-making activity to be enhanced by the application identified Plans to characterize better the decision-making activity developed.
	3	Proof of Application Concept (Viability Established)	A convincing case for the viability of the application concept made Detailed characterization of user decision-making process completed (e.g., pre-application baseline performance, mechanisms, and limitations) Components of the application tested and validated independently
Development, testing, and validation	4	Initial Integration and Verification (Prototype/Plan)	Components of the eventual application system were brought together, and technical integration issues worked out Organizational challenges and human process issues identified and managed.
	5	Validation in Relevant Environment (Potential Determined)	Application components integrated into a functioning prototype application system with realistic supporting elements The application system's potential to improve the decision-making activity determined and articulated (e.g., projected impacts on cost, functionality, delivery time, etc.)
	6	Demonstration in Relevant Environment (Potential Demonstrated)	Prototype application system beta-tested in a simulated operational environment Projected improvements in the performance of decision-making activity demonstrated in a simulated operational environment
Integration into the partner's system	7	Application Prototype in Partner's Decision-Making (Functionality Demonstrated)	Prototype application system integrated into end-users operational environment Prototype application functionality tested & demonstrated in decision-making activity
	8	Application Completed and Qualified (Functionality Proven)	Finalized application system was tested, proven operational, and shown to operate as expected within the user's environment Application qualified and approved by user for use in decision-making activity User documentation and training completed.
	9	Approved, Operational Deployment and Use in Decision Making (Sustained Use)	Sustained use of application system in a decision-making context

C.5. ISWAT Application Usability Levels (AULs)

Table C5. Application Usability Level (AUL) definitions (discussed in §7.5).

Phase	AUL	Level description
I) Discoverability and viability	1	Basic research
	2	Establishment of users and their requirements
	3	Assess viability and current state of the art
II) Development, testing, and validation	4	Initial integration and verification
	5	Demonstration in the relevant context
	6	Completed validation
III) Implementation and integration into operation	7	Application prototype
	8	Validation in relevant context
	9	Approved for on-demand use

References

Abbas, F., Ameen, M.A., 2023. Evaluation of ionospheric and solar proxy indices for IRI-plas 2020 model over pakistan and japan during different solar activity epochs. Adv. Space Res. 72 (12), 5551–5562. https://doi.org/10.1016/j.asr.2022.06.044.

Akmaev, R.A., Newman, A., Codrescu, M., Schulz, C., Nerney, E., 2010. D-RAP model validation: I. Scientific. Report https://www.ngdc.noaa.gov/stp/drap/DRAP-V-Report1.pdf.

Al Shidi, Q., Pulkkinen, T.I., Welling, D., Tóth, G., 2024. Accuracy of global geospace simulations: influence of solar wind monitor location and solar wind driving. Space Weather 22, e2023SW003747. https://doi.org/10.1029/2023SW003747.

Al Shidi, Q., Pulkkinen, T., Tóth, G., Brenner, A., Zou, S., Gjerloev, J., 2022. A large simulation set of geomagnetic storms—Can simulations predict ground magnetometer station observations of magnetic field perturbations?. Space Weather 20 e2022SW003049. https://doi.org/10.1029/2022SW003049.

Amerstorfer, T., Hinterreiter, J., Reiss, M.A., Möstl, C., Davies, J.A., Bailey, R.L., Weiss, A.J., Dumbović, M., Bauer, M., Amerstorfer, U. V., Harrison, R.A., 2021. Evaluation of CME arrival prediction using ensemble modeling based on heliospheric imaging observations. Space Weather 19, e2020SW002553. https://doi.org/10.1029/2020SW002553.

Amerstorfer, T., Möstl, C., Hess, P., Temmer, M., Mays, M.L., Reiss, M. A., Lowrance, P., Bourdin, P.-A., 2018. Ensemble prediction of a halo coronal mass ejection using heliospheric imagers. Space Weather 16 (7), 784–801. https://doi.org/10.1029/2017SW001786.

Anderson, J.L., Mertens, C.J., Grajewski, B., Luo, L., Tseng, C.Y., Cassinelli II., R.T., 2014. Flight attendant radiation dose from solar particle events. Aviat. Space Environ. Med. 85 (8), 828–832. https://doi.org/10.3357/ASEM.3989.2014.

Angling, M.J., Nogués-Correig, O., Nguyen, V., Vetra-Carvalho, S., Bocquet, F.X., Nordstrom, K., Melville, S.E., Savastano, G., Mohanty, S., Masters, D., 2021. Sensing the ionosphere with the Spire radio occultation constellation. J. Space Weather Space Clim. 11 (56), 56. https://doi.org/10.1051/swsc/2021040.

Arge, C.N., Pizzo, V.J., 2000. Improvement in the prediction of solar wind conditions using near-real time solar magnetic field updates. J. Geophys. Res. 105, 10465–10480. https://doi.org/10.1029/1999JA000262.

Arge, C.N., Henney, C.J., Hernandez, I.G., Toussaint, W.A., Koller, J., Godinez, H.C., 2013. Modeling the corona and solar wind using ADAPT maps that include far-side observations. Big Island, Hawaii, AIP Conf. Proc. 1539 (1), 11–14. https://doi.org/10.1063/1.4810977.

Arge, C.N., Odstrčil, D., Pizzo, V.J., Mayer, L.R., 2003. Improved method for specifying solar wind speed near the sun, Pisa, Italy. AIP Conf. Proc. 679 (1), 190–193. https://doi.org/10.1063/1.1618574.

Bain, H.M., Steenburgh, R.A., Onsager, T.G., Stitely, E.M., 2021. Operational solar flare forecasting with GOES X-ray sensor. Space Weather 19, e2020SW002670. https://doi.org/10.1029/2020SW002670.

Baker, D.N., 2000. The occurrence of operational anomalies in spacecraft and their relationship to space weather. IEEE Trans. Plasma Sci. 28 (6), 2007–2016. https://doi.org/10.1109/27.902228.

Baker, D.N., McPherron, R.L., Cayton, T.E., Klebesadel, R.W., 1990. Linear prediction filter analysis of relativistic electron properties at 6.6 R_E. J. Geophys. Res. 95 (A9), 15133–15140. https://doi.org/10.1029/JA095iA09p15133.

Balch, C.C., Jing, C., Kelbert, A., Arons, P., Richardson, K., 2023. Geoelectric field model validation in the Southern California Edison system: case study. In: 2023 IEEE Energy Conversion Congress and Exposition (ECCE). IEEE. pp. 6107–6114. doi: 10.1109/ECCE53617.2023.10362883.

Balch, C., Caher, M., Kobet, G., Grant, I., Kelbert, A., 2024. Verification of a 3-dimensional geoelectric field model for geomagnetic disturbance and geomagnetically induced current studies. C2-10439-2024. https://www.e-cigre.org/publications/detail/c2-10439-2024-verification-of-a-3-dimensional-geoelectric-field-model-for-geomagnetic-disturbance-and-geomagnetically-induced-current-studies.html.

Barnard, L., Owens, M.J., Scott, C.J., Lockwood, M., de Koning, C.A., Amerstorfer, T., Hinterreiter, J., Möstl, C., Davies, J.A., Riley, P., 2021. Quantifying the uncertainty in CME kinematics derived from geometric modeling of heliospheric imager data. Space Weather 19, e2021SW002841. https://doi.org/10.1029/2021SW002841.

Barnard, L., Owens, M., Scott, C., Lang, M., Lockwood, M., 2023. SIR-HUXt—A particle filter data assimilation scheme for CME time-elongation profiles. Space Weather 21, e2023SW003487. https://doi.org/10.1029/2023SW003487.

Barnes, D., Palmerio, E., Amerstorfer, T., Asvestari, E., Barnard, L., Bauer, M., Čalogović, J., Hess, P., Kay, C., Cappello, G., Kenny, K., 2025. Tomographic Inversion of Synthetic White-Light Images: Observing Coronal Mass Ejections in 3D. Sixth PUNCH (Polarimeter to UNify the Corona and Heliosphere) Science Meeting. Presentation: https://cpaess.ucar.edu/sites/default/files/punch-6-presentation/barnes-david.pdf; Recording: https://youtu.be/VPvgE-PN3sM?si=Az0cGjaZdS500EGo&t=3736.

Barnes, G., Leka, K.D., Schrijver, C.J., Colak, T., Qahwaji, R., Ashamari, O.W., Yuan, Y., Zhang, J., McAteer, R.T.J., Bloomfield, D.S., Higgins, P.A., Gallagher, P.T., Falconer, D.A., Georgoulis, M.K., Wheatland, M.S., Balch, C., Dunn, T., Wagner, E.L., 2016. A comparison of flare forecasting methods. I. Results from the "All-Clear" workshop. Astrophys. J. 829 (2), 89. https://doi.org/10.3847/0004-637X/829/2/89.

Bauer, M., Amerstorfer, T., Hinterreiter, J., Weiss, A.J., Davies, J.A., Möstl, C., Amerstorfer, U.V., Reiss, M.A., Harrison, R.A., 2021. Predicting CMEs using ELEvoHI with STEREO-HI beacon data. Space Weather 19, e2021SW002873. https://doi.org/10.1029/2021SW002873.

Bauer, P., Thorpe, A., Brunet, G., 2015. The quiet revolution of numerical weather prediction. Nature 525, 47–55. https://doi.org/10.1038/nature14956.

Beggan, C.D., Eastwood, J.P., Eggington, J.W.B., Forsyth, C., Freeman, M.P., Henley, E., Heyns, M., Hüber, T.J., Jackson, D.R., LaMoury, A.T., Richardson, G.S., Smith, A.W., Thomson, A.W.P., 2025. Implementing an operational cloud-based now- and forecasting system for space weather ground effects in the UK. Space Weather 23 (5), e2025SW004364. https://doi.org/10.1029/2025SW004364.

Bilitza, D., Altadill, D., Truhlik, V., Shubin, V., Galkin, I., Reinisch, B., Huang, X., 2017. International Reference Ionosphere 2016: from ionospheric climate to real-time weather predictions. Space Weather 15, 418–429. https://doi.org/10.1002/2016SW001593.

Bergin, A., Chapman, S.C., Gjerloev, J.W., 2020. *AE*, D_{ST}, and their SuperMAG counterparts: the effect of improved spatial resolution in geomagnetic indices. J. Geophys. Res. Space Phys. 125, e2020JA027828. https://doi.org/10.1029/2020JA027828.

Bessho, N., Chen, L.-J., Hesse, M., 2016. Electron distribution functions in the diffusion region of asymmetric magnetic reconnection. Geophys. Res. Lett. 43 (5), 1828–1836. https://doi.org/10.1002/2016GL067886.

Bialke and Hansell, 2017. A newly discovered branch of the fault tree explaining systemic reaction wheel failures and anomalies, European Space Mechanisms and Tribology Symposium, https://www.esmats.eu/esmatspapers/pastpapers/pdfs/2017/bialke.pdf.

Bialke, 2018. A Discussion of Friction Anomaly Signatures in Response to Electrical Discharge in Ball Bearings, European Space Mechanisms and Tribology Symposium, https://esmats.eu/amspapers/pastpapers/pdfs/2018/bialke.pdf.

Bisi, M.M., Shea, M.A., 2023. COSPAR space weather roadmap 2022–2024: scientific research and applications. Adv. Space Res. 72 (12), 5159–5160. https://doi.org/10.1016/j.asr.2023.10.045, ISSN: 0273-1177.

Bisi, M.M, Forte, B., Stephen, M., Jackson, D., Fallows, R.A., Jackson, B.V., Odstrčil, D., 2023. Radio Investigations for Space Environment Research (RISER): an overview. AGU Fall Meeting 2023, San Francisco, CA, 11–15 December 2023, Session: SPA-Aeronomy / Advances in Radio Frequency Propagation Modeling and Applications, Poster No. 309, id. SA43B-309, 2023AGUFMSA43B.309B. https://ui.adsabs.harvard.edu/abs/2023AGUFMSA43B.309B.

Bisi, M.M., Breen, A.R., Jackson, B.V., Fallows, R.A., Walsh, A.P., Mikić, Z., Riley, P., Owen, C.J., Gonzalez-Esparza, A., Aguilar-Rodriguez, E., Morgan, H., Jensen, E.A., Wood, A.G., Owens, M.J., Tokumaru, M., Manoharan, P.K., Chashei, I.V., Giunta, A.S., Linker, J.A., Shishov, V.I., Tyul'bashev, S.A., Agalya, G., Glubokova, S.K., Hamilton, M.S., Fujiki, K., Hick, P.P., Clover, J.M., Pintér, B., 2010. From the Sun to the Earth: the 13 May 2005 coronal mass ejection. Sol. Phys. 265, 49–127. https://doi.org/10.1007/s11207-010-9602-8.

Bisi, M.M., Fallows, R.A., Sobey, C., Eftekhari, T., Jensen, E.A., Jackson, B.V., Yu, H.-S., Hick, P.P., Odstrčil, D., Tokumaru, M., 2016. Worldwide Interplanetary Scintillation (IPS) and Heliospheric Faraday Rotation Plans and Progress. In: The IPSP South African National Space Agency Space Weather Research Forum, Hermanus, South Africa, 20 January.

Bisi, M.M., Gonzalez-Esparza, J.A., Jackson, B., Aguilar-Rodriguez, E., Tokumaru, M., Chashei, I., Tyul'bashev, S., Manoharan, P., Fallows, R., Chang, O., Yu, H.-S., Fujiki, K., Shishov, V., Barnes, D., 2017. The Worldwide Interplanetary Scintillation (IPS) Stations (WIPSS). Network in Support of Space-Weather Science and Forecasting. Paper Presented at Geophysical Res. Abstracts, EGU General Assembly.

Bloomfield, D.S., Higgins, P.A., McAteer, R.T.J., Gallagher, P.T., 2012. Toward reliable benchmarking of solar flare forecasting methods. ApJL 747, L41. https://doi.org/10.1088/2041-8205/747/2/L41.

Bobra, M.G., Sun, X., Hoeksema, J.T., Turmon, M., Liu, Y., Hayashi, K., Barnes, G., Leka, K.D., 2014. The helioseismic and magnetic imager (HMI) vector magnetic field pipeline: SHARPs – Space-Weather HMI active region patches. Sol. Phys. 289 (3549–3578), 2014. https://doi.org/10.1007/s11207-014-0529-3.

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. Sol. Terr. Phys. 177, 2–11. https://doi.org/10.1016/j.jastp.2017.03.014.

Boyd, A., O'Brien, T.P., Cox, J., Larsen, B., 2026. Environmental specification accuracy requirements for anomaly resolution in various orbits. Adv. Space Res. https://doi.org/10.1016/j.asr.2023.03.017, this issue.

Breen, A.R., Fallows, R.A., Bisi, M.M., Jones, R.A., Jackson, B.V., Kojima, M., Dorrian, G.D., Middleton, H.R., Thomasson, P., Wannberg, G., 2008. The Solar Eruption of 2005 May 13 and its Effects: long-baseline interplanetary scintillation observations of the earth-directed coronal mass ejection. ApJ 683 (L79), L79–L82. https://doi.org/10.1086/591520.

Bruinsma, S., Sutton, E., Solomon, S.C., Fuller-Rowell, T., Fedrizzi, M., 2018. Space weather modeling capabilities assessment: neutral density for orbit determination at low Earth orbit. Space Weather 16 (11), 1806–1816. https://doi.org/10.1029/2018SW002027.

Bruinsma, S., Boniface, C., Sutton, E., Fedrizzi, M., 2021. Thermosphere modeling capabilities assessment: geomagnetic storms. J. Space Weather Space Clim. 11, 12. https://doi.org/10.1051/swsc/2021002.

Bruinsma, S., Siemes, C., Emmert, J.T., Mlynczak, M.G., 2023. Description and comparison of 21st century thermosphere data. Adv. Space Res. 72 (12), 5476–5489. https://doi.org/10.1016/j.asr.2022.09.038.

Bruinsma, S., Laurens, S., 2024. Thermosphere model assessment for geomagnetic storms from 2001 to 2023. J. Space Weather Space Clim. 14, 28. https://doi.org/10.1051/swsc/2024027.

Bruinsma, S., Dudok de Wit, T., Fuller-Rowell, T., Garcia-Sage, K., Mehta, P., Schiemenz, F., Shprits, Y.Y., Vasile, R., Yue, J., Elvidge, S., 2026. Thermosphere and satellite drag. Adv. Space Res. https://doi.org/10.1016/j.asr.2023.05.011, this issue.

Burch, J.L., Phan, T.D., 2016. Magnetic reconnection at the dayside magnetopause: Advances with MMS. Geophys. Res. Lett. 43, 8327–8338. https://doi.org/10.1002/2016GL069787.

Burt, J., Smith, B., 2012. Deep space climate observatory: the DSCOVR mission. In: 2012 IEEE Aerospace Conference, pp. 1–13. https://doi.org/10.1109/AERO.2012.6187025.

Čalogović, J., Dumbović, M., Sudar, D., Vršnak, B., Martinić, K., Temmer, M., Veronig, A.M., 2021. Probabilistic drag-based ensemble model (DBEM) evaluation for heliospheric propagation of CMEs. Sol. Phys. 296, 114. https://doi.org/10.1007/s11207-021-01859-5.

Campi, C., Benvenuto, F., Massone, A.M., Bloomfield, D.S., Georgoulis, M.K., Piana, M., 2019. Feature ranking of active region source properties in solar flare forecasting and the uncompromised stochasticity of flare occurrence. Astrophys J 883 (2), 150. https://doi.org/10.3847/1538-4357/ab3c26.

Camporeale, E., Shprits, Y., Chandorkar, M., Drozdov, A., Wing, S., 2016. On the propagation of uncertainties in radiation belt simulations. Space Weather 14, 982–992. https://doi.org/10.1002/2016SW001494.

Camporeale, E., 2019. The challenge of machine learning in space weather: nowcasting and forecasting. Space Weather 17, 1166–1207. https://doi.org/10.1029/2018SW002061.

Cargill, P.J., 2004. On the aerodynamic drag force acting on interplanetary coronal mass ejections. Sol. Phys. 221, 135–149. https://doi.org/10.1023/B:SOLA.0000033366.10725.a2.

Carter, B.A., Yizengaw, E., Retterer, J.M., Francis, M., Terkildsen, M., Marshall, R., Norman, R., Zhang, K., 2014. An analysis of the quiet time day-to-day variability in the formation of postsunset equatorial plasma bubbles in the Southeast Asian region. J. Geophys. Res. Space Phys. 119, 3206–3223. https://doi.org/10.1002/2013JA019570.

Cash, M.D., Biesecker, D.A., Pizzo, V., de Koning, C.A., Millward, G., Arge, C.N., Henney, C.J., Odstrčil, D., 2015. Ensemble modeling of the 23 July 2012 coronal mass ejection. Space Weather 13, 611–625. https://doi.org/10.1002/2015SW001232.

Chand, A.E., Bowden, G., Brown, M., Ridley, A.J., 2025. Global sensitivity analysis of nitric oxide–related chemical reaction rates in the global ionosphere thermosphere model. Space Weather 23 (4), e2024SW004171. https://doi.org/10.1029/2024SW004171.

Chang, H., Morton, Y.J., Dittmann, T., Hunt, D., Durgonics, T., Braun, J., Weiss, J.-P., 2025. Assessment of scintillation data from PlanetiQ and Spire Global radio occultation missions. J. Geophys. Res. Space Phys. 130, e2024JA033543. https://doi.org/10.1029/2024JA033543.

Chen, H., Sachdeva, N., Huang, Z., van der Holst, B., Manchester IV, W. B., Jivani, A., Zou, S., Chen, Y., Huan, X., Tóth, G., 2024. Decent estimate of CME arrival time from a data-assimilated ensemble in the alfvén wave solar atmosphere model (DECADE-AWSoM). ESS Open Archive. https://doi.org/10.22541/essoar.172745166.65966095/v2.

Chen, J., Cargill, P.J., Palmadesso, P.J., 1997. Predicting solar wind structures and their geoeffectiveness. J. Geophys. Res. 102 (A7), 14701–14720. https://doi.org/10.1029/97JA00936.

Chen, M.W., O'Brien, T.P., Lemon, C.L., Guild, T.B., 2018. Effects of uncertainties in electric field boundary conditions for ring current simulations. J. Geophys. Res. Space Phys. 123, 638–652. https://doi.org/10.1002/2017JA024496.

Cherniak, I., Zakharenkova, I., Braun, J.J., Wu, Q., Pedatella, N.M., Schreiner, W.S., Weiss, J.-P., Hunt, D.C., 2021. Accuracy assessment of the quiet-time ionospheric F2 peak parameters as derived from COSMIC-2 multi-GNSS radio occultation measurements. J. Space Weather Space Clim. 11. https://doi.org/10.1051/swsc/2020080.

Choi, H.-S., Lee, J., Cho, K.-S., Kwak, Y.-S., Cho, I.-H., Park, Y.-D., Kim, Y.-H., Baker, D.N., Reeves, G.D., Lee, D.-K., 2011. Analysis of GEO spacecraft anomalies: space weather relationships. Space Weather 9, S06001. https://doi.org/10.1029/2010SW000597.

Chou, M.-Y., Lin, C.C.H., Tsai, H.F., Lin, C.Y., 2017. Ionospheric electron density inversion for Global Navigation Satellite Systems radio occultation using aided Abel inversions. J. Geophys. Res. Space Phys. 122, 1386–1399. https://doi.org/10.1002/2016JA023027.

Chou, M.-Y., Yue, J., McDonald, S., Sassi, F., Tate, J., Pedatella, N., Harvey, V.L., 2024. Modeling the post-midnight equatorial plasma bubbles with SAMI3/SD-WACCM-X: large-scale wave structure. J. Geophys. Res. Space Phys. 129, e2024JA033023. https://doi.org/10.1029/2024JA033023.

Chou, M.-Y., Yue, J., Wang, J., Huba, J. D., El Alaoui, M., Kuznetsova, M. M., Rastätter, L., Shim, J.S., Fang, T., Meng, X., Fuller-Rowell, D., Retterer, J.M., 2023. Validation of ionospheric modeled TEC in the equatorial ionosphere during the 2013 March and 2021 November geomagnetic storms. Space Weather 21, e2023SW003480. doi: 10.1029/2023SW003480.

Choudhary, D.P., 2019. Contemporary space weather need: going beyond the global oscillation network group. Space Weather 17, 375–378. https://doi.org/10.1029/2018SW002097.

Clarke, M., 1964. Two topics in radiophysics. i. some observations of discrete radio sources. ii. an investigation of ionospheric irregularities using a radio signal from an artificial satellite. University of Cambridge (Unpublished Ph.D. Thesis).

COARDS, 1995. Conventions for the standardization of NetCDF files. NOAA Ferret Support, Pacific Marine Environment Laboratory. https://ferret.pmel.noaa.gov/Ferret/documentation/coards-netcdf-conventions.

Coleman, C.J., 1998. A ray tracing formulation and its application to some problems in over-the-horizon radar. Radio Sci. 33 (4), 1187–1197. https://doi.org/10.1029/98RS01523.

Coleman, C.J., 2011. Point-to-point ionospheric ray tracing by a direct variational method. Radio Sci. 46 (5), RS5016. https://doi.org/10.1029/2011RS004748.

Coleman, C.J., 2017. Analysis and modeling of radio wave propagation. Cambridge University Press, Cambridge.

Corti, C., Kuznetsova, M.M., Reiss, M.A., Yue, J., Karpen, J., Arge, C. N., Bacchini, F., Bard, C., Bruinsma, S., Caplan, R.M., Daldorff, L.K. S., Deka, P.J., DeVore, C.R., Elvidge, S., Ganushkina, N., Huba, J.D., Jackson, B.V., Jordanova, V., Linker, J.A., Liu, H., Luhmann, J.G., Markidis, S., Mayank, P., Merkin, V., Moens, N., Odstrcil, D., Omelchenko, Y.A., Palmroth, M., Poedts, S., Ridley, A.J., Shou, Y., Tenishev, V., Themens, D.R., Tóth, G., Wang, W., Wilhelm, R.-P., Young, M.A., Cecconi, B., Chou, M.-Y., De Zeeuw, D., Delzanno, G. L., Didigu, C., El Alaoui, M., Fung, S. F., Green, J., Huang, Z., Jian, L.K., Landwer, L.J., Lesko, M., MacNeice, P., Masson, A., Mays, M. L., Mehta, P.M., Miesch, M.S., Palmerio, E., Petrenko, M., Provornikova, E., Rastätter, L., Rusaitis, L., Sachdeva, N., Samara, E., Sur, D., Taktakishvili, A., Topper, J., Tsui, T., Verbeke, C., Wang, J., Wiegand, C., Wiltberger, M., Zheng, Y., Bisi, M.M., Georgoulis, M.K., Kodikara, T., Pulkkinen, T., Chartier, A., da Silva, D., Faturahman, A., Garcia-Sage, K., Kondrashov, D., Ledvina, V.E., Liu, W., Pandey, C., Resnick, E., Shi, C., Weigel, R.S., Whitman, K., Zakharenkova, I., and Zhang, K., (2026) Advancing Heliophysics and Space Weather Modeling through Open Science. Submitted to *Space Weather Journal*, 2025SW004922, ESS Open Archive, January 12, 2026. doi: 10.22541/essoar.176824639.92354528/v1.

Culverwell, I.D., Healy, S.B., Elvidge, S., 2024. One-dimensional variational ionospheric retrieval using Radio Occultation bending angles: 1. Theory. Space Weather 22, e2023SW003572. https://doi.org/10.1029/2023SW003572.

Davies, E.E., Rüdisser, H.T., Amerstorfer, U.V., Möstl, C., Bauer, M., Weiler, E., Amerstorfer, T., Majumdar, S., Hess, P., Weiss. A.J., 2024. Flux Rope Modeling of the 2022 September 5 Coronal Mass Ejection Observed by Parker Solar Probe and Solar Orbiter from 0.07 to 0.69 au. ApJ 973 51. doi: 10.3847/1538-4357/ad64cb.

Delzanno, G.L., Borovsky, J.E., 2022. The need for a system science approach to global magnetospheric models. Front. Astron. Space Sci. 9, 808629. https://doi.org/10.3389/fspas.2022.808629.

Delzanno, G.L., Isola, B., Lao, C., Borovsky, J.E., Sorathia, K., Merkin, V.G., Koshkarov, O., McCubbin, A., Garretson, J., Arnold, H., Lin, D., 2025. Validation of a global geospace model with a systems science approach based on canonical correlation analysis. Geophys. Res. Lett. 52, e2025GL115589. https://doi.org/10.1029/2025GL115589.

DeVore, C.R., Boris, J.P., Sheeley Jr., N.R., 1984. The concentration of the large-scale solar magnetic field by a meridional surface flow. Sol. Phys. 92 (1–2), 1–14. https://doi.org/10.1007/BF00157230.

Devos, A., Verbeeck, C., Robbrecht, E., 2014. Verification of space weather forecasting at the Regional Warning Center in Belgium. J. Space Weather Space Clim. 4, A29. https://doi.org/10.1051/swsc/2014025.

Dimmock, A.P., Hietala, H., Zou, Y., 2020. Compiling magnetosheath statistical data sets under specific solar wind conditions: lessons learnt from the dayside kinetic southward IMF GEM challenge. Earth Space Sci. 7 (6), e2020EA001095. https://doi.org/10.1029/2020EA001095.

Drozdov, A.Y., Shprits, Y.Y., Aseev, N.A., Kellerman, A.C., Reeves, G. D., 2017. Dependence of radiation belt simulations to assumed radial diffusion rates tested for two empirical models of radial transport. Space Weather 15 (1), 150–162. https://doi.org/10.1002/2016SW001426.

Dumbović, M., Čalogović, J., Martinić, K., Vršnak, B., Sudar, D., Temmer, M., Veronig, A., 2021. Drag-based model (DBM). Tools for forecast of coronal mass ejection arrival time and speed. Front. Astron. Space Sci. 8, 639986. https://doi.org/10.3389/fspas.2021.639986.

Dumbović, M., Čalogović, J., Vršnak, B., Temmer, M., Mays, M.L., Veronig, A., Piantschitsch, I., 2018. The drag-based ensemble model (DBEM) for coronal mass ejection propagation. Astrophys J 854 (2), 180. https://doi.org/10.3847/1538-4357/aaaa66.

Dungey, J.W., 1961. Interplanetary magnetic field and the auroral zones. Phys. Rev. Lett. 6 (2), 47–48. https://doi.org/10.1103/PhysRevLett.6.47.

Eastes, R.W., McClintock, W.E., Burns, A.G., Anderson, D.N., Andersson, L., Aryal, S., et al., 2020. Initial observations by the GOLD mission. J. Geophys. Res. Space Phys. 125 (7), e2020JA027823. https://doi.org/10.1029/2020JA027823.

Eastes, R.W., McClintock, W.E., Burns, A.G., Anderson, D.N., Andersson, L., Codrescu, M., et al., 2017. The global-scale observations of the limb and disk (GOLD) mission. Space Sci. Rev. 212 (1–2), 383–408. https://doi.org/10.1007/s11214-017-0392-2.

Eastwood, J.P., Biffis, E., Hapgood, M.A., Green, L., Bisi, M.M., Bentley, R.D., Wicks, R., McKinnell, L.-A., Gibbs, M., Burnett, C., 2017. The economic impact of space weather: where do we stand? Risk Anal. 37, 206–218. https://doi.org/10.1111/risa.12765.

ECMWF, 2017. Anomaly correlation of ECMWF 500 hPa height forecasts. Retrieved from https://web.archive.org/web/20170318125441/ http://www.ecmwf.int/en/forecasts/charts/medium/anomaly-correlation-ecmwf-500hpa-height-forecasts (https://web.archive.org/web/20170318132842/ http://stream.ecmwf.int/data/atls12/data/data02/scratch/ps2png-atls12-a82bacafb5c306db76464b-c7e824bb75-JdNo7H.png) on 2026-01-04.

ECMWF, 2025. Anomaly Correlation Coefficient, https://confluence.ecmwf.int/display/FUG/Section+6.2.2+Anomaly+Correlation+Coefficient, webpage retrieved 2025-04-18.

ECSS, 2023. Glossary of terms of the European Cooperation for Space Standardization ECSS system, ECSS-S-ST-00-01C, Rev.1, October 11, 2023, https://ecss.nl/wp-content/uploads/2023/10/ECSS-S-ST-00-01C_Rev.1(11October2023).pdf.

El Alaoui, M., Richard, R.L., Ashour-Abdalla, M., Goldstein, M.L., 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 (12), 7752–7761.

Elvidge, S., Healy, S.B., Culverwell, I.D., 2024. One-dimensional variational ionospheric retrieval using radio occultation bending angles: 2. Validation. Space Weather 22, e2023SW003571. https://doi.org/10.1029/2023SW003571.

Evans, J.S., Lumpe, J.D., Eastes, R.W., Correira, J., Aryal, S., Laskar, F., et al., 2024. Disk images of neutral temperature from the global-scale observations of the limb and disk (GOLD) mission. J. Geophys. Res. Space Phys. 129, e2024JA032424. https://doi.org/10.1029/2024JA032424.

Fallows, R.A., Iwai, K., Jackson, B.V., Zhang, P., Bisi, M.M., Zucca, P., 2023. Application of novel interplanetary scintillation visualisations using LOFAR: a case study of merged CMEs from September 2017. Adv. Space Res. 72 (12), 5311–5327. https://doi.org/10.1016/j.asr.2022.08.076, ISSN 0273-1177.

Ferranti, L., Corti, S., Janousek, M., 2015. Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector. Q.J.R. Meteorol. Soc. 141, 916–924. https://doi.org/10.1002/qj.2411.

Fiori, R.A.D., Kumar, V., Boteler, D.H., Terkildsen, M.B., 2022. Occurrence of moderate to severe level space weather conditions likely to impact high frequency radio wave propagation used by aviation. J. Space Weather Space Clim. 12 (21). https://doi.org/10.1051/swsc/2022017.

Florios, K., Kontogiannis, I., Park, S.-H., Guerra, J.A., Benvenuto, F., Bloomfield, D.S., Georgoulis, M.K., 2018. Forecasting solar flares using magnetogram-based predictors and machine learning. Sol. Phys. 293 (2), 28. https://doi.org/10.1007/s11207-018-1250-4.

Forte, B., Allbrook, T., Arnold, A., Astin, I., Vani, B.C., Monico, J.F.G., Shimabukuro, M.H., Koulouri, A., John, H.M., 2026. Methodology for the characterisation of the impact of TEC fluctuations and scintillation on ground positioning quality over South America and North Europe, with implications for forecasts. Adv. Space Res. 2026. https://doi.org/10.1016/j.asr.2024.02.033, this issue.

Fox, N.J., Velli, M.C., Bale, S.D., Decker, R., Driesman, A., Howard, R.A., Kasper, J.C., Kinnison, J., Kusterer, M., Lario, D., Lockwood, M.K., McComas, D.J., Raouafi, N.E., Szabo, A., 2016. The Solar Probe Plus mission: humanity's first visit to our star. Space Sci. Rev. 204 (1–4), 7–48. https://doi.org/10.1007/s11214-015-0211-6.

Fuller-Rowell, D., Fang, T. W., Steenburgh, R., Fuller-Rowell, T., 2022. Observation- and model-driven GNSS-related products at SWPC and plans for the future. 44th COSPAR Scientific Assembly (Vol. 44, p. 3472). https://ui.adsabs.harvard.edu/abs/2022cosp...44.3472F/abstract.

Fung, S.F., Masson, A., Bargatze, L.F., King, T., Ringuette, R., Candey, R.M., Wiegand, C., Jian, L.K., De Zeeuw, D., Muglach, K., McGranaghan, R.M., Aaron Roberts, D., Cecconi, B., André, N., Génot, V., Vandegriff, J., Reiss, M.A., 2023. SPASE metadata as a building block of a heliophysics science-enabling framework. Adv. Space Res. 72 (12), 5707–5752. https://doi.org/10.1016/j.asr.2023.09.066.

Fusaro, R.L., 1994. Lubrication of Space Systems, NASA TM-106392. https://ntrs.nasa.gov/api/citations/19940024896/downloads/19940024896.pdf.

Gallagher, P.T., Moon, Y.J., Wang, H., 2002. Active-region monitoring and flare forecasting – I. Data processing and first results. Sol. Phys. 209, 171–183. https://doi.org/10.1023/A:1020950221179.

Ganushkina, N.Y., Sillanpää, I., Welling, D.T., Haiducek, J., Liemohn, M., Dubyagin, S., Rodriguez, J.V., 2019. Validation of Inner Magnetosphere Particle Transport and Acceleration Model (IMPTAM) with long-term GOES MAGED measurements of keV electron fluxes at geostationary orbit. Space Weather 17, 687–708. https://doi.org/10.1029/2018SW002028.

Gapper, G.R., Hewish, A., Purvis, A., Duffett-Smith, P.J., 1982. Observing interplanetary disturbances from the ground. Nature 296, 633–636. https://doi.org/10.1038/296633a0.

Génot, V., Budnik, E., Jacquey, C., Bouchemit, M., Renard, B., Dufourg, N., André, N., Cecconi, B., Pitout, F., Lavraud, B., Fedorov, A., Ganfloff, M., Plotnikov, I., Modolo, R., Lormant, N., Si Hadj Mohand, H., Tao, C., Besson, B., Heulet, D., Boucon, D., Durand, J., Bourrel, N., Brzustowski, Q., Jourdane, N., Hitier, R., Garnier, P., Grison, B., Aunai, N., Jeandet, A., Cabrolie, F., 2021. Automated Multi-Dataset Analysis (AMDA): an on-line database and analysis tool for heliospheric and planetary plasma data. Planet. Space Sci. 201, 105214. https://doi.org/10.1016/j.pss.2021.105214.

Gjerloev, J.W., 2009. A global ground-based magnetometer initiative. Eos Trans. AGU 90 (27), 230–231. https://doi.org/10.1029/2009EO270002.

Georgoulis, M.K., Rust, D.M., 2007. Quantitative forecasting of Major Solar Flares. ApJ 661 (1), L109. https://doi.org/10.1086/518718.

Georgoulis, M.K., Bloomfield, D.S., Piana, M., Massone, A.M., Soldati, M., Gallagher, P.T., Pariat, E., Vilmer, N., Buchlin, E., Baudin, F., Csillaghy, A., Sathiapal, H., Jackson, D.R., Alingery, P., Benvenuto, F., Campi, C., Florios, K., Gontikakis, C., Guennou, C., Guerra, J.A., Kontogiannis, I., Latorre, V., Murray, S.A., Park, S.-H., von Stachelski, S., Torbica, A., Vischi, D., Worsfold, M., 2021. The flare likelihood and region eruption forecasting (FLARECAST) project: flare forecasting in the big data and machine learning era. J. Space Weather Space Clim. 11, 39. https://doi.org/10.1051/swsc/2021023.

Gilleland, E., Ahijevych, D., Brown, B.G., Casati, B., Ebert, E.E., 2009. Intercomparison of spatial forecast verification methods. Weather Forecast. 24 (5), 1416–1430. https://doi.org/10.1175/2009WAF2222269.1.

Gonzalez, W.D., Tsurutani, B.T., Clúa de Gonzalez, A.L., 1999. Interplanetary origin of geomagnetic storms. Space Sci. Rev. 88, 529–562. https://doi.org/10.1023/A:1005160129098.

Good, S.W., Kilpua, E.K.J., LaMoury, A.T., Forsyth, R.J., Eastwood, J.P., Möstl, C., 2019. Self-similarity of ICME flux ropes: observations by radially aligned spacecraft in the inner heliosphere. J. Geophys. Res. Space Phys. 124, 4960–4982. https://doi.org/10.1029/2019JA026475.

Gordeev, E., Sergeev, V., Honkonen, I., Kuznetsova, M., Rastätter, L., Palmroth, M., Janhunen, P., Tóth, G., Lyon, J., Wiltberger, M., 2015. Assessing the performance of community available global MHD models using key system parameters and empirical relationships. Space Weather 13 (12), 868–884. https://doi.org/10.1002/2015SW001307.

Gregory, J., 2003 The CF metadata standard. https://cfconventions.org/Data/cf-documents/overview/article.pdf.

Guo, J., Wang, B., Whitman, K., Plainaki, C., Zhao, L., Bain, H.M., Cohen, C., Dalla, S., Dumbovic, M., Janvier, M., Jun, I., Luhmann, J., Malandraki, O.E., Leila Mays, M., Rankin, J.S., Wang, L., Zheng, Y., 2026. Particle radiation environment in the heliosphere: status, limitations, and recommendations. Adv. Space Res. https://doi.org/10.1016/j.asr.2024.03.070, this issue.

Haiducek, J.D., Welling, D.T., Ganushkina, N.Y., Morley, S.K., Ozturk, D.S., 2017. SWMF global magnetosphere simulations of January 2005: geomagnetic indices and cross-polar cap potential. Space Weather 15 (12), 1567–1587. https://doi.org/10.1002/2017SW001695.

Haiducek, J.D., Welling, D.T., Morley, S.K., Ganushkina, N.Y., Chu, X., 2020. Using multiple signatures to improve accuracy of substorm identification. J. Geophys. Res. Space Phys. 125 (4), e2019JA027559. https://doi.org/10.1029/2019JA027559.

Hajj, G.A., Romans, L.J., 1998. Ionospheric electron density profiles obtained with the global positioning system: results from the GPS/MET experiment. Radio Sci. 33, 175–190. https://doi.org/10.1029/97RS03183.

Halford, A.J., Kellerman, A.C., Garcia-Sage, K., Klenzing, J., Carter, B.A., McGranaghan, R.M., Guild, T., Cid, C., Henney, C.J., Ganushkina, N.Y., Burrell, A.G., Terkildsen, M., Welling, D.T., Murray, S.A.,

Leka, K.D., McCollough, J.P., Thompson, B.J., Pulkkinen, A., Fung, S.F., Bingham, S., Bisi, M.M., Liemohn, M.W., Walsh, B.M., Morley, S.K., 2019. Application usability levels: a framework for tracking project product progress. J. Space Weather Space Clim. 2019 (9), A34. https://doi.org/10.1051/swsc/2019030.

Hands, A.D.P., Ryden, K.A., Meredith, N.P., Glauert, S.A., Horne, R.B., 2018. Radiation effects on satellites during extreme space weather events. Space Weather 16, 1216–1226. https://doi.org/10.1029/2018SW001913.

Harrison, R.A., Davies, J.A., Biesecker, D., Gibbs, M., 2017. The application of heliospheric imaging to space weather operations: lessons learned from published studies. Space Weather 15, 985–1003. https://doi.org/10.1002/2017SW001633.

Hassell, D., Gregory, J., Blower, J., Lawrence, B.N., Taylor, K.E., 2017. A data model of the climate and Forecast metadata conventions (CF-1.6) with a software implementation (cf-python v2.1). Geosci. Model Dev. 10, 4619–4646. https://doi.org/10.5194/gmd-10-4619-2017.

Hathaway, D.H., Upton, L.A., Mahajan, S.S., 2022. Variations in differential rotation and meridional flow within the Sun's surface shear layer 1996–2022. Front. Astron. Space Sci. 9 (419), 1007290. https://doi.org/10.3389/fspas.2022.1007290.

Hegde, D.V., Kim, T.K., Pogorelov, N.V., Jones, S.V., Arge, C.N., 2025. Ensemble modeling of the solar wind flow with boundary conditions governed by synchronic photospheric magnetograms. I. Multipoint validation in the inner heliosphere. Astrophys. J. 988, 154. https://doi.org/10.3847/1538-4357/addf33.

Henley, E.M, Jackson, D.R., Bisi, M.M., Kuznetsova, M.M., Reiss, M.A., 2026. Looking Down: Where Terrestrial Weather May Help Space Weather, submitted to Advances in Space Research, this issue.

Henley, E. M., and Andries, J. A. (2025). Formats and Metadata for operational space weather data: progress on clarifying scopes and next steps. CGMS-53 presentation: https://www.cgms-info.org/agendas/agendas/CGMS53WG; presentation retrieved on 2025-04-19 from https://www.cgms-info.org/Agendas/GetWpFile.ashx?wid=5d1bee1f-200a-4192-9e17-c5fc14ae5c33&aid=df1a6fc9-59dc-4461-8bed-512fae79a1b7.

Henley, E.M., Bingham, S., and Bocquet, F.-X., 2024. Model verification at the Met Office: closing the R2O2R loop with continuous verification. ESWW presentation retrieved on 2025-04-19 from https://esww2024.org/wp-content/uploads/fusion-forms/672c7fe293a7b.pptx.

Henley, E.M., Pope, E.C.D., 2017. Cost-loss analysis of ensemble solar wind forecasting: space weather use of terrestrial weather tools. Space Weather 15, 1562–1566. https://doi.org/10.1002/2017SW001758.

Henney, C.J., Hock, R.A., Schooley, A.K., Toussaint, W.A., White, S.M., Arge, C.N., 2015. Forecasting solar extreme and far ultraviolet irradiance. Space Weather 13 (141–153), 2015. https://doi.org/10.1002/2014SW001118.

Hesse, M., Aunai, N., Sibeck, D., Birn, J., 2014. On the electron diffusion region in planar, asymmetric, systems. Geophys. Res. Lett. 41 (24), 8673–8680. https://doi.org/10.1002/2014GL061586.

Hewish, A., Scott, P.F., Wills, D., 1964. Interplanetary scintillation of small diameter radio sources. Nature 203, 1214–1217. https://doi.org/10.1038/2031214a0.

Hickmann, K.S., Godinez, H.C., Henney, C.J., Arge, C.N., 2015. Data assimilation in the ADAPT photospheric flux transport model. Sol. Phys. 290, 1105–1118. https://doi.org/10.1007/s11207-015-0666-3.

Hietala, H., Dimmock, A.P., Zou, Y., Garcia-Sage, K., 2020. The challenges and rewards of running a geospace environment modeling challenge. J. Geophys. Res. Space Phys. 125 (3), e2019JA027642. https://doi.org/10.1029/2019JA027642.

Hill, F., 2018. The global oscillation network group facility—an example of research to operations in space weather. Space Weather 16 (10), 1488–1497. https://doi.org/10.1029/2018SW002001.

Hoque, M.M., Yuan, L., Prol, F.S., Hernández-Pajares, M., Notarpietro, R., Jakowski, N., Olivares Pulido, G., Von Engeln, A., Marquardt, C., 2023. A new method of electron density retrieval from MetOp-A's truncated radio occultation measurements. Rem. Sens. (Basel) 15 (5), 1424. https://doi.org/10.3390/rs15051424.

Hsu, C.-T., Matsuo, T., Wang, W., Liu, J.-Y., 2014. Effects of inferring unobserved thermospheric and ionospheric state variables by using an Ensemble Kalman Filter on global ionospheric specification and forecasting. J. Geophys. Res. Space Phys. 119, 9256–9267. https://doi.org/10.1002/2014JA020390.

Hu, S., Kim, M.-H.-Y., McClellan, G.E., Cucinotta, F.A., 2009. Modeling the acute health effects of astronauts from exposure to large solar particle events. Health Phys. 96 (4), 465–476.

Hua, M., Bortnik, J., Kellerman, A.C., Camporeale, E., Ma, Q., 2023. Ensemble modeling of radiation belt electron acceleration by chorus waves: dependence on key input parameters. Space Weather 21, e2022SW003234. https://doi.org/10.1029/2022SW003234.

Huang, X., Reinisch, B.W., 1982. Automatic calculation of electron density profiles from digital ionograms 2. true height inversion of topside ionograms with the profile-fitting method. Radio Sci. 17, 837–844. https://doi.org/10.1029/RS017i004p00837.

Hudson, H.S., 2020. Solar flare build-up and release. Sol. Phys. 295, 132. https://doi.org/10.1007/s11207-020-01698-w.

ICAO, 2019. Manual on Space Weather Information in Support of International Air Navigation. Doc 10100, 1st edition, ISBN 978-92-9258-662-1, International Civil Aviation Organization.

Ishii, M., Berdermann, J., Forte, B., Hapgood, M., Bisi, M.M., Romano, V., 2026a. Space weather impact on radio communication and navigation. Adv. Space Res. 2026. https://doi.org/10.1016/j.asr.2024.01.043, this issue.

Ishii, M., Costa, J.E.R., Kuznetsova, M.M., Andries, J., Gopalswamy, N., Belehaki, A., Alfonsi, L., Shiokawa, K., Stanislawska, I., Bingham, S., Kalegaev, V., Tobiska, W.K., Rees, D., Glover, A., Spann, J.F., 2026b. Pathways to global coordination in space weather: international organizations, initiatives, and space agencies. Adv. Space Res. 2026. https://doi.org/10.1016/j.asr.2024.06.017, this issue.

Ishii, M., Costa, J.E.R., Kuznetsova, M.M., Bisi, M.M., Mezae, A., Veronig, A.M., Denardini, C.M., Plainaki, C., Dasso, S., Temmer, M., Molinalm, M.G., Berdermann, Yoon, K., Gonzalez-Esparza, J.A., Valdivia, J.A., Supnithi, P., Marshall, R., Bruinsma, S.L., Onsager, T., Lanabere, V., Pötzi, W., Bouya, Z., Rodger, C.J., Bothmer, V., Luo, B., Nandy, D., Martini, D., Wintoft, P., Kero, J., Jackson, D., Kholodkov, K., Rabiu, B., Cid, C., Alfonsi, L., Raulin, J.P., Boteler, D., Castill, D.A.S., Tulunay, Y., 2026c. Global landscape of space weather observations, research and operations. Adv. Space Res. 2026. https://doi.org/10.1016/j.asr.2025.06.021, this issue.

Iwai, K., Fallows, R.A., Bisi, M.M., Shiota, D., Jackson, B.V., Tokumaru, M., Fujiki, K., 2023. Magnetohydrodynamic simulation of coronal mass ejections using interplanetary scintillation data observed from radio sites ISEE and LOFAR. Adv. Space Res. 72 (12), 5328–5340. https://doi.org/10.1016/j.asr.2022.09.028.

Jackson, B.V., Buffington, A., Cota, L., Odstrčil, D., Bisi, M.M., Fallows, R.F., Tokumaru, M., 2020. Iterative tomography: a key to providing time-dependent 3-D reconstructions of the inner heliosphere and the unification of space weather forecasting techniques. Front. Astron. Space Sci. 7, 568429. https://doi.org/10.3389/fspas.2020.568429.

Jackson, B.V., Odstrčil, D., Yu, H.-S., Hick, P.P., Buffington, A., Mejia-Ambriz, J.C., Kim, J., Hong, S., Kim, Y., Han, J., Tokumaru, M., 2015. The UCSD IPS solar wind boundary and its use in the ENLIL 3D-MHD prediction model. Space Weather 13, 104–115. https://doi.org/10.1002/2014SW001130.

Jackson, B.V., Tokumaru, M., Fallows, R.A., Bisi, M.M., Fujiki, K., Chashei, I., Bashev, S. Tyul, Chang, O., Barnes, D., Buffington, A., Cota, L., Bracamontes, M., 2023. Interplanetary scintillation (IPS) analyses during LOFAR campaign mode periods that include the first three Parker Solar Probe close passes of the Sun. Adv. Space Res. 72 (12), 5341–5360. https://doi.org/10.1016/j.asr.2022.06.029.

Jakowski, N., Hoque, M.M., 2019. Estimation of spatial gradients and temporal variations of the total electron content using ground-based GNSS measurements. Space Weather 17, 339–356. https://doi.org/10.1029/2018SW002119.

Jakowski, N., Wehrenpfennig, A., Heise, S., Reigber, C., Luhr, H., Grunwaldt, L., Meehan, T.K., 2002. GPS radio occultation measurements of the ionosphere from CHAMP: early results. Geophys. Res. Lett. 29 (10), 5. https://doi.org/10.1029/2001GL014364.

Jensen, E.A., Hick, P.P., Bisi, M.M., Jackson, B.V., Clover, J., Mulligan, T., 2010. Faraday rotation response to coronal mass ejection structure. Sol. Phys. 265, 31–48. https://doi.org/10.1007/s11207-010-9543-2.

Jensen, T. 2023. Building a Community Verification and Diagnostics Framework for the UFS. UFS seminar series, https://www.ufs.epic.noaa.gov/2023/03/metplus-building-a-community-verification-and-diagnostics-framework-for-the-ufs/ Recording: https://youtu.be/xMnrznyjR50 ; slides: https://www.ufs.epic.noaa.gov/wp-content/uploads/2023/03/UFS-Seminar-Jensen-METplus-9Mar23.pptx.

Jensen, T., Halley Gotway, J., Harrold, M., Vigh, J., Row, M., Kalb, C., Fisher, H., Goodrich, L., Adriaansen, D., Win-Gildenmeister, M., McCabe, G., Prestopnik, J., Opatz, J., Frimel, J., Blank, L., Arbetter, T., 2024. The METplus Version 6.0.0 User's Guide. Developmental Testbed Center. Available at: https://github.com/dtcenter/METplus/releases.

Jolliffe, I.T., Stephenson, D.B., 2011. Forecast Verification: A Practitioner's Guide in Atmospheric Science, 2nd ed. Wiley, New Jersey, USA.

Jones, W.R., Jansen, M.J., 2006. Lubrication for Space Applications. Handbook of Lubrication and Tribology: Volume I Application and Maintenance, Second Edition. Boca Raton, CRC Press, ISBN: 9780849320958. doi: 10.1201/9781420003840.

Jun, I., Garrett, H., Kim, W., Zheng, Y., Fung, S.F., Corti, C., Ganushkina, N., Guo, J., 2026. A review on radiation environment pathways to impacts: radiation effects, relevant empirical environment models, and future needs. Adv. Space Res. 2026. https://doi.org/10.1016/j.asr.2024.03.079.

JWGFVR (2025). Forecast verification FAQ webpage, WMO WWRP/WGNE Joint WorkingGroup on Forecast Verification Research (JWGFVR), https://cawcr.gov.au/projects/verification/#FAQs, retrieved 2025-04-17.

Kahler, S.W., Darsey, H., 2021. Exploring contingency skill scores based on event sizes. Space Weather 19, e2020SW002604. https://doi.org/10.1029/2020SW002604.

Kaiser, M.L., 2005. The STEREO mission: an overview. Adv. Space Res. 36 (8), 1483–1488. https://doi.org/10.1016/j.asr.2004.12.066.

Kalafatoglu Eyiguler, E.C., Shim, J.S., Kuznetsova, M.M., Kaymaz, Z., Bowman, B.R., Codrescu, M.V., et al., 2019. Quantifying the storm time thermospheric neutral density variations using model and observations. Space Weather 17 (2), 269–284. https://doi.org/10.1029/2018SW002033.

Kay, C., Davies, E., 2025. H2-02 The magnetic structure of CMEs. Presentation at ISWAT 2025. https://docs.google.com/presentation/d/1xw-l_x9mbOGlq-foW2wqUOLwYRHqK0OhW1X3C9avekc.

Kay, C., Gopalswamy, N., Reinard, A., Opher, M., 2017. Predicting the magnetic field of earth-impacting CMEs. ApJ 835, 117. https://doi.org/10.3847/1538-4357/835/2/117.

Kay, C., Palmerio, E., Riley, P., Mays, M.L., Nieves-Chinchilla, T., Romano, M., et al., 2024. Updating measures of CME arrival time errors. Space Weather 22, e2024SW003951. https://doi.org/10.1029/2024SW003951.

Keller, C.U., Harvey, J.W., Giampapa, M.S., 2003. SOLIS: an innovative suite of synoptic instruments. In: Keil, S.L., Avakyan, S.V. (Eds.), Innovative Telescopes and Instrumentation for Solar Astrophysics, Vol. 4853. Society of Photo-Optical Instrumentation Engineers (SPIE), pp. 194–204. https://doi.org/10.1117/12.460373.

Kelley, M.C., Wong, V.K., Aponte, N., Coker, C., Mannucci, A.J., Komjathy, A., 2009. Comparison of COSMIC occultation-based electron density profiles and TIP observations with Arecibo incoherent scatter radar data. Radio Sci. 44, RS4011. https://doi.org/10.1029/2008RS004087.

King, J.H., Papitashvili, N.E., 2005. Solar wind spatial scales in and comparisons of hourly Wind and ACE plasma and magnetic field data. J. Geophys. Res. Space Phys. 110, A02104. https://doi.org/10.1029/2004JA010649.

Klenzing, J., Halford, A.J., Kellerman, A., Thompson, B., 2023. Using application usability levels to support tracking the health of heliophysics. Front. Astron. Space Sci. 10, 1144053. https://doi.org/10.3389/fspas.2023.1144053.

Kniezewski, K. L., S. J. Schonfeld, and C. J. Henney (2024). Nowcasting Solar EUV Irradiance With Photospheric Magnetic Fields and the Mg II Index, Space Weather, 22, e2023SW003772, 2024. DOI: 10.1029/2023SW003772.

Kooi, J.E., Wexler, D.B., Jensen, E.A., Kenny, K.N., Nieves-Chinchilla, T., Wilson III, L.B., Wood, B.E., Jian, L.K., Fung, S.F., Pevtsov, A., Gopalswamy, N., Manchester, W.B., 2022. Modern faraday rotation studies to probe the solar wind. Front. Astron. Space Sci. 9, 841866. https://doi.org/10.3389/fspas.2022.841866.

Kraaikamp, E., Verbeeck, C., 2015. Solar Demon – an approach to detecting flares, dimmings, and EUV waves on SDO/AIA images. J. Space Weather Space Clim. 5, A18. https://doi.org/10.1051/swsc/2015019.

Kress, B.T., Rodriguez, J.V., Buzulukova, N.Y., Redmon, R.J., Machol, J.L., Fiorello, J.L., Roza, M.A., Meloy, R.M., 2024. Relationship between GOES-R series spacecraft operational anomalies and in situ 30 eV–3-MeV electron measurement. IEEE Trans. Plasma Sci. 52 (5), 1610–1618. https://doi.org/10.1109/TPS.2024.3390658.

Krieger, A.S., Timothy, A.F., Roelof, E.C., 1973. A coronal hole and its identification as the source of a high velocity solar wind stream. Sol. Phys. 29, 505–525. https://doi.org/10.1007/BF00150828.

Kubicka, M., Möstl, C., Amerstorfer, T., Boakes, P.D., Feng, L., Eastwood, J.P., Törmänen, O., 2016. Prediction of geomagnetic storm strength from inner heliospheric in situ observations. ApJ 833, 255. https://doi.org/10.3847/1538-4357/833/2/255.

Kubyshkina, M., Sergeev, V.A., Tsyganenko, N.A., Zheng, Y., 2019. Testing efficiency of empirical, adaptive, and global MHD magnetospheric models to represent the geomagnetic field in a variety of conditions. Space Weather 17, 672–686. https://doi.org/10.1029/2019SW002157.

Kusano, K., Iju, T., Bamba, Y., Inque, S., 2020. A physics-based method that can predict imminent large solar flares. Science 369 (6503), 587–591. https://doi.org/10.1126/science.aaz2511.

Kuznetsova, M. M., Bisi, M. M., Opgenoorth, H., Reiss, M. A., Mays, M.. Leila, Alfonsi, L., Arge, C. Nick, Barnes, D., Belehaki, A., Berger, T. E., Brecht, A., Bruinsma, S., Chang, O., Corti, C., Davies, E., Dayeh, M. A., DiBraccio, G., Drozdov, A. Y., Frissell, Garcia-Sage, K., N. A., Guo, J., Henney, C. J., Henley, E., Hoque, M., Hozumi, K., Jun, I., Kahre, M., Kay, C., Klimchuk, J. A., Krista, L., Mehta, P. M., Morgan, J., Muglach, K., Perri, B., Vanlommel, P., Verbeke, C., Vourlidas, A., Resnick, E., Samara, E., Sanchez-Cano, B., Tsagouri, I., Yue, J., Zhang, P., Zheng, Y., Zucca, P., Beser, K., Bingham, S., Blum, T. L., Chhabra, S., D'Huys, E., Egeland, R., Fung, S. F., Glover, A., Grande, M., Gonzi, S., Halford, A., Heynderickx, D., Janssens, J., Linker, J., Ma, Y., Mann, I. R., Majumdar, S., Masson, A., Mendoza, M., Mevius, M., Mertens, C., Murray, S. A., Nandi, D., Paouris, E., Pevtsov, A. A., Posner, A., Richardson, I. G., Robinson, R., Sachdeva, N., Schiewe, T., Scolini, C., Temmer, M., Tóth, G., Wang, D., Whitman, K., Wiegand, C., Yue, C., Zhao, L., Levisohn, M., 2026. The COSPAR International Space Weather Action Teams (ISWAT) Initiative: A Driving Force for the COSPAR Space Weather Roadmap, submitted to Advances in Space Research, this issue.

Lapenta, G., Berchem, J., Zhou, M., Walker, R.J., El-Alaoui, M., Goldstein, M.L., Paterson, W.R., Giles, B.L., Pollock, C.J., Russell, C. T., Strangeway, R.J., Ergun, R.E., Khotyaintsev, Y.V., Torbert, R.B., Burch, J.L., 2017. On the origin of the crescent-shaped distributions observed by MMS at the magnetopause. J. Geophys. Res. Space Phys. 122 (2), 2024–2039. https://doi.org/10.1002/2016JA023290.

Laskar, F.I., Sutton, E.K., Lin, D., Greer, K.R., Aryal, S., Cai, X., Pedatella, N.M., Eastes, R.W., Wang, W., Codrescu, M.V., McClintock, W.E., 2023. Thermospheric temperature and density variability during 3–4 February 2022 minor geomagnetic storm. Space Weather 21, e2022SW003349. https://doi.org/10.1029/2022SW003349.

Lee, E., Wilber, M., Parks, G.K., Min, K.W., Lee, D.-Y., 2004. Modeling of remote sensing of thin current sheet. Geophys. Res. Lett. 31 (21), L21801. https://doi.org/10.1029/2004GL020331.

Lei, J., Syndergaard, S., Burns, A.G., Solomon, S.C., Wang, W., Zeng, Z., Roble, R.G., Wu, Q., Kuo, Y.-H., Holt, J.M., Zhang, S.-R., Hysell, D. L., Rodrigues, F.S., Lin, C.H., 2007. Comparison of COSMIC ionospheric measurements with ground-based observations and model predictions: preliminary results. J. Geophys. Res. Space Phys. 112 (A7), A07308. https://doi.org/10.1029/2006JA012240.

Leka, K.D., Barnes, G., 2018. Solar flare forecasting: present methods and challenges. In: Buzulukova, N. (Ed.), Extreme events in geospace: Origins, predictability, and consequences. Elsevier, Amsterdam, pp. 65–98. https://doi.org/10.1016/B978-0-12-812700-1.00003-0.

Leka, K.D., Park, S.-H., Kusano, K., Andries, J., Barnes, G., Bingham, S., Bloomfield, D.S., McCloskey, A.E., Delouille, V., Falconer, D., Gallagher, P.T., Georgoulis, M.K., Kubo, Y., Lee, K., Lee, S., Lobzin, V., Mun, J., Murray, S.A., Hamad Nageem, T.A.M., Qahwaji, R., Sharpe, M., Steenburgh, R.A., Steward, G., Terkildsen, M., 2019a. A comparison of flare forecasting methods. III. Systematic behaviors of operational solar flare forecasting systems. Astrophys. J. 881 (2), 101. https://doi.org/10.3847/1538-4357/ab2e11.

Leka, K.D., Park, S.-H., Kusano, K., Andries, J., Barnes, G., Bingham, S., Bloomfield, D.S., McCloskey, A.E., Delouille, V., Falconer, D., Gallagher, P.T., Georgoulis, M.K., Kubo, Y., Lee, K., Lee, S., Lobzin, V., Mun, J., Murray, S.A., Hamad Nageem, T.A.M., Qahwaji, R., Sharpe, M., Steenburgh, R.A., Steward, G., Terkildsen, M., 2019b. A comparison of flare forecasting methods. ii. benchmarks, metrics, and performance results for operational solar flare forecasting systems. ApJS 243, 36. https://doi.org/10.3847/1538-4365/ab2e12.

Liemohn, M., Ganushkina, N.Y., De Zeeuw, D.L., Rastätter, L., Kuznetsova, M., Welling, D.T., Tóth, G., Ilie, R., Gombosi, T.I., van der Holst, B., 2018. Real-time SWMF at CCMC: assessing the Dst output from continuous operational simulations. Space Weather 16, 1583–1603. https://doi.org/10.1029/2018SW001953.

Liemohn, M.W., Shane, A.D., Azari, A.R., Petersen, A.K., Swiger, B.M., Mukhopadhyay, A., 2021. RMSE is not enough: guidelines to robust data-model comparisons for magnetospheric physics. J. Atmos. Sol. Terr. Phys. 218, 105624. https://doi.org/10.1016/j.jastp.2021.105624.

Liemohn, M.W., Rastätter, L., Halford, A.J., Zheng, Y., Garcia-Sage, K. S., Redmon, R., Vines, S.K., 2025. Guide for conducting "community challenges" in space physics. Earth Space Sci. 12, e2024EA004138. https://doi.org/10.1029/2024EA004138.

Lin, C.H., Liu, J.Y., Fang, T.W., Chang, P.Y., Tsai, H.F., Chen, C.H., Hsiao, C.C., 2007a. Motions of the equatorial ionization anomaly crests imaged by FORMOSAT-3/COSMIC. Geophys. Res. Lett. 34, L19101. https://doi.org/10.1029/2007GL030741.

Lin, C.H., Hsiao, C.C., Liu, J.Y., Liu, C.H., 2007b. Longitudinal structure of the equatorial ionosphere: time evolution of the four-peaked EIA structure. J. Geophys. Res. 112, A12305. https://doi.org/10.1029/2007JA012455.

Lin, C.-Y., Lin, C.-C.-H., Liu, J.-Y., Rajesh, P.K., Matsuo, T., Chou, M.-Y., et al., 2020. The early results and validation of FORMOSAT-7/COSMIC-2 space weather products: global ionospheric specification and Ne-aided Abel electron density profile. J. Geophys. Res. Space Phys. 125, e2020JA028028. https://doi.org/10.1029/2020JA028028.

Linker, J.A., Heinemann, S.G., Temmer, M., Owens, M.J., Caplan, R.M., Arge, C.N., Asvestari, E., Delouille, V., Downs, C., Hofmeister, S.J., Jebaraj, I.C., Madjarska, M.S., Pinto, R.F., Pomoell, J., Samara, E., Scolini, C., Vršcnak, B., 2021. Coronal hole detection and open magnetic flux. Astrophys. J. 918 (1), 21. https://doi.org/10.3847/1538-4357/ac090a.

Linker, J.A., Mikić, Z., Biesecker, D.A., Forsyth, R.J., Gibson, S.E., Lazarus, A.J., Lecinski, A., Riley, P., Szabo, A., Thompson, B.J., 1999. Magnetohydrodynamic modeling of the solar corona during whole Sun month. J. Geophys. Res. 104, 9809–9830. https://doi.org/10.1029/1998JA900159.

Liu, G., Rowland, D.E., Gan, Q., Liu, H.-L., Klenzing, J.H., England, S. L., Eastes, R.W., 2023. Thermospheric temperature and ΣO/N2 variations as observed by GOLD and compared to MSIS and WACCM-X simulations during 2019–2020 at deep solar minimum. J. Geophys. Res. Space Phys. 128, e2023JA031560. https://doi.org/10.1029/2023JA031560.

Loto'aniu, P.T.M., Romich, K., Rowland, W., Codrescu, S., Biesecker, D., Johnson, J., Singer, H.J., Szabo, A., Stevens, M., 2022. Validation of the DSCOVR spacecraft mission space weather solar wind products. Space Weather 20, e2022SW003085. https://doi.org/10.1029/2022SW003085.

Lugaz, N., Al-Haddad, N., Zhuang, B., Möstl, C., Davies, E.E., Farrugia, C.J., Banu, S.A., Weiler, E., Galvin, A.B., 2025. The need for a sub-L1 space weather research mission: current knowledge gaps on coronal mass ejections. Space Weather 23, e2024SW004189. https://doi.org/10.1029/2024SW004189.

Macha, M., Senajova, D., Giles, T., Calviani, M., Girard, S., Ferrari, M., 2025. Effects of radiation dose on lubricants: a review of experimental studies. ACS Appl. Mat. Interf. 17 (10), 14773–14800. https://doi.org/10.1021/acsami.4c21220.

MacNeice, P., 2009a. Validation of community models: identifying events in space weather model timelines. Space Weather 7 (6), S06004. https://doi.org/10.1029/2009SW000463.

MacNeice, P., 2009b. Validation of community models: 2. Development of a baseline using the Wang-Sheeley-Arge model. Space Weather 7 (12), S12002. https://doi.org/10.1029/2009SW000489.

MacNeice, P., 2018. On the need to automate support for quality assessment studies of space weather models. Space Weather 16 (11), 1627–1634. https://doi.org/10.1029/2018SW002039.

MacNeice, P., Jian, L.K., Antiochos, S.K., Arge, C.N., Bussy-Virat, C.D., DeRosa, M.L., Jackson, B.V., Linker, J.A., Mikić, Z., Owens, M.J., Ridley, A.J., Riley, P., Savani, N., Sokolov, I., 2018. Assessing the quality of models of the ambient solar wind. Space Weather 16 (11), 1644–1667. https://doi.org/10.1029/2018SW002040.

Mankins, J.C., 2009. Technology readiness assessments: a retrospective. Acta Astronaut. 65 (9), 1216–1223. https://doi.org/10.1016/j.actaastro.2009.03.058.

Mankins, J.C., 1995. Technology readiness levels. A White Paper, Advanced Concepts Office, Office of Space Access and Technology, NASA.

Manu, V., Balan, N., Zhang, Q.-H., Xing, Z.-Y., 2023. Double superposed epoch analysis of geomagnetic storms and corresponding solar wind and IMF in solar cycles 23 and 24. Space Weather 21, e2022SW003314. https://doi.org/10.1029/2022SW003314.

Masson, A., Fung, S.F., Camporeale, E., Kuznetsova, M.M., Poedts, S., Barnum, J., Ringuette, R., De Zeeuw, D., Polson, S., Sadykov, V.M., Navarro, V., Thomas, B., Caplan, R.M., Linker, J., Rastätter, L., Wiegand, C., McGranaghan, R.M., Petrenko, M., Didigu, C., Reerink, J., Ireland, J., Cecconi, B., 2026. Heliophysics and space weather information architecture and innovative solutions: current status and ways forward. Adv. Space Res. https://doi.org/10.1016/j.asr.2024.05.052, this issue.

Mays, M.L., Taktakishvili, A., Pulkkinen, A., MacNeice, P.J., Rastätter, L., Odstrčil, D., Jian, L.K., Richardson, I.G., LaSota, J.A., Zheng, Y., Kuznetsova, M.M., 2015. Ensemble modeling of CMEs using the WSA-ENLIL+Cone model. Sol. Phys. 290 (6), 1775–1814. https://doi.org/10.1007/s11207-015-0692-1.

Mazur, J.E., O'Brien, T.P., 2012. Comment on "Analysis of GEO spacecraft anomalies: space weather relationships" by Ho-Sung Choi et al. Space Weather 10, S03003. https://doi.org/10.1029/2011SW000738.

Meadors, G.D., Jones, S.I., Hickmann, K.S., Arge, C.N., Godinez-Vasquez, H.C., Henney, C.J., 2020. Data assimilative optimization of WSA source surface and interface radii using particle filtering. Space Weather 18, e2020SW002464. https://doi.org/10.1029/2020SW002464.

Mehta, P.M., Paul, S.N., Crisp, N.H., Sheridan, P.L., Siemes, C., March, G., Bruinsma, S., 2023. Satellite drag coefficient modeling for thermosphere science and mission operations. Adv. Space Res. 72 (12), 5443–5459. https://doi.org/10.1016/j.asr.2022.05.064.

Meier, M.M., Copeland, K., Matthiä, D., Mertens, C.J., Schennetten, K., 2018. First steps toward the verification of models for the assessment of the radiation exposure at aviation altitudes during quiet space weather conditions. Space Weather 16, 1269–1276. https://doi.org/10.1029/2018SW001984.

Mikić, Z., Linker, J.A., Schnack, D.D., Lionello, R., Tarditi, A., 1999. Magnetohydrodynamic modeling of the global solar corona. Phys. Plasmas 6 (5), 2217–2224. https://doi.org/10.1063/1.873474.

Millward, G., Biesecker, D., Pizzo, V., de Koning, C.A., 2013. An operational software tool for the analysis of coronagraph images: determining CME parameters for input into the WSA-Enlil heliospheric model. Space Weather 11, 57–68. https://doi.org/10.1002/swe.20024.

Millward, G., Miesch, M., Gopala, M., Martinkus, C., Englyst, A., Marsh, M., Bocquet. F.-X., 2024. PyCAT: The SWPC/UKMO Python CME Analysis Tool. Poster at the Space Weather Workshop. https://cpaess.ucar.edu/sites/default/files/poster-2024-sww/Millward-George.pdf.

Milošić, D., Temmer, M., Heinemann, S.G., Poladchikova, T., Veronig, A., Vršnak, B., 2023. Improvements to the Empirical Solar Wind Forecast (ESWF) model. Sol. Phys. 298, 45. https://doi.org/10.1007/s11207-022-02102-5.

Minow, J.I., Jordanova, V.K., Pitchford, D., Ganushkina, N.Y., Zheng, Y., Luca Delzanno, G., Jun, I., Kim, W., 2026. ISWAT spacecraft surface charging review, Advances in Space Research, 2026, this issue. doi: 10.1016/j.asr.2024.08.058.

Morgan, J., McCauley, P.I., Waszewski, A., Ekers, R., Chhetri, R., 2023. Detection and characterization of a coronal mass ejection using interplanetary scintillation measurements from the murchison wide-field array. Space Weather 21, e2022SW003396. https://doi.org/10.1029/2022SW003396.

Morley, S.K., Brito, T.V., Welling, D.T., 2018a. Measures of model performance based on the log accuracy ratio. Space Weather 16, 69–88. https://doi.org/10.1002/2017SW001669.

Morley, S.K., Welling, D.T., Woodroffe, J.R., 2018b. Perturbed input ensemble modeling with the space weather modeling framework. Space Weather 16, 1330–1347. https://doi.org/10.1029/2018SW002000.

Morley, S.K., 2020. Challenges and opportunities in magnetospheric space weather prediction. Space Weather 18 (3), e2018SW002108. https://doi.org/10.1029/2018SW002108.

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. https://doi.org/10.1002/2017SW001735.

Müller, D., St. Cyr, O.C., Zouganelis, I., Gilbert, H.R., Marsden, R., Nieves-Chinchilla, T., Antonucci, E., Auchère, F., Berghmans, D., Horbury, T.S., Howard, R.A., Krucker, S., Maksimovic, M., Owen, C. J., Rochus, P., Rodriguez-Pacheco, J., Romoli, M., Solanki, S.K., Bruno, R., Carlsson, M., Fludra, A., Harra, L., Hassler, D.M., Livi, S., Louarn, P., Peter, H., Schuhle, U., Teriaca, L., del Toro Iniesta, J. C., Wimmer-Schweingruber, R.F., March, E., Velli, M., De Groof, A., Walsh, A., Williams, D., 2020. The solar orbiter mission. Sci. Overview. Astron. Astrophys. 642, A1. https://doi.org/10.1051/0004-6361/202038467.

Murray, S.A., 2018. The importance of ensemble techniques for operational space weather forecasting. Space Weather 16, 777–783. https://doi.org/10.1029/2018SW001861.

Murray, S.A., Bingham, S., Sharpe, M., Jackson, D.R., 2017. Flare forecasting at the met office space weather operations centre. Space Weather 15, 577–588. https://doi.org/10.1002/2016SW001579.

National Research Council, 2000. From Research to Operations in Weather Satellites and Numerical Weather Prediction: Crossing the Valley of Death. Washington, DC: The National Academies Press. doi: 10.17226/9948.

National Research Council, 2003. Satellite Observations of the Earth's Environment: Accelerating the Transition of Research to Operations. Washington, DC: The National Academies Press. doi: 10.17226/10658.

National Research Council, 2006. Completing the Forecast: Characterizing and 2682 Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts (p. 2683 11699). National Academies Press. Washington, D.C. https://doi.org/10.17226/11699

Nelson, A., Koizumi, K., Uzo-Okoro, E., Erickson, M., Winn, M., Meehan, J., 2022. Space Weather Research-to-Operations and Operations-to-Research Framework, National Science & Technology Council, USA. https://web.archive.org/web/20250117192757/ https://www.whitehouse.gov/wp-content/uploads/2022/03/03-2022-Space-Weather-R2O2R-Framework.pdf.

Nguyen, G., Aunai, N., Fontaine, D., Le Pennec, E., Van den Bossche, J., Jeandet, A., Bakkali, B., Vignoli, L., Regaldo-Saint Blancard, B., 2019. Automatic detection of interplanetary coronal mass ejections from in situ data: a deep learning approach. Astrophys. J. 874 (2), 145. https://doi.org/10.3847/1538-4357/ab0d24.

Niehof, J.T., Morley, S.K., Welling, D.T., Larsen, B.A., 2022. The SpacePy space science package at 12 years. Front. Astron. Space Sci. 9, 1023612. https://doi.org/10.3389/fspas.2022.1023612.

Núñez, M., Nieves-Chinchilla, T., Pulkkinen, A., 2016. Prediction of shock arrival times from CME and flare data. Space Weather 14 (8), 544–562. https://doi.org/10.1002/2016SW001361.

Odstrčil, D., 2003. Modeling 3-D solar wind structure. Adv. Space Res. 32 (4), 497–506.

Odstrčil, D., 2023. Heliospheric 3-D MHD ENLIL simulations of multi-CME and multi-spacecraft events. Front. Astron. Space Sci. 10, 1226992. https://doi.org/10.3389/fspas.2023.1226992.

Odstrčil, D., Pizzo, V.J., 1999a. Three-dimensional propagation of CMEs in a structured solar wind flow: 1. CME launched within the streamer belt. J. Geophys. Res. 104, 483–492. https://doi.org/10.1029/1998JA900019.

Odstrčil, D., Pizzo, V.J., 1999b. Three-dimensional propagation of coronal mass ejections in a structured solar wind flow: 2. CME launched adjacent to the streamer belt. J. Geophys. Res. 104, 493–504. https://doi.org/10.1029/1998JA900038.

Odstrčil, D., Riley, P., Zhao, X.P., 2004. Numerical simulation of the 12 May 1997 interplanetary CME event. J. Geophys. Res. Space Phys. 109 (A2), A02116. https://doi.org/10.1029/2003JA010135.

Odstrčil, D., Pizzo, V.J., Arge, C.N., 2005. Propagation of the 12 May 1997 interplanetary coronal mass ejection in evolving solar wind structures. J. Geophys. Res. 110, A02106. https://doi.org/10.1029/2004JA010745.

Ogilvie, K.W., Desch, M.D., 1997. The Wind spacecraft and its early scientific results. Adv. Space Res. 20 (4–5), 559–568. https://doi.org/10.1016/S0273-1177(97)00439-0.

Olechowski, A.L., Joglekar, N., Eppinger, S.D., 2015. Technology readiness levels at 40: A study of state-of-the-art use, challenges, and opportunities. In: 2015 Portland International Conference on Management of Engineering and Technology (PICMET), August 2-6 2015, Institute of Electrical and Electronics Engineers (IEEE), September 2015. doi: 10.1109/PICMET.2015.7273196.

Opgenoorth, H.J., Robinson, R., Ngwira, C.M., Garcia Sage, K., Kuznetsova, M., El Alaoui, M., Boteler, D., Gannon, J., Weygand, J., Merkin, V., Nykyri, K., Kosar, B., Welling, D., Eastwood, J., Eggington, J., Heyns, M., Kaggwa Kwagala, N., Sur, D., Gjerloev, J., 2026. Earth's geomagnetic environment—progress and gaps in understanding, prediction, and impacts. Adv. Space Res., 2026, this issue. doi: 10.1016/j.asr.2024.05.016.

Owens, M.J., 2018. Time-window approaches to space-weather forecast metrics: a solar wind case study. Space Weather 16 (11), 1847–1861. https://doi.org/10.1029/2018SW002059.

Owens, M.J., Riley, P., 2017. Probabilistic solar wind forecasting using large ensembles of near-Sun conditions with a simple one-dimensional "upwind" scheme. Space Weather 15 (11), 1461–1474. https://doi.org/10.1002/2017SW001679.

Owens, M.J., 2020. Coherence of Coronal Mass Ejections in Near-Earth Space. Sol Phys 295, 148. https://doi.org/10.1007/s11207-020-01721-0.

Owens, M.J., Arge, C.N., Spence, H.E., Pembroke, A., 2005. An event-based approach to validating solar wind speed predictions: high-speed enhancements in the Wang–Sheeley–Arge model. J. Geophys. Res.: Space Phys. 110 (A12), A12105. https://doi.org/10.1029/2005JA011343.

Owens, M.J., Spence, H.E., McGregor, S., Hughes, W.J., Quinn, J.M., Arge, C.N., Riley, P., Linker, J., Odstrčil, D., 2008. Metrics for solar wind prediction models: comparison of empirical, hybrid, and physics-based schemes with 8 years of L1 observations. Space Weather 6 (8), S08001. https://doi.org/10.1029/2007SW000380.

Owens, M.J., Challen, R., Methven, J., Henley, E., Jackson, D.R., 2013. A 27 day persistence model of near-earth solar wind conditions: a long lead-time forecast and a benchmark for dynamical models. Space Weather 11 (5), 225–236. https://doi.org/10.1002/swe.20040.

Owens, M.J., Riley, P., Horbury, T.S., 2017. Probabilistic solar wind and geomagnetic forecasting using an analogue ensemble or "Similar Day" approach. Sol. Phys. 292 (69), 2017. https://doi.org/10.1007/s11207-017-1090-7.

Öztürk, D.S., Garcia-Sage, K., Connor, H.K., 2020. All hands on deck for ionospheric modeling. EOS 101. https://doi.org/10.1029/2020EO144365.

Palmerio, E., Scolini, C., Barnes, D., Magdalenić, J., West, M.J., Zhukov, A.N., Rodriguez, L., Mierla, M., Good, S.W., Morosan, D.E., Kilpua, E.K.J., Pomoell, J., Poedts, S., 2019. Multipoint study of successive coronal mass ejections driving moderate disturbances at 1 au. Astrophys. J. 878 (1), 37. https://doi.org/10.3847/1538-4357/ab1850.

Paouris, E., Mavromichalaki, H., 2017. Effective acceleration model for the arrival time of interplanetary shocks driven by coronal mass ejections. Sol. Phys. 292 (12), 180. https://doi.org/10.1007/s11207-017-1212-2.

Park, S.-H., Leka, K.D., Kusano, K., Andries, J., Barnes, G., Bingham, S., Bloomfield, D.S., McCloskey, A.E., Delouille, V., Falconer, D., Gallagher, P.T., Georgoulis, M.K., Kubo, Y., Lee, K., Lee, S., Lobzin, V., Mun, J., Murray, S.A., Hamad Nageem, T.A.M., Qahwaji, R., Sharpe, M., Steenburgh, R.A., Steward, G., Terkildsen, M., 2020. A comparison of flare forecasting methods. IV. Evaluating consecutive-day forecasting patterns. Astrophys J 890 (2), 124. https://doi.org/10.3847/1538-4357/ab65f0.

Pesnell, W.D., Thompson, B.J., Chamberlin, P.C., 2012. The Solar Dynamics Observatory (SDO). Sol. Phys. 275 (1–2), 3–15. https://doi.org/10.1007/s11207-011-9841-3.

Pevtsov, A.A., Nandy, D., Usoskin, I., Pevtsov, A.A., Corti, C., Lefèvre, L., Owens, M., Li, G., Krivova, N., Saha, C., Perri, B., Brun, A.S., Strugarek, A., Dayeh, M.A., Nagovitsyn, Y.A., Erdélyi, R., 2026. Long-term solar variability: ISWAT S1 cluster review for COSPAR space weather roadmap. Adv. Space Res. https://doi.org/10.1016/j.asr.2023.08.034, this issue.

Phoenix, D.B., Mertens, C.J., Gronoff, G.P., Tobiska, K., 2024. Characterization of radiation exposure at aviation flight altitudes using the Nowcast of Aerospace Ionizing Radiation System (NAIRAS). Space Weather 22, e2024SW003869. https://doi.org/10.1029/2024SW003869.

Pogorelov, N.V., Arge, C.N., Caplan, R.M., Colella, P., Linker, J.A., Singh, T., Van Straalen, B., Upton, L., Downs, C., Gebhart, C., Hegde, D.V., Henney, C., Jones, S., Johnston, C., Kim, T.K., Marble, A., Raza, S., Stulajter, M.M., Turtle, J., 2024. Space weather with quantified uncertainties: improving space weather predictions with data-driven models of the solar atmosphere and inner heliosphere. J. Phys. Conf. Series, IOP Publishing 2742, 012013. https://doi.org/10.1088/1742-6596/2742/1/012013.

Pomoell, J., Poedts, S., 2018. EUHFORIA: European heliospheric forecasting information asset. J. Space Weather Space Clim. 8, A35. https://doi.org/10.1051/swsc/2018020.

Pope, E., 2016. Verification of predictions of CME arrival time at L1. Retrieved from https://ccmc.gsfc.nasa.gov/static/files/esww2016_cme_verfctn.pdf ; accessed 2025-05-07.

Prikryl, P., Jayachandran, P.T., Mushini, S.C., Richardson, I.G., 2014. High-latitude GPS phase scintillation and cycle slips during high-speed solar wind streams and interplanetary coronal mass ejections: a superposed epoch analysis. Earth Planet Sp 66, 62. https://doi.org/10.1186/1880-5981-66-62.

Pulkkinen, A., Bernabeu, E., Thomson, A., Viljanen, A., Pirjola, R., Boteler, D., Eichner, J., Cilliers, P.J., Welling, D., Savani, N.P., Weigel, R.S., Love, J.J., Balch, C., Ngwira, C.M., Crowley, G., Schultz, A., Kataoka, R., Anderson, B., Fugate, D., Simpson, J.J., MacAlester, M., 2017. Geomagnetically induced currents science, engineering, and applications readiness. Space Weather 15 (7), 828–856. https://doi.org/10.1002/2016SW001501.

Pulkkinen, A., Rastätter, L., Kuznetsova, M., Singer, H., Balch, C., Weimer, D., Tóth, G., Ridley, A., Gombosi, T., Wiltberger, M., Raeder, J., Weigel, R., 2013. Community-wide validation of geospace model ground magnetic field perturbation predictions to support model transition to operations. Space Weather 11, 369–385. https://doi.org/10.1002/swe.20056.

Rasca, A.P., Farrell, W.M., MacDowall, R.J., Bale, S.D., Kasper, J.C., 2021. Near-sun switchback boundaries: dissipation with solar distance. Astrophys J 916, 84. https://doi.org/10.3847/1538-4357/ac079f.

Rastätter, L., Shim, J.S., Kuznetsova, M.M., Kilcommons, L.M., Knipp, D.J., Codrescu, M., Fuller-Rowell, T., Emery, B., Weimer, D.R., Cosgrove, R., Wiltberger, M., Raeder, J., Li, W., Tóth, G., Welling, D., 2016. GEM-CEDAR challenge: poynting flux at DMSP and modeled Joule heat. Space Weather 14 (2), 113–135. https://doi.org/10.1002/2015SW001238.

Rastätter, L., Wiegand, C.P., Mullinix, R.E., MacNeice, P.J., 2019. Comprehensive assessment of models and events using library tools (CAMEL) framework: time series comparisons. Space Weather 17 (6), 845–860. https://doi.org/10.1029/2018SW002043.

Raza, S.A.Z., Singh, T., Pogorelov, N.V., 2025. The effect of uncertainties in reproducing the ambient solar wind at Earth on forecasting CME arrival times. Space Weather 23, e2025SW004371. https://doi.org/10.1029/2025SW004371.

Reinisch, B.W., Huang, H., 1983. Automatic calculation of electron density profiles from digital ionograms: 3. Proc. Bottoms. Ionogr. Radio Sci. 18 (3), 477–492. https://doi.org/10.1029/RS018i003p00477.

Reinisch, B., Galkin, I., 2011. Global ionospheric radio observatory (GIRO). Earth Planets Space 63 (4), 377–381. https://doi.org/10.5047/eps.2011.03.001.

Reiss, M.A., Muglach, K., Möstl, C., Arge, C.N., Bailey, R., Delouille, V., Garton, T.M., Hamada, A., Hofmeister, S., Illarionov, E., Jarolim, R., Kirk, M.S.F., Kosovichev, A., Krista, L., Lee, S., Lowder, C., MacNeice, P.J., Veronig, A.COSPAR ISWAT Coronal Hole Boundary Working Team, 2021a. The observational uncertainty of coronal hole boundaries in automated detection schemes. Astrophys. J. 913 (1), 28. https://doi.org/10.3847/1538-4357/abf2c8.

Reiss, M.A., Möstl, C., Bailey, R.L., Rüdisser, H.T., Amerstorfer, U.V., Amerstorfer, T., Weiss, A.J., Hinterreiter, J., Windisch, A., 2021b. Machine learning for predicting the B_z magnetic field component from upstream in situ observations of solar coronal mass ejections. Space Weather 19, e2021SW002859. https://doi.org/10.1029/2021SW002859.

Reiss, M.A., Muglach, K., Mason, E., Davies, E.E., Chakraborty, S., Delouille, V., Downs, C., Garton, T.M., Grajeda, J.A., Hamada, A., Heinemann, S.G., Hofmeister, S., Illarionov, E., Jarolim, R., Krista, L., Lowder, C., Verwichte, E., Arge, C.N., Boucheron, L.E., Foullon, C., Kirk, M.S., Kosovichev, A., Leisner, A., Möstl, C., Turtle, J., Veronig, A., 2024. A community data set for comparing automated coronal hole detection schemes. Astrophys. J. Suppl. Ser. 271 (1), 6. https://doi.org/10.3847/1538-4365/ad1408.

Reiss, M.A., Muglach, K., Mullinix, R., Kuznetsova, M.M., Wiegand, C., Temmer, M., Arge, C.N., Dasso, S., Fung, S.F., González-Avilés, J.J., Gonzi, S., Jian, L., MacNeice, P., Möstl, C., Owens, M., Perri, B., Pinto, R.F., Rastätter, L., Riley, P., Samara, E., 2023. Unifying the

validation of ambient solar wind models. Adv. Space Res. 72 (12), 5275–5286. https://doi.org/10.1016/j.asr.2022.05.026.

Reiss, M.A., Temmer, M., Veronig, A.M., Nikolić, L., Vennerstrom, S., Schoengassner, F., Hofmeister, S.J., 2016. Verification of high-speed solar wind stream forecasts using operational solar wind models. Space Weather 14 (7). https://doi.org/10.1002/2016SW001390, 14, 495–510.

Rew, R., Davis, G., Emmerson, S., Cormack, C., Caron, J., Pincus, R., Hartnett, E., Heimbigner, D., Appel, L., Fisher, W., 1989. Unidata NetCDF [Computer software]. UCAR/NCAR - Unidata. doi: 10.5065/D6H70CW6.

Richardson, I.G., Cane, H.V., 2011. Geoeffectiveness (*Dst* and *Kp*) of interplanetary coronal mass ejections during 1995–2009 and implications for storm forecasting. Space Weather 9, S07005. https://doi.org/10.1029/2011SW000670.

Richardson, I.G., Mays, M.L., Thompson, B.J., 2018. Prediction of solar energetic particle event peak proton intensity using a simple algorithm based on CME speed and direction and observations of associated solar phenomena. Space Weather 16, 1862–1881. https://doi.org/10.1029/2018SW002032.

Ridley, A.J., Deng, Y., Tóth, G., 2006. The global ionosphere–thermosphere model. J. Atmos. Sol. Terr. Phys. 68 (8), 839–864. https://doi.org/10.1016/j.jastp.2006.01.008.

Ridley, A.J., De Zeeuw, D.L., Rastätter, L., 2016. Rating global magnetosphere model simulations through statistical data-model comparisons. Space Weather 14, 819–834. https://doi.org/10.1002/2016SW001465.

Riley, P., Ben-Nun, M., 2021. On the sources and sizes of uncertainty in predicting the arrival time of interplanetary coronal mass ejections using global MHD models. Space Weather 19 (6), e02775. https://doi.org/10.1029/2021SW002775.

Riley, P., Gosling, J.T., Pizzo, V.J., 1997. A two-dimensional simulation of the radial and latitudinal evolution of a solar wind disturbance driven by a fast, high-pressure coronal mass ejection. J. Geophys. Res. 102 (A7), 14677–14686. https://doi.org/10.1029/97JA01131.

Riley, P., Ben-Nun, M., Linker, J.A., Owens, M.J., Horbury, T.S., 2017. Forecasting the properties of the solar wind using simple pattern recognition. Space Weather 15, 526–540. https://doi.org/10.1002/2016SW001589.

Riley, P., Mays, L., Andries, J., Amerstorfer, T., Biesecker, D., Delouille, V., Dumbović, M., Feng, X., Henley, E., Linker, J.A., Möstl, C., Nuñez, M., Pizzo, V., Temmer, M., Tobiska, W.K., Verbeke, C., West, M.J., Zhao, X., 2018. Forecasting the arrival time of coronal mass ejections: analysis of the CCMC CME scoreboard. Space Weather 16, 1245–1260. https://doi.org/10.1029/2018SW001962.

Riley, P., Reiss, M.A., Möstl, C., 2023. Which upstream solar wind conditions matter most in predicting B_z within coronal mass ejections. Space Weather 21, e2022SW003327. https://doi.org/10.1029/2022SW003327.

Roberts, D.A., Thieman, J., Génot, V., King, T., Gangloff, M., Perry, C., Wiegand, C., De Zeeuw, D., Fung, S.F., Cecconi, B., Hess, S., 2018. The SPASE data model: a metadata standard for registering, finding, accessing, and using Heliophysics data obtained from observations and modeling. Space Weather 16, 1899–1911. https://doi.org/10.1029/2018SW002038.

Robinson, R., Zhang, Y., Garcia-Sage, K., Fang, X., Verkhoglyadova, O. P., Ngwira, C., Bingham, S., Kosar, B., Zheng, Y., Kaeppler, S., Liemohn, M., Weygand, J.M., Crowley, G., Merkin, V., McGranaghan, R., Mannucci, A.J., 2019. Space weather modeling capabilities assessment: auroral precipitation and high-latitude ionospheric electrodynamics. Space Weather 17, 212–215. https://doi.org/10.1029/2018SW002127.

Rodriguez, J.V., Kress, B.T., Buzulukova, N.Y., Redmon, R.J., 2025. Solar wind and magnetospheric conditions for satellite anomalies attributed to shallow internal charging. Space Weather 23, e2024SW004112. https://doi.org/10.1029/2024SW004112.

Rodriguez, L., Scolini, C., Mierla, M., Zhukov, A.N., West, M.J., 2020. Space weather monitor at the L5 point: a case study of a CME observed with STEREO B. Space Weather 18, e2020SW002533. https://doi.org/10.1029/2020SW002533.

Rodwell, M.J., Richardson, D.S., Parsons, D.B., Wernli, H., 2018. Flow-dependent reliability: a path to more skillful ensemble forecasts. Bull. Amer. Meteor. Soc. 99, 1015–1026. https://doi.org/10.1175/BAMS-D-17-0027.1.

Rodwell, M.J., Palmer, T.N., 2007. Using NWP to assess climate models. ECMWF Tech. Memoranda 518. https://doi.org/10.21957/74cdjc70y, DOI: 10.1051/0004-6361:200811101.

Roudier, T., Rieutord, M., Brito, D., Rincon, F., Malherbe, J.M., Meunier, N., Berger, T., Frank, Z., 2009. Mesoscale dynamics on the Sun's surface from HINODE observations. Astronomy and Astrophysics 495 (3), 945–952. https://doi.org/10.1051/0004-6361:200811101.

Runov, A., Angelopoulos, V., Khurana, K., Liu, J., Balikhin, M., Artemyev, A.V., 2023. Properties of quiet magnetotail plasma sheet at lunar distances. J. Geophys. Res. Space Phys. 128 (11), e2023JA031908. https://doi.org/10.1029/2023JA031908.

Ryan, D.F., Dominique, M., Seaton, D., Stegen, K., White, A., 2016. Effects of flare definitions on the statistics of derived flare distributions. A&A 592, A133. https://doi.org/10.1051/0004-6361/201628130.

Samara, E., Laperre, B., Kieokaew, R., Temmer, M., Verbeke, C., Rodriguez, L., Magdalenić, J., Poedts, S., 2022. Dynamic time warping as a means of assessing solar wind time series. Astrophys. J. 927 (2), 187. https://doi.org/10.3847/1538-4357/ac4af6.

Samara, E., Provornikova, E., Arge, C.N., McCubbin, A.J., Merkin, V. G., 2025. Why do solar wind models get it wrong? Understanding the capabilities of time-dependent solar wind simulations. Astrophys. J. 994 (2), 236. https://doi.org/10.3847/1538-4357/ae0b67.

Sauer, H.H., Wilkinson, D.C., 2008. Global mapping of ionospheric HF/VHF radio wave absorption due to solar energetic protons. Space Weather 6, S12002. https://doi.org/10.1029/2008SW000399.

Saunders, R., Hocking, J., Turner, E., Rayer, P., Rundle, D., Brunel, P., Vidot, J., Roquet, P., Matricardi, M., Geer, A., Bormann, N., Lupu, C., 2018. An update on the RTTOV fast radiative transfer model (currently at version 12). Geosci. Model Dev. 11, 2717–2737. https://doi.org/10.5194/gmd-11-2717-2018.

Savani, N.P., Vourlidas, A., Szabo, A., Mays, M.L., Richardson, I.G., Thompson, B.J., Pulkkinen, A., Evans, R., Nieves-Chinchilla, T., 2015. Predicting the magnetic vectors within coronal mass ejections arriving at Earth: 1. Initial Architecture. Space Weather 13, 374–385. https://doi.org/10.1002/2015SW001171.

Savani, N.P., Vourlidas, A., Richardson, I.G., Szabo, A., Thompson, B.J., Pulkkinen, A., Mays, M.L., Nieves-Chinchilla, T., Bothmer, V., 2017. Predicting the magnetic vectors within coronal mass ejections arriving at Earth: 2. Geomagn. Res., Space Weather 15, 441–461. https://doi.org/10.1002/2016SW001458.

Sawyer, C., Warwick, J.W., Dennett, J.T., 1986. Solar Flare Prediction. Colorado Assoc. Univ. Press, Boulder, CO, p. 1986.

Scherliess, L., Tsagouri, I., Yizengaw, E., Bruinsma, S., Shim, J.S., Coster, A., Retterer, J.M., 2019. The international community coordinated modeling center space weather modeling capabilities assessment: overview of ionosphere/thermosphere activities. Space Weather 17, 527–538. https://doi.org/10.1029/2018SW002036.

Schou, J., Borrero, J.M., Norton, A.A., Tomczyk, S., Elmore, D., Card, G.L., 2012. Polarization calibration of the helioseismic and magnetic imager (HMI) onboard the solar dynamics observatory (SDO). Sol. Phys. 275 (1–2), 327–355. https://doi.org/10.1007/s11207-010-9639-8.

Schreiner, W.S., Sokolovskiy, S.V., Rocken, C., 1999. Analysis and validation of GPS/MET radio occultation data in the ionosphere. Radio Sci. 34, 949–966. https://doi.org/10.1029/1999RS900034.

Schreiner, W.S., Weiss, J.P., Anthes, R.A., Braun, J., Chu, V., Fong, J., Hunt, D., Kuo, Y., Meehan, T., Serafino, W., Sjoberg, J., Sokolovskiy, S., Talaat, E., Wee, T.K., Zeng, Z., 2020. COSMIC-2 radio occultation constellation: first results. Geophys. Res. Lett. 47, e2019GL086841. https://doi.org/10.1029/2019GL086841.

Schreiner, W., Rocken, C., Sokolovskiy, S., Syndergaard, S., Hunt, D., 2007. Estimates of the precision of GPS radio occultations from the COSMIC/FORMOSAT-3 mission. Geophys. Res. Lett. 34, L04808. https://doi.org/10.1029/2006GL027557.

Schrijver, C.J., De Rosa, M.L., 2003. Photospheric and heliospheric magnetic fields. Sol. Phys. 212 (1), 165–200. https://doi.org/10.1023/A:1022908504100.

Schrijver, C.J., Kauristie, K., Aylward, A.D., Denardini, C.M., Gibson, S. E., Glover, A., Gopalswamy, N., Grande, M., Hapgood, M., Heynderickx, D., Jakowski, N., Kalegaev, V.V., Lapenta, G., Linker, J.A., Liu, S., Mandrini, C.H., Mann, I.R., Nagatsuma, T., Nandy, D., Obara, T., O'Brien, T.P., Onsager, T., Opgenoorth, H.J., Terkildsen, M., Valladares, C.E., Vilmer, N., 2015. Understanding space weather to shield society: a global road map for 2015–2025 commissioned by COSPAR and ILWS. Adv. Space Res. 55 (12), 2745–2807. https://doi.org/10.1016/j.asr.2015.03.023.

Scolini, C., Rodriguez, L., Mierla, M., Pomoell, J., Poedts, S., 2019. Observation-based modelling of magnetised coronal mass ejections with EUHFORIA. Astron. Astrophys. 626, A122. https://doi.org/10.1051/0004-6361/201935053.

Sharpe, M.A., Murray, S.A., 2017. Verification of space weather forecasts issued by the met office space weather operations centre. Space Weather 15, 1383–1395. https://doi.org/10.1002/2017SW001683.

Sharpe, M.A., 2013. Verification of marine forecasts using an objective area forecast verification system. Met. Apps 20, 224–235. https://doi.org/10.1002/met.1341.

Shaifullah, G.M., Magdalenic, J., Tiburzi, C., Jebaraj, I., Samara, E., Zucca, P., 2023. Validation of heliospheric modeling algorithms through pulsar observations II: simulations with EUHFORIA. Adv. Space Res. 72 (12), 5298–5310. https://doi.org/10.1016/j.asr.2022.07.003.

Sheeley Jr., N.R., DeVore, C.R., Boris, J.P., 1985. Simulations of the mean solar magnetic field during sunspot cycle-21. Sol. Phys. 98 (2), 219–239. https://doi.org/10.1007/BF00152457.

Shen, F., Shen, C., Xu, M., Liu, Y., Feng, X., Wang, Y., 2022. Propagation characteristics of coronal mass ejections (CMEs) in the corona and interplanetary space. Rev. Mod. Plasma Phys. 6 (1), 8. https://doi.org/10.1007/s41614-022-00069-1.

Shim, J.S., Kuznetsova, M., Rastätter, L., Bilitza, D., Butala, M., Codrescu, M., Emery, B.A., Foster, B., Fuller-Rowell, T., Huba, J., Mannucci, A.J., Pi, X., Ridley, A., Scherliess, L., Schunk, R.W., Sojka, J.J., Stephens, P., Thompson, D.C., Weimer, D., Zhu, L., Sutton, E., 2012. CEDAR Electrodynamics Thermosphere Ionosphere (ETI) Challenge for systematic assessment of ionosphere/thermosphere models: electron density, neutral density, NmF2, and hmF2 using space-based observations. Space Weather 10 (10), S10004. https://doi.org/10.1029/2012SW000851.

Shim, J.S., Kuznetsova, M., Rastätter, L., Bilitza, D., Butala, M., Codrescu, M., Emery, B.A., Foster, B., Fuller-Rowell, T., Huba, J., Mannucci, A.J., Pi, X., Ridley, A., Scherliess, L., Schunk, R.W., Sojka, J.J., Stephens, P., Thompson, D.C., Weimer, D., Zhu, L., Anderson, D., Chau, J.L., Sutton, E., 2014. Systematic evaluation of ionosphere/thermosphere (I/T) models: CEDAR electrodynamics thermosphere ionosphere (ETI) challenge (2009–2010). In: Modeling the ionosphere-thermosphere system. AGU Geophysical Monograph Series. doi: 10.1002/9781118704417.ch13.

Shim, J.S., Kuznetsova, M., Rastätter, L., Hesse, M., Bilitza, D., Butala, M., Codrescu, M., Emery, B., Foster, B., Fuller-Rowell, T., Huba, J., Mannucci, A.J., Pi, X., Ridley, A., Scherliess, L., Schunk, R.W., Stephens, P., Thompson, D.C., Zhu, L., Anderson, D., Chau, J.L., Sojka, J.J., Rideout, B., et al., 2011. CEDAR electrodynamics thermosphere ionosphere (ETI) challenge for systematic assessment of ionosphere/thermosphere models: NmF2, hmF2, and vertical drift using ground-based observations. Space Weather 9 (12). https://doi.org/10.1029/2011SW000727.

Shim, J.S., Rastätter, L., Kuznetsova, M., Bilitza, D., Codrescu, M., Coster, A.J., Emery, B.A., Fedrizzi, M., Förster, M., Fuller-Rowell, T., Gardner, L.C., Goncharenko, L., Huba, J., McDonald, S.E., Mannucci, A.J., Namgaladze, A.A., Pi, X., Prokhorov, B.E., Ridley, A.J., Scherliess, L., Schunk, R.W., Sojka, J.J., Zhu, L., 2017. CEDAR-GEM challenge for systematic assessment of ionosphere/thermosphere models in predicting TEC during the 2006 December storm event. Space Weather 15 (10), 1238–1256. https://doi.org/10.1002/2017SW001649.

Shim, J.S., Tsagouri, I., Goncharenko, L., Rastätter, L., Kuznetsova, M., Bilitza, D., Codrescu, M., Coster, A.J., Solomon, S.C., Fedrizzi, M., Förster, M., Fuller-Rowell, T., Gardner, L.C., Huba, J., Namgaladze, A.A., Prokhorov, B.E., Ridley, A.J., Scherliess, L., Schunk, R.W., Sojka, J.J., Zhu, L., 2018. Validation of ionospheric specifications during geomagnetic storms: TEC and foF2 during the 2013 March storm event. Space Weather 16 (11), 1686–1701. https://doi.org/10.1029/2018SW002034.

Shim, J.S., Song, I.-S., Jee, G., Kwak, Y.-S., Tsagouri, I., Goncharenko, L., McInerney, J., Vitt, A., Rastätter, L., Yue, J., Chou, M., Codrescu, M., Coster, A.J., Fedrizzi, M., Fuller-Rowell, T., Ridley, A.J., Solomon, S.C., Habarulema, J.B., 2023. Validation of ionospheric specifications during geomagnetic storms: TEC and foF2 during the 2013 March storm event-II. Space Weather 21, e2022SW003388. https://doi.org/10.1029/2022SW003388.

Shiota, D., Kataoka, R., 2016. Magnetohydrodynamic simulation of interplanetary propagation of multiple coronal mass ejections with internal magnetic flux rope (SUSANOO-CME). Space Weather 14 (2), 56–75. https://doi.org/10.1002/2015SW001308.

Siemes, C., Encarnação, J., Doornbos, E., van den IJssel, J., Kraus, J., Pereštý, R., Grunwaldt, L., Apelbaum, G., Flury, J., Holmdahl Olsen, P.E., 2016. Swarm accelerometer data processing from raw accelerations to thermospheric neutral densities. Earth Planets Space 68, 9. https://doi.org/10.1186/s40623-016-0474-5.

Simmons, A.J., Hollingsworth, A., 2002. Some aspects of the improvement in skill of numerical weather prediction. Q.J.R. Meteorol. Soc. 128, 647–677. https://doi.org/10.1256/003590002321042135.

Simms, L.E., Ganushkina, N.Y., Van der Kamp, M., Balikhin, M., Liemohn, M.W., 2023. Predicting geostationary 40–150 keV electron flux using ARMAX (an autoregressive moving average transfer function), RNN (a recurrent neural network), and logistic regression: a comparison of models. Space Weather 21, e2022SW003263. https://doi.org/10.1029/2022SW003263.

Singh, T., Yalim, M.S., Pogorelov, N.V., 2018. A data constrained model for coronal mass ejections using the graduated cylinder shell method. Astrophys. J. 864, 18. https://doi.org/10.3847/1538-4357/aad3b4.

Singh, T., Yalim, M.S., Pogorelov, N.V., Gopalswamy, N., 2019. Simulating solar coronal mass ejections constrained by observations of their speed and poloidal flux. Astrophys. J. Lett. 875, L17. https://doi.org/10.3847/2041-8213/ab14e9.

Singh, T., Yalim, M.S., Pogorelov, N.V., Gopalswamy, N., 2020. A modified spheromak model suitable for coronal mass ejection simulations. Astrophys. J. 894, 49. https://doi.org/10.3847/1538-4357/ab845f.

Singh, T., Benson, B., Raza, S.A.Z., Kim, T.K., Pogorelov, N.V., Smith, W.P., Arge, C.N., 2023. Improving the arrival time estimates of coronal mass ejections by using magnetohydrodynamic ensemble modeling, heliospheric imager data, and machine learning. Astrophys. J. 948 (2), 78. https://doi.org/10.3847/1538-4357/acc10a.

Singh, T., Kim, T.K., Pogorelov, N.V., Arge, C.N., 2022. Ensemble simulations of the 2012 July 12 coronal mass ejection with the constant-turn flux rope model. Astrophys. J. 933 (2), 123. https://doi.org/10.3847/1538-4357/ac73f3.

Smith, A.W., Forsyth, C., Rae, I.J., Garton, T.M., Jackman, C.M., Bakrania, M., Shore, R.M., Richardson, G.S., Beggan, C.D., Heyns, M.J., Eastwood, J.P., Thomson, A.W.P., Johnson, J.M., 2022. On the considerations of using near real time data for space weather hazard forecasting. Space Weather 20, e2022SW003098. https://doi.org/10.1029/2022SW003098.

Sorathia, K.A., Merkin, V.G., Panov, E.V., Zhang, B., Lyon, J.G., Garretson, J., Ukhorskiy, A.Y., Ohtani, S., Sitnov, M., Wiltberger, M., 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 (14), e2020GL088227. https://doi.org/10.1029/2020GL088227.

Stankov, S.M., Jodogne, J.C., Kutiev, I., Stegen, K., Warnant, R., 2012. Evaluation of automatic ionogram scaling for use in real-time ionospheric density profile specification: dourbes DGS-256/ARTIST-4 performance. Ann. Geophys. 55 (2), 283–291. https://doi.org/10.4401/ag-4976.

Stankov, S.M., Verhulst, T.G.W., Sapundjiev, D., 2023. Automatic ionospheric weather monitoring with DPS-4D ionosonde and ARTIST-5 autoscaler: system performance at a mid-latitude observatory. Radio Sci. 58, e2022RS007628. https://doi.org/10.1029/2022RS007628.

Stone, E.C., Frandsen, A.M., Mewaldt, R.A., Christian, E.R., Margolies, D., Ormes, J.F., Snow, F., 1998. The advanced composition explorer. Space Sci. Rev. 86, 1–22. https://doi.org/10.1023/A:1005082526237.

Storz, M.F., Bowman, B.R., Branson, M.J.I., Casali, S.J., Tobiska, W.K., 2005. High accuracy satellite drag model (HASDM). Adv. Space Res. 36 (12), 2497–2505. https://doi.org/10.1016/j.asr.2004.02.020.

Subbotin, D.A., Shprits, Y.Y., 2009. Three-dimensional modeling of the radiation belts using the Versatile Electron Radiation Belt (VERB) code. Space Weather 7, S10001. https://doi.org/10.1029/2008SW000452.

Sutton, E.K., 2018. A new method of physics-based data assimilation for the quiet and disturbed thermosphere. Space Weather 16, 736–753.

Syndergaard, S., 2002. A new algorithm for retrieving GPS radio occultation total electron content. Geophys. Res. Lett. 29 (16), 2002. https://doi.org/10.1029/2001GL014478.

Taha, A.A., Hanbury, A., 2015. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med. Imaging 15 (1), 1–28. https://doi.org/10.1186/s12880-015-0068-x.

Temmer, M., Scolini, C., Richardson, I.G., Heinemann, S.G., Paouris, E., Vourlidas, A., Bisi, M.M., writing teams:, Al-Haddad, N., Amerstorfer, T., Barnard, L., Buresova, D., Hofmeister, S.J., Iwai, K., Jackson, B.V., Jarolim, R., Jian, L.K., Linker, J.A., Lugaz, N., Manoharan, P. K., Mays, M.L., Mishra, W., Owens, M.J., Palmerio, E., Perri, B., Pomoell, J., Pinto, R.F., Samara, E., Singh, T., Sur, D., Verbeke, C., Veronig, A.M., Zhuang, B., 2026. CME propagation through the heliosphere: Status and future of observations and model development. Adv. Space Res., this issue. doi: 10.1016/j.asr.2023.07.003.

Thernisien, A., 2011. Implementation of the graduated cylindrical shell model for the three-dimensional reconstruction of coronal mass ejections. Astrophys. J. Suppl. Ser. 194 (2), 33. https://doi.org/10.1088/0067-0049/194/2/33.

Thompson, R.L., Watt, C.E.J., Williams, P.D., 2020. Accounting for variability in ULF wave radial diffusion models. J. Geophys. Res. Space Phys. 125, e2019JA027254. https://doi.org/10.1029/2019JA027254.

Thomson, A.W.P., 1996. Non-linear predictions of *Ap* by activity class and numerical value. PAGEOPH 146, 163–193. https://doi.org/10.1007/BF00876675.

Thomson, A.W.P., 2000. Evaluating space weather forecasts of geomagnetic activity from a user perspective. Geophys. Res. Lett. 27 (24), 4049–4052. https://doi.org/10.1029/2000GL011908.

Thomson, A.W.P., King, J.A., Clarke, E., and Clark, T.D.G., 2001. Improved Predictions of Solar and Geomagnetic Activity with Application to ESA/LEO Satellite Operations. Paper presented at 'Space Weather Workshop: Looking Towards a Future European Space Weather Programme', ESTEC, https://geomag.bgs.ac.uk/documents/estec171201a.pdf.

Tiburzi, C., Jackson, B.V., Cota, L., Shaifullah, G.M., Fallows, R.A., Tokumaru, M., Zucca, P., 2023. Validation of heliospheric modeling algorithms through pulsar observations I: interplanetary scintillation-based tomography. Adv. Space Res. 72 (12), 5287–5297. https://doi.org/10.1016/j.asr.2022.04.070.

Tobiska, W.K., Didkovsky, L., Judge, K., Weiman, S., Bouwer, D., Bailey, J., Atwell, B., Maskrey, M., Mertens, C., Zheng, Y., Shea, M., Smart, D., Gersey, B., Wilkins, R., Bell, D., Gardner, L., Fuschino, R., 2018. Analytical representations for characterizing the global aviation radiation environment based on model and measurement databases. Space Weather 16, 1523–1538. https://doi.org/10.1029/2018SW001843.

Tóth, G., van der Holst, B., Sokolov, I.V., De Zeeuw, D.L., Gombosi, T. I., Fang, F., Manchester, W.B., Meng, X., Najib, D., Powell, K.G., Stout, Q.F., Glocer, A., Ma, Y., Opher, M., 2012. Adaptive numerical algorithms in space weather modeling. J. Comput. Phys. 231 (3), 870–903. https://doi.org/10.1016/j.jcp.2011.02.006.

Tóth, G., Sokolov, I.V., Gombosi, T.I., Chesney, D.R., Clauer, C.R., de Zeeuw, D.L., Hansen, K.C., Kane, K.J., Manchester, W.B., Oehmke, R.C., Powell, K.G., Ridley, A.J., Roussev, I.I., Stout, Q.F., Volberg, O., Wolf, R.A., Sazykin, S., Chan, A., Yu, B., Kóta, J., 2005. Space Weather Modeling Framework: a new tool for the space science community. J. Geophys. Res. Space Phys. 110 (A12), A12226. https://doi.org/10.1029/2005JA011126.

Tsagouri, I., Anna Belehaki, David R. Themens, Norbert Jakowski, Tim Fuller-Rowell, Mainul M. Hoque, Grzegorz Nykiel, Wojciech J. Miloch, Claudia Borries, Anna Morozova, Teresa Barata, William Engelke, Ja-Soon Shim, 2026a. Ionosphere variability I: Advances in observational, monitoring and detection capabilities, Advances in Space Research, 2026, this issue. doi: 10.1016/j.asr.2023.07.024.

Tsagouri, I., David R. Themens, Anna Belehaki, Ja-Soon Shim, Mainul M. Hoque, Grzegorz Nykiel, Claudia Borries, Anna Morozova, Teresa Barata, Wojciech J. Miloch, 2026b. Ionosphere variability II: Advances in theory and modeling, Advances in Space Research, 2026, this issue. doi: 10.1016/j.asr.2023.07.056.

Tsagouri, I., Belehaki, A., 2023. Assessment of solar wind driven ionospheric storm forecasts: the case of the solar wind driven autoregression model for ionospheric forecast (SWIF). Adv. Space Res. 72 (12), 5577–5586. https://doi.org/10.1016/j.asr.2022.06.047.

Tsagouri, I., Goncharenko, L., Shim, J.S., Belehaki, A., Buresova, D., Kuznetsova, M.M., 2018. Assessment of current capabilities in modeling the ionospheric climatology for space weather applications: FoF2 and hmF2. Space Weather 16 (12), 1930–1945. https://doi.org/10.1029/2018SW002035.

Tsurutani, B.T., Gonzalez, W.D., Gonzalez, A.L.C., Guarnieri, F.L., Gopalswamy, N., Grande, M., Kamide, Y., Kasahara, Y., Lu, G., Mann, I., McPherron, R., Soraas, F., Vasyliunas, V., 2006. Corotating solar wind streams and recurrent geomagnetic activity: a review. J. Geophys. Res. Space Phys. 111, A07S01. https://doi.org/10.1029/2005JA011273.

Tu, W., Li, W., Albert, J.M., Morley, S.K., 2019. Quantitative assessment of radiation belt modeling. J. Geophys. Res. Space Phys. 124 (2), 898–904. https://doi.org/10.1029/2018JA026414.

Turner, H., Lang, M., Owens, M., Smith, A., Riley, P., Marsh, M., Gonzi, S., 2023. Solar wind data assimilation in an operational context: use of near-real-time data and the forecast value of an L5 monitor. Space Weather 21, e2023SW003457. https://doi.org/10.1029/2023SW003457.

Upton, L., Hathaway, D.H., 2014. Predicting the Sun's polar magnetic fields with a surface flux transport model. Astrophys. J. 780 (1), 5. https://doi.org/10.1088/0004-637X/780/1/5.

van den IJssel, J., Visser, P., 2007. Performance of GPS-based accelerometry: CHAMP and GRACE. Adv. Space Res. 39 (10), 1597–1603. https://doi.org/10.1016/j.asr.2006.12.027.

van der Holst, B., Sokolov, I.V., Meng, X., Jin, M., Manchester IV, W.B., Tóth, G., Gombosi, T.I., 2014. Alfvén wave solar model (AWSoM): coronal heating. Astrophys. J. 782 (2), 81. https://doi.org/10.1088/0004-637X/782/2/81.

Verbeke, C., Mays, M.L., Kay, C., Riley, P., Palmerio, E., Dumbović, M., Mierla, M., Scolini, C., Temmer, M., Paouris, E., Balmaceda, L.A., Cremades, H., Hinterreiter, J., 2023. Quantifying errors in 3D CME parameters derived from synthetic data using white-light reconstruction techniques. Adv. Space Res. 72 (12), 5243–5262. https://doi.org/10.1016/j.asr.2022.08.056.

Verbeke, C., Mays, M.L., Temmer, M., Bingham, S., Steenburgh, R., Dumbović, M., Núñez, M., Jian, L.K., Hess, P., Wiegand, C., Taktakishvili, A., Andries, J., 2019. Benchmarking CME arrival time

and impact: progress on metadata, metrics, and events. Space Weather 17 (1), 6–26. https://doi.org/10.1029/2018SW002046.

Verbeke, C., Schmieder, B., Démoulin, P., Dasso, S., Grison, B., Samara, E., Scolini, C., Poedts, S., 2022. Over-expansion of coronal mass ejections modelled using 3D MHD EUHFORIA simulations. Adv. Space Res. 70 (6), 1663–1683. https://doi.org/10.1016/j.asr.2022.06.013.

Vigh, J., 2023. Going Beyond the Terrestrial: Space Weather Verification Using METplus, Unifying Innovations in Forecasting Capabilities Workshop. https://epic.noaa.gov/eventsposts/uifcw-2023/, Slides: https://epic.noaa.gov/wp-content/uploads/2023/08/UIFCW-2023-Thu-4.-Vigh_20230719_SWPC-RT_UIFCW_presentation_v2.pptx; recording: https://vimeo.com/855853449?share=copy#t=2880.216.

Vigh, J., Jensen, T.L., Onsager, T., Maruyama, N., Steenburgh, R., Fuller-Rowell, D., Wang, J., Codrescu, M., Fuller-Rowell, T., 2021. Developing a space weather verification system using METplus. Am. Meteorol. Soc. 101, recording: https://ams.confex.com/ams/101ANNUAL/prelim.cgi/Paper/384237.

Vigh, J., Onsager, T., Jensen, T. L., Steenburgh, R., Maruyama, N., Centinello, F., Fuller-Rowell, D., Codrescu, M., Wang, J., Fuller-Rowell, T., Brown, B.G., Gotway, J.E.H., Burek, T.V., McCabe, G., 2020. Developing a space weather verification system using METplus. International Workshop on Verification Methods 2020. https://open-sky.ucar.edu/islandora/object/conference%3A3520.

Viljanen, A., Nevanlinna, H., Pajunpää, K., Pulkkinen, A., 2001. Time derivative of the horizontal geomagnetic field as an activity indicator. Ann. Geophys. 19, 1107–1118. https://doi.org/10.5194/angeo-19-1107-2001.

Vögler, A., Shelyag, S., Schüssler, M., Cattaneo, F., Emonet, T., Linde, T., 2005. Simulations of magneto-convection in the solar photosphere. Equations, methods, and results of the MURaM code. Astron. Astrophys. 429, 335–351. https://doi.org/10.1051/0004-6361:20041507.

Vourlidas, A., Patsourakos, S., Savani, N.P., 2019. Predicting the geoeffective properties of coronal mass ejections: current status, open issues and path forward. Philos. Trans. A Math Phys. Eng. Sci. 377 (2148). https://doi.org/10.1098/rsta.2018.0096.

Vršnak, B., Zic, T., Vrbanec, D., Temmer, M., Rollett, T., Möstl, C., Veronig, A., Čalogović, J., Dumbović, M., Lulic, S., Moon, Y.-J., Shanmugaraju, A., 2013. Propagation of interplanetary coronal mass ejections: the drag-based model. Sol. Phys. 285 (1–2), 295–315. https://doi.org/10.1007/s11207-012-0035-4.

Waldron, Z.C., Garcia-Sage, K., Thayer, J.P., Sutton, E.K., Ray, V., Rowlands, D.D., Lemoine, F.G., Luthcke, S.B., Kuznetsova, M., Ringuette, R., Rastätter, L., Berland, G.D., 2024. Assessing thermospheric neutral density models using GEODYN's precision orbit determination. SpaceWeather 22, e2023SW003603. https://doi.org/10.1029/2023SW003603.

Walsh, B.M., Bhakyapaibul, T., Zou, Y., 2019. Quantifying the uncertainty of using solar wind measurements for geospace inputs. J. Geophys. Res. Space Phys. 124, 3291–3302. https://doi.org/10.1029/2019JA026507.

Wang, D., Shprits, Y.Y., Zhelavskaya, I.S., Effenberger, F., Castillo, A., Drozdov, A.Y., Aseev, N.A., Cervantes, S., 2020. The effect of plasma boundaries on the dynamic evolution of relativistic radiation belt electrons. J. Geophys. Res. Space Phys. 125, e2019JA027422. https://doi.org/10.1029/2019JA027422.

Wang, Y.-M., Nash, A.G., Sheeley Jr., N.R., 1989. Evolution of the Sun's polar fields during sunspot cycle 21: poleward surges and long-term behavior. Astrophys. J. 347, 529–541. https://doi.org/10.1086/168143.

Wang, Y., Shen, C.L., Wang, S., Ye, P.Z., 2003. An empirical formula relating the geomagnetic storm's intensity to the interplanetary parameters: −VBz and Δt. Geophys. Res. Lett. 30, 2039. https://doi.org/10.1029/2003GL017901.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612. https://doi.org/10.1109/TIP.2003.819861.

Waszewski, A., Morgan, J.S., Chhetri, R., Ekers, R., Cheung, M.C.M., Bhat, N.D.R., Johnston-Hollitt, M., 2023. Resolving moving heliospheric structures using interplanetary scintillation observations with the Murchison Widefield Array. Space Weather 21, e2023SW003570. https://doi.org/10.1029/2023SW003570.

Watt, C.E.J., Allison, H.J., Thompson, R.L., Bentley, S.N., Meredith, N. P., Glauert, S.A., Horne, R.B., Rae, I.J., 2021. The implications of temporal variability in wave-particle interactions in Earth's radiation belts. Geophys. Res. Lett. 48, e2020GL089962. https://doi.org/10.1029/2020GL089962.

Weigel, R.S., Vandegriff, J., Faden, J., King, T., Roberts, D.A., Harris, B., Candey, R., Lal, N., Boardsen, S., Lindholm, C., Lindholm, D., Baltzer, T., Brown, L.E., Grimes, E.W., Cecconi, B., Génot, V., Renard, B., Masson, A., Martinez, B., 2021. HAPI: an API standard for accessing Heliophysics time series data. J. Geophys. Res. Space Phys. 126, e2021JA029534. https://doi.org/10.1029/2021JA029534.

Weiler, E., Möstl, C., Davies, E.E., Veronig, A.M., Amerstorfer, U.V., Amerstorfer, T., Le Louëdec, J., Bauer, M., Lugaz, N., Haberle, V., Rüdisser, H.T., Majumdar, S., Reiss, M., 2025. First observations of a geomagnetic superstorm with a sub-L1 monitor. Space Weather 23, e2024SW004260. https://doi.org/10.1029/2024SW004260.

Welling, D.T., Ngwira, C.M., Opgenoorth, H., Haiducek, J.D., Savani, N. P., Morley, S.K., Cid, C., Weigel, R.S., Weygand, J.M., Woodroffe, J. R., Singer, H.J., Rosenqvist, L., Liemohn, M.W., 2018. Recommendations for next-generation ground magnetic perturbation validation. Space Weather 16, 1912–1920. https://doi.org/10.1029/2018SW002064.

Wharton, S.J., Millward, G.H., Bingham, S., Henley, E.M., Gonzi, S., Jackson, D.R., 2019. Incorporation of Heliospheric Imagery into the CME Analysis Tool for improvement of CME forecasting. Space Weather 17, 1312–1328. https://doi.org/10.1029/2019SW002166.

White, W.W., Schoendorf, J.A., Siebert, K.D., Maynard, N.C., Weimer, D.R., Wilson, G.L., Sonnerup, B.U.O., Siscoe, G.L., Erickson, G.M., 2001. MHD simulation of magnetospheric transport at the mesoscale. In: Song, P., Singer, H., Siscoe, G., (Eds.), Space Weather. AGU, Washington, D. C., 2001.

Whitman, K., Egeland, R., Richardson, I.G., Allison, C., Quinn, P., Barzilla, J., Kitiashvili, I., Sadykov, V., Bain, H.M., Dierckxsens, M., Mays, M.L., Tadesse, T., Lee, K.T., Semones, E., Luhmann, J.G., Núñez, M., White, S.M., Kahler, S.W., Ling, A.G., Smart, D.F., Shea, M.A., Tenishev, V., Boubrahimi, S.F., Aydin, B., Martens, P., Angryk, R., Marsh, M.S., Dalla, S., Crosby, N., Schwadron, N.A., Kozarev, K., Gorby, M., Young, M.A., Laurenza, M., Cliver, E.W., Alberti, T., Stumpo, M., Benella, S., Papaioannou, A., Anastasiadis, A., Sandberg, I., Georgoulis, M.K., Ji, A., Kempton, D., Pandey, C., Li, G., Hu, J., Zank, G.P., Lavasa, E., Giannopoulos, G., Falconer, D., Kadadi, Y., Fernandes, I., Dayeh, M.A., Muñoz-Jaramillo, A., Chatterjee, S., Moreland, K.D., Sokolov, I.V., Roussev, I.I., Taktakishvili, A., Effenberger, F., Gombosi, T., Huang, Z., Zhao, L., Wijsen, N., Aran, A., Poedts, S., Kouloumvakos, A., Paassilta, M., Vainio, R., Belov, A., Eroshenko, E.A., Abunina, M.A., Abunin, A.A., Balch, C. C., Malandraki, O., Karavolos, M., Heber, B., Labrenz, J., Kühl, P., Kosovichev, A.G., Oria, V., Nita, G.M., Illarionov, E., O'Keefe, P.M., Jiang, Y., Fereira, S.H., Ali, A., Paouris, E., Aminalragia-Giamini, S., Jiggens, P., Jin, M., Lee, C.O., Palmerio, E., Bruno, A., Kasapis, S., Wang, X., Chen, Y., Sanahuja, B., Lario, D., Jacobs, C., Strauss, D.T., Steyn, R., van den Berg, J., Swalwell, B., Waterfall, C., Nedal, M., Miteva, R., Dechev, M., Zucca, P., Engell, A., Maze, B., Farmer, H., Kerber, T., Barnett, B., Loomis, J., Grey, N., Thompson, B.J., Linker, J.A., Caplan, R.M., Downs, C., Török, T., Lionello, R., Titov, V., Zhang, M., Hosseinzadeh, P., 2023. Review of solar energetic particle prediction models. Adv. Space Res. 72 (12), 5161–5242. https://doi.org/10.1016/j.asr.2022.08.006.

Whitman, K., Egeland, R., Allison, C., Quinn, P., Stegeman, L., (2026), Validation of Solar Energetic Particle Forecasting Models for Space Radiation Operations with SPHINX and VIVID, NASA Technical Report TP-20260000463, January 1, 2026, https://doi.org/10.64631/WDZT5355, https://ntrs.nasa.gov/citations/20260000463.

Wilks, Daniel S., 2020. Statistical Methods in the Atmospheric Sciences. 4th ed., Amsterdam, Elsevier. doi: 10.1016/C2017-0-03921-6.

Wintoft, P., Wik, M., 2018. Evaluation of Kp and Dst predictions using ACE and DSCOVR solar wind data. Space Weather 16, 1972–1983. https://doi.org/10.1029/2018SW001994.

WMO, 2023. Manual on the WMO Integrated Processing and Prediction System, Annex IV to the WMO Technical Regulations. WMO-No. 485. 2023 edition. ISBN: 978-92-63-10485-4, https://library.wmo.int/idurl/4/35703.

Wold, A.M., Mays, M.L., Taktakishvili, A., Jian, L.K., Odstrčil, D., MacNeice, P., 2018. Verification of real-time WSA-ENLIL+Cone simulations of CME arrival-time at the CCMC from 2010 to 2016. J. Space Weather Space Clim. 8, A17. https://doi.org/10.1051/swsc/2018005.

Wrench, D., Parashar, T.N., Oughton, S., de Lange, K., Frean, M., 2024. What is the Reynolds number of the solar wind? Astrophys. J. 961 (2), 182. https://doi.org/10.3847/1538-4357/ad118e.

Wu, D.L., 2020. Ionospheric S4 scintillations from GNSS Radio Occultation (RO) at slant path. Rem. Sens. (Basel) 12 (15), 2373. https://doi.org/10.3390/rs12152373.

Yu, Y., Rastätter, L., Jordanova, V.K., Zheng, Y., Engel, M., Fok, M.-C., Kuznetsova, M.M., 2019. Initial results from the GEM challenge on the spacecraft surface charging environment. Space Weather 17, 299–312. https://doi.org/10.1029/2018SW002031.

Yue, X., Schreiner, W.S., Lei, J., Sokolovskiy, S.V., Rocken, C., Hunt, D. C., Kuo, Y.H., 2010. Error analysis of Abel retrieved electron density profiles from radio occultation measurements. Ann. Geophys. 28, 217–222. https://doi.org/10.5194/angeo-28-217-2010.

Zakharenkova, I., Cherniak, I., Braun, J.J., Weiss, J.-P., Wu, Q., VanHove, T., Hunt, D., Sleziak-Sallee, M., 2025. Unveiling ionospheric response to the May 2024 superstorm with low-Earth-orbit satellite observations. Space Weather 23, e2024SW004245. https://doi.org/10.1029/2024SW004245.

Zhang, B., Sorathia, K.A., Lyon, J.G., Merkin, V.G., Garretson, J.S., Wiltberger, M., 2019. GAMERA: a three-dimensional finite-volume MHD solver for non-orthogonal curvilinear geometries. Astrophys. J. Suppl. Series 244 (1), 20. https://doi.org/10.3847/1538-4365/ab3a4c.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595. https://doi.org/10.1109/CVPR.2018.00068.

Zheng, Y., Ganushkina, N.Y., Jiggens, P., Jun, I., Meier, M., Minow, J.I., O'Brien, T.P., Pitchford, D., Shprits, Y., Tobiska, W.K., Xapsos, M. A., Guild, T.B., Mazur, J.E., Kuznetsova, M.M., 2019. Space radiation and plasma effects on satellites and aviation: quantities and metrics for tracking performance of space weather environment models. Space Weather 17, 1384–1403. https://doi.org/10.1029/2018SW002042.

Zheng, Y., 2025. Multi-Purpose Model Validation Efforts for Space Plasma and Radiation Environment in the Near-Earth Region. Zenodo. April 16, 2025. doi: 10.5281/zenodo.15232712.

Zheng, Y., Jun, I., Tu, W., Shprits, Y., Kim, W., Matthiä, D., Meier, M. M., Kent Tobiska, W., Miyoshi, Y., Jordanova, V.K., Ganushkina, N. Y., Tenishev, V., O'Brien, T.P., Brunet, A., Maget, V., Guo, J., Wang, D., Horne, R.B., Glauert, S., Haas, B., Drozdov, A.Y. 2026. Overview, progress and next steps for our understanding of the near-earth space radiation and plasma environment: science and applications. Adv. Space Res., this issue. doi: 10.1016/j.asr.2024.05.017.

Zhou, Y., Feng, X., 2017. Numerical study of the propagation characteristics of coronal mass ejections in a structured ambient solar wind. J. Geophys. Res. Space Phys. 122 (2), 1451–1462. https://doi.org/10.1002/2016JA023053.