Earth and Space Science

RESEARCH ARTICLE

10.1029/2024EA004138

Guide for Conducting “Community Challenges” in Space Physics

Michael W. Liemohn[1], **Lutz Rastätter**[2], **Alexa J. Halford**[2], **Yihua Zheng**[2], **Katherine S. Garcia-Sage**[2], **Robert Redmon**[3], **and Sarah K. Vines**[4]

[1]Department of Climate and Space Sciences and Engineering, University of Michigan, Ann Arbor, MI, USA, [2]NASA Goddard Space Flight Center, Greenbelt, MD, USA, [3]NOAA Center for AI and National Centers for Environmental Information, National Oceanic and Atmospheric Administration, Boulder, CO, USA, [4]Southwest Research Institute, San Antonio, TX, USA

Key Points:

- A history is presented of the “community challenges” conducted over the past 3 decades within the Geospace Environment Modeling Program
- Key recommendations and lessons learned from past challenge leaders, as well as suggestions from the research community, are presented
- Additional resources that might aid in the successful running of a community challenge are given, including a summary of metrics options

Correspondence to:

M. W. Liemohn,
liemohn@umich.edu

Citation:

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 and Space Science*, *12*, e2024EA004138. https://doi.org/10.1029/2024EA004138

Received 10 DEC 2024
Accepted 19 APR 2025

Author Contributions:

Conceptualization: Michael W. Liemohn, Lutz Rastätter, Alexa J. Halford, Yihua Zheng, Katherine S. Garcia-Sage, Robert Redmon, Sarah K. Vines
Funding acquisition: Michael W. Liemohn
Investigation: Michael W. Liemohn, Lutz Rastätter, Alexa J. Halford, Yihua Zheng, Katherine S. Garcia-Sage, Robert Redmon, Sarah K. Vines
Methodology: Michael W. Liemohn, Lutz Rastätter, Alexa J. Halford, Yihua Zheng, Katherine S. Garcia-Sage, Robert Redmon, Sarah K. Vines
Resources: Michael W. Liemohn
Visualization: Michael W. Liemohn, Robert Redmon
Writing – original draft: Michael W. Liemohn

Abstract

The Geospace Environment Modeling (GEM) program regularly issues “community challenges” in which researchers examine a particular space physics phenomenon or geomagnetic activity event, often running numerical models to assess dominant processes and understand the timing and relationship of observed signatures. The GEM Methods and Validation Resource Group helps those GEM focus group leaders running challenges to maximize participation and optimize scientific return from the significant time investment of these endeavors. This article gives a brief history of GEM community challenges and details those best practices that lead to an inclusive and valuable experience.

Plain Language Summary

Over 30 years ago, a group of space scientists set out to coordinate efforts toward the creation of a community-wide numerical modeling resource. This led to the formation of the Geospace Environment Modeling program, and one of the regular activities of this program is the instigation of “community challenges.” These challenges typically select a particular geospace activity interval or a physical process and then rally the research community to participate in the analysis of this phenomenon. The practice has led to substantial new knowledge of Earth's space environment and significant advancements in numerical modeling capabilities of this region. Here, we describe the history of these community challenges, highlight the lessons learned, and collect the best practices that maximize participation and optimize scientific return.

1. Introduction

The Geospace Environment Modeling (GEM) Program has existed for over 30 years as a funded element of the Geospace section within the National Science Foundation (NSF). It strives to spur action within the magnetospheric physics research community on specific topics. A key tool used by GEM community researchers is the “community challenge,” a grassroots approach to focus effort on a particular plasma process or interesting interval of geomagnetic activity (Hietala et al., 2020; Lyons et al., 1996). A hallmark of the GEM program since its very beginning, a number of community challenges have been conducted throughout the intervening decades. Much about magnetospheric physics—and fundamental plasma physics—has been learned from these challenges. Moreover, lessons have been learned about conducting these challenges. This article coalesces those lessons into a best practices guide for running a GEM community challenge. Our goal in writing this is to aid not only the community challenges of the magnetospheric physics research community but also similar community-wide efforts in other disciplines.

1.1. History of the GEM Program

To fully appreciate and understand the concept of community challenges, it is useful to first introduce the GEM program and its organizational structure. In response to a congressional promise of doubling the foundation's budget, the Global Geosciences initiative was created (NSF, 1987). As described by Roederer (1988), the magnetospheric physics research community heeded the call of NSF for ambitious new ideas. This summary of a gathering in August 1987 defines the GEM program as an interconnected sequence of research campaigns—as graphically depicted in Figure 1—that leveraged the new technology of online data centers and supercomputing resources. It detailed an ambitious timeline of robust magnetospheric physics advancements in the 1990–1995 timeline. As seen in Figure 1, the GEM program would tackle everything “magnetospheric” in nature, from the dayside interaction with the solar wind, to nightside magnetotail dynamics, to near-Earth convection and

Writing – review & editing: Michael W. Liemohn, Lutz Rastätter, Alexa J. Halford, Yihua Zheng, Katherine S. Garcia-Sage, Robert Redmon, Sarah K. Vines

ionospheric coupling. Campaigns focused on individual topics and questions within the GEM community. These campaigns often used community challenges to help push forward the understanding within these topics.

This initial meeting and report were followed by two pre-program workshops in 1989, one on magnetospheric theory and the other on observations. Anonymous (1990), perhaps written by or in conjunction with the GEM Steering Committee chair at the time, George Siscoe, describes a critical and unique element of the GEM program philosophy: "Unlike most NSF programs, this one was, and is, goal directed. It is aimed at developing a geospace general circulation model (GGCM)." It further states: "The coordinated attack on the magnetosphere's poorly understood region leads, after five intensive, problem-focused campaigns, to a sixth, assembly and-assessment campaign, to codify and test the GGCM." This sixth campaign was targeted at the creation and testing of the GGCM. Rather than one model emerging from the effort, several were created. In the early 2000s, two working groups on "modeling methods" and "metrics and validation" were initiated. These merged and eventually became the Methods and Validation (M&V) Resource Group (RG) within the GEM program in 2018.

Soon after the first full GEM summer workshop in 1991, a full definition of the GEM program was provided by Dusenbery and Siscoe (1992). This emphasized a distributed leadership structure that fostered community buy-in of coordinated research efforts. Regarding the creation of the GGCM, data-model comparison arose as an integral component of the process: "A major objective of the GEM program is to bring these two camps together so that the models can tell the experimenters what is happening physically and the data collected can help to constrain and better improve the models."

Three more Eos articles further explained the concept of the GEM program. Lotko (1993) listed the milestones of the GEM program, detailing the two initial campaigns. The first was the Boundary Layer and Cusp campaign, focusing on the dayside magnetosphere, and the second was the Substorm campaign, focusing on this enigmatic yet fundamental phenomenon of the nightside magnetosphere. Siscoe and Fedder (1994) introduced the plan for creating the GGCM, listing the steps needed to make a coupled code. The final one in the series, Lyons and de la Beaujardiére (1994), described the first geomagnetic activity interval "chosen for detailed analysis": 27–29 January 1992. This interval was the basis of the first "community challenge" and was the main focus of the Boundary Layer and Cusp campaign throughout that group's term within the GEM program.

The GEM program flourished through its first 5 years (Spence, 1996). Not only did the community gather for a dedicated GEM workshop every summer, but it also instituted a half-day "GEM mini-workshop" preceding the annual meeting of the American Geophysical Union, held in December. This allowed regular progress toward coordination of the community-wide efforts. The challenges were central to these discussions and the twice-a-year meetings of interested researchers enabled regular action toward GEM program goals. While some researchers were funded by the NSF specifically for GEM-related research or leadership, many participated on multiple programs. Regarding observations, ground-based instrumentation for magnetospheric and ionospheric physics has largely been funded by NSF programs and, because of the magnetic field connection between the ionosphere and magnetosphere, measurements of the Earth's upper atmosphere are directly relevant to the GEM program (Lyons, 1995). Additionally, various magnetospheric-focused NASA satellite missions were critical to GEM program advancement.

The GEM program did not end in 1995. Many in the magnetospheric community appreciated the informal atmosphere of the GEM workshops, the grassroots involvement of researchers across a range of demographics, and the collaborative nature of the challenge activities. The campaigns cycled through all magnetospheric regions with some ending and others beginning throughout the next decade. In 2008, the GEM Steering Committee approved a significant restructure of the program, replacing the campaigns—each containing three to five working groups—with a new Focus Group (FG) format with Research Areas. Each FG, competitively proposed by the community and selected by the Steering Committee, has a maximum term of 5 years, enforcing a regular shift in the content and emphasis of the GEM program and workshop. One of these FGs was on Metrics and Validation, centered on quantifying GGCM accuracy and raising awareness of the underlying techniques used in data-model comparisons. Moreover, it enabled the other FGs to effectively run their community challenges. This FG was reproposed each time it approached expiration, with changes in leadership with each of these proposals. Because it serves a different purpose than a standard FG, a new structure was initiated in 2020, the permanently standing Resource Group (RG), of which the Methods and Validation (M&V) Resource Group was the first. Figure 2 shows the modern organizational structure of the GEM program Research Areas with respect to

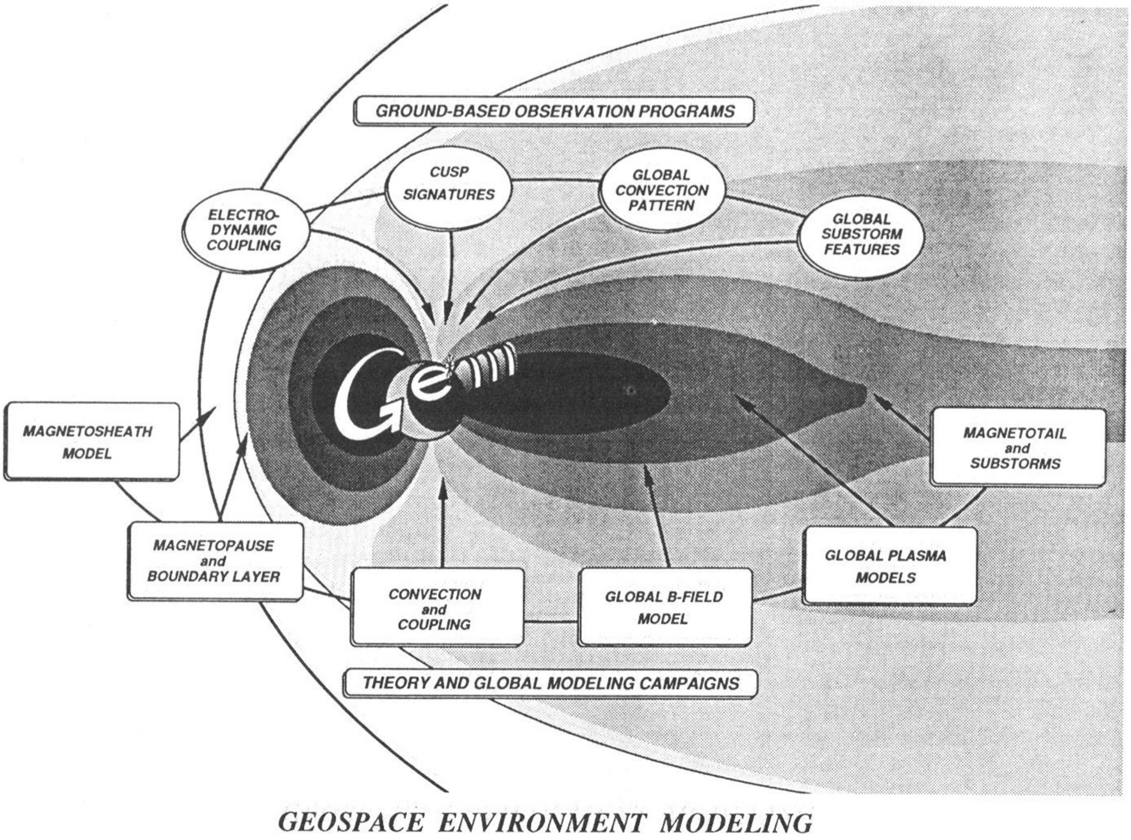
GROUND-BASED OBSERVATION PROGRAMS
CUSP SIGNATURES
GLOBAL CONVECTION PATTERN
ELECTRO-DYNAMIC COUPLING
GLOBAL SUBSTORM FEATURES
MAGNETOSHEATH MODEL
MAGNETOTAIL and SUBSTORMS
MAGNETOPAUSE and BOUNDARY LAYER
CONVECTION and COUPLING
GLOBAL B-FIELD MODEL
GLOBAL PLASMA MODELS
THEORY AND GLOBAL MODELING CAMPAIGNS
GEM
GEOSPACE ENVIRONMENT MODELING

Figure 1. A schematic of the regions and processes within the scope of the Geospace Environment Modeling (GEM) program (Roederer, 1988).

magnetospheric region, as well as a listing of the other, more cross-cutting and integrative activities conducted with the scope of the GEM program.

1.2. History of GEM Challenges

There have been at least 14 community challenges throughout the GEM program era. These are summarized in Table 1, listing the years, main approach, and key summary papers for each. Let us briefly summarize some of the main points along the timeline of GEM challenges over the last 3 decades.

The first challenge, often called at the time the Grand Challenge because of its novel approach to broad-based, yet coordinated, participation in a large-scale project, rallied community effort around the January 1992 interval. As summarized by Lyons et al. (1996) and Lyons (1998), this interval was driven by a strong interplanetary magnetic field (IMF) that rotated through all B_Y-B_Z clock angles (where the clock angle is given by $\theta = \tan^{-1}(B_Y/B_Z)$ and notes the contribution of one IMF component over the other in solar wind-magnetosphere interactions), allowing for a robust assessment of convection patterns in these different activity states. Several studies from this challenge further exemplify the coordinated approach to understanding this interval, including several that bring multiple data sets together (e.g., Hill & Toffoletto, 1998; Weiss et al., 1995) and those focused on comparing the emerging global models with these data (e.g., Fedder et al., 1998; Raeder et al., 1998; Winglee et al., 1997).

The other kick-off campaign of the GEM program initiated the first Substorm Challenge, as introduced by Raeder and Maynard (2001). Substorms are a class of geomagnetic activity originating from the magnetotail, and remain a topic of active research. While Raeder and Maynard (2001) focused on a single substorm event in 1996, many other intervals were considered to better contextualize the timing and intensity of particular substorm signatures (e.g., Lyons, McPherron et al., 2001). Again, several studies focused on the coordinated analysis of disparate data sets (e.g., Lyons, Ruohoniemi, and Lu, 2001; Sigsbee et al., 2001; Weimer, 2001) while others used this event to

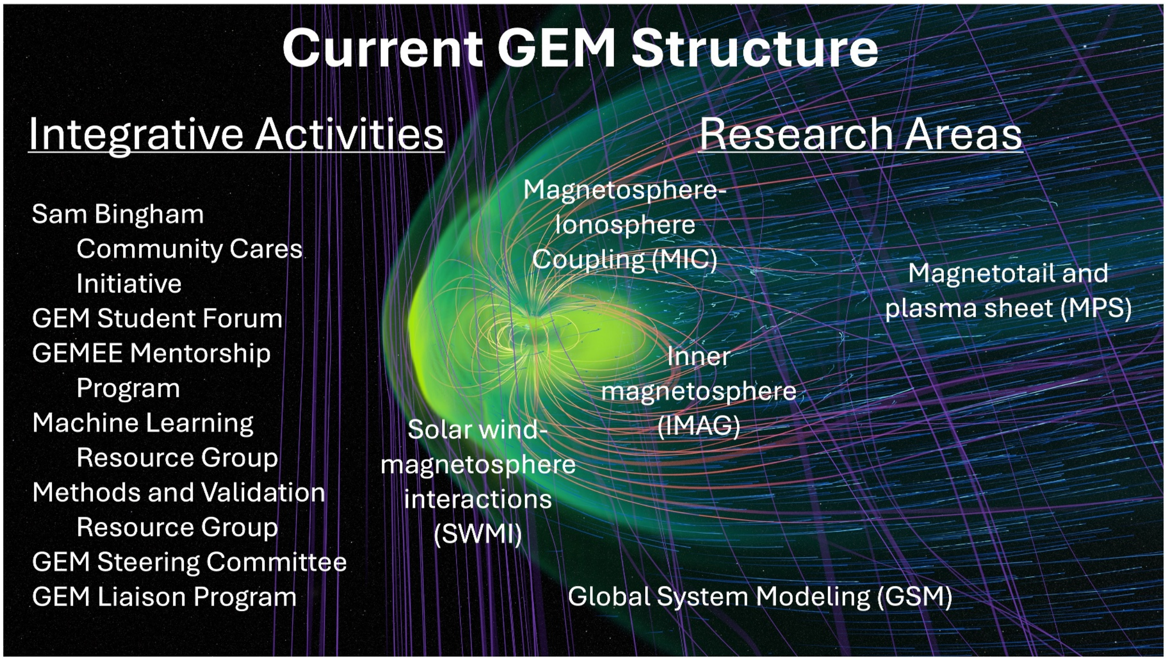
Current GEM Structure
Integrative Activities
Research Areas
Sam Bingham Community Cares Initiative
GEM Student Forum
GEMEE Mentorship Program
Machine Learning Resource Group
Methods and Validation Resource Group
GEM Steering Committee
GEM Liaison Program
Magnetosphere-Ionosphere Coupling (MIC)
Magnetotail and plasma sheet (MPS)
Inner magnetosphere (IMAG)
Solar wind-magnetosphere interactions (SWMI)
Global System Modeling (GSM)

Figure 2. The current GEM program structure, listing the research areas that cover the full range of magnetospheric regions and processes, plus a listing of the integrative activities within the GEM program. Background image courtesy of the NASA Visualization Studio, showing the magnetospheric topology as produced by the Multiscale Atmosphere-Geospace Environment (MAGE) model (Lin et al., 2022).

Table 1
GEM Challenges

Challenge name	Publication years	Approach	Summary paper(s)
Grand Challenge	Late 1990s	Centered on the January 1992 event	Lyons et al. (1996) and Lyons (1998)
Substorm Challenge	Early 2000s	Centered on a single substorm in 1996	Raeder and Maynard (2001)
Inner Magnetosphere and Storms Events	Early 2000s	Three storms: May 1997 September 1998, October 1998	No single summary paper produced
Reconnection Challenge	Early 2000s	Idealized forced reconnection scenario	Birn et al. (2001)
Inner Magnetosphere/Storms Assessment Challenge	Mid-2000s	Two storms for ring current, two storms for radiation belt	Liemohn (2006)
2008–2009 GEM Modeling Challenge	Early 2010s	Analyzed by CCMC; four storms, ranging from “small” to “super”	Pulkkinen et al. (2010, 2011)
SWPC Challenge	Early to mid 2010s	Conducted by CCMC; used six events, including 2 “surprise” events	Pulkkinen et al. (2013)
MHD-Empirical Modeling Comparison	Mid 2010s	Conducted by CCMC; idealized runs across a range of solar wind inputs	Gordeev et al. (2015)
Ionospheric Conductance Challenge	Late 2010s to early 2020s	Analysis by CCMC; centered on December 2006 event	Rastätter et al. (2016), Shim et al. (2018)
Spacecraft Surface Charging Challenge	Late 2010s	Centered on the March 2013 storm	Yu, Rastätter, et al. (2019)
Inner Magnetosphere Cross Energy-Population Interactions Challenge	Mid to late 2010s	Focused on science of cross-population coupling in the inner magnetosphere	Yu, Liemohn, et al. (2019)
Radiation Belt Challenge	Mid to late 2010s	Focused on model comparison with Van Allen Probes observations	Tu et al. (2019)
Dayside Kinetic Challenge	Early 2020s	Centered on THEMIS and MMS magnetopause and magnetosheath observations from 2017 to 2018	Dimmock et al. (2020) and Hietala et al. (2020)
Magnetotail at Lunar Distances Challenge	Early 2020s	Centered on ARTEMIS data	Runov et al. (2023)

test the ability of global models to accurately reproduce substorm features (e.g., Raeder et al., 2001; Ridley et al., 2002).

A third community challenge of the 1990s was the selection of geomagnetic storms by the Inner Magnetosphere and Storms (IM/S) Campaign. One major thrust of this effort was the hot (∼keV) ions that carry most of the electric current in the inner magnetosphere during large geomagnetic storms, with analysis of the events from this challenge leading to the theory that loss through the dayside magnetosphere through convection on open drift paths dominates early decay of the ring current (e.g., Liemohn et al., 1999) and that the ring current is highly asymmetric during the main phase of storms (e.g., Kozyra & Liemohn, 2003; Liemohn et al., 2001). The energetic (∼keV to MeV) electrons were the other charged particle population of interest for this challenge, with resulting studies leading to new understanding of microburst losses of the radiation belts (e.g., Lorentzen et al., 2001; McAdams et al., 2001; O'Brien et al., 2004).

Arguably the most successful and impactful community challenge of the GEM program is the Reconnection Challenge (Birn et al., 2001). Conceived as a simple head-to-head assessment of different types of numerical approaches to an idealized yet classic problem of magnetic reconnection, this activity led to substantial new understanding about the microphysics of this ubiquitous process. The collection of studies from this challenge covered the full range of relevant physics, from ideal magnetohydrodynamics (MHD) to extensions that included additional physics representations, such as Hall MHD (e.g., Ma & Bhattacharjee, 2001), current-dependent resistivity (Otto, 2001), anisotropic pressure (Birn & Hesse, 2001), nongyrotropic pressure (Kuznetsova et al., 2001), and full particle-in-cell modeling (e.g., Hesse et al., 2001; Pritchett, 2001; Shay et al., 2001). This challenge continued to have a lasting effect on magnetic reconnection studies for years afterward (e.g., Birn et al., 2005), with many of these studies still having profound influence on the field to this day.

In its waning years, the IM/S campaign conducted a second community challenge, selecting two intense geomagnetic storm intervals that were well observed by the Imager for Magnetopause-to-Aurora Global Exploration (IMAGE) mission. Focusing on the plasmasphere (e.g., Goldstein et al., 2005; Liemohn et al., 2006), ring current (e.g., Chen et al., 2006; Ganushkina et al., 2006; Jordanova et al., 2006; Milillo et al., 2006), and radiation belts (Huang et al., 2006; Miyoshi et al., 2006), it especially emphasized advancing knowledge about the physical processes connecting these plasma populations, as summarized by Liemohn (2006).

With the advent of the Metrics and Validation FG, the practice of reporting community challenge results switched to a mode of providing model output/predictions to the Community Coordinated Modeling Center (CCMC) and allowing an independent investigator to lead the assessment. This practice yielded a number of significant studies quantifying magnetospheric numerical models against various data sets. For example, Pulkkinen et al. (2010, 2011) focused on ground-based and geosynchronous satellite data-model comparisons for MHD modeling results of the 2008–2009 event challenge, Rastätter et al. (2011, 2013) further evaluated these event simulations against geosynchronous magnetic fields and the Dst index, and Honkonen et al. (2013) continued this effort with comparisons of the four MHD model results against additional satellite observations throughout the magnetosphere. Perhaps the most influential of these CCMC-led studies is that of Pulkkinen et al. (2013), for which the models themselves were passed to CCMC staff (i.e., not just the model outputs) and led to the selection of one of the models for implementation as an operational space weather prediction asset by NOAA (the “SWPC Challenge”). Gordeev et al. (2015) presented a systematic analysis of MHD models against empirical magnetic field models, selecting a large number of magnetospheric shape parameters for analysis. A follow-on study by Collado-Vega et al. (2023) conducted a similar analysis of magnetopause locations but for real events and compared against all available satellite data. The joint challenge on ionospheric conductance between the GEM program and its sibling organization, the ionospheric-focused Coupling Energetics and Dynamics of Atmospheric Regions (CEDAR) program, was also run this way, with Rastätter et al. (2016) focusing on high-latitude ionospheric data-model comparisons, and then Shim et al. (2017, 2018, 2023) conducting a series of data-model comparison studies on the global ionospheric state (Özturk et al., 2020). Additional studies arose from this challenge but the efforts were decentralized, to the point that the resulting papers do not even mention the GEM challenge, even though they grew out of those discussions (e.g., Khazanov et al., 2019). A final challenge to mention, on spacecraft surface charging, used this approach of gathering results into a single study (Yu, Rastätter, et al., 2019).

In recent years, several FGs have gone back to the decentralized format of community challenges. One challenge, focusing on the inner magnetosphere (summarized by Yu, Liemohn, et al., 2019), allowed participants to publish their studies as the results became available, rather than imposing a deadline through a single collection (e.g., Chen et al., 2014; Glocer et al., 2016; Tu et al., 2013). A similar approach was taken for another challenge on radiation belt dynamics (summarized by Tu et al., 2019), specifically focusing on comparisons with Van Allen Probes observations (e.g., Albert et al., 2018; Brito & Morley, 2017; Engebretson et al., 2018; Ma et al., 2018). Another centered on the kinetic processes in the dayside outer magnetosphere, magnetopause, and magnetosheath and used the journal special collection as motivation for community involvement (Dimmock et al., 2020; Hietala et al., 2020). This approach resulted in substantial advancements about transient phenomena at the magnetopause and bow shock, as well as a comprehensive collection of data-model and model-to-model comparisons (e.g., Chen et al., 2020; Guo et al., 2020; Trattner et al., 2020; Zhang & Zong, 2020). Finally, a very recent challenge centered on the magnetotail at lunar distances, taking advantage of the Acceleration, Reconnection, Turbulence and Electrodynamics of Moon's Interaction with the Sun (ARTEMIS) spacecraft in orbit around the Moon, which passes through the magnetotail in a region that is still not well understood (e.g., Kamaletdinov et al., 2024; Runov et al., 2023).

1.3. Learning Lessons From Past GEM Challenges

It is clear from the discussion above that many papers, findings, and collaborations have emerged from the group-wide challenges undertaken by the GEM community. A variety of management styles have been implemented to foster participation from researchers, different approaches have been used regarding inter-code comparisons and data-model metrics, and the timelines of these challenges have spanned from a single year to a steady stream over several years. From all of these experiences, we have learned much about how to create a space for researchers to want to work together toward a common objective. While there has been no single methodology for conducting a challenge, it is useful to assess what has worked well and what has caused frustration with respect to running and participating in these group efforts. Section 2 presents the lessons learned from past challenge leaders, as documented in summary articles and retrospectives that they have written. Section 3 shares additional ideas from the GEM community, as collected through a series of conversations at the GEM workshops in recent years. Section 4 covers some GEM-specific logistics of starting and running a community challenge within this program. A summary is presented in Section 5, including a synthesis of all of the best practices covered in the earlier sections and some suggestions for follow-up bibliometric studies on the effectiveness of challenges for advancing science.

2. Best Practices for Running Community Challenges: Lessons From Past Challenge Leaders

2.1. Summarizing Published Advice

When considering a community challenge, the first and arguably foremost decision is on the scope of the effort. A good example of such a statement of purpose is found in Pulkkinen et al. (2010): “The fundamental purpose of the GEM 2008–2009 Challenge is to quantify, for the given evaluation setup, the current state of the space physics modeling capability and to address the differences between various modeling approaches.” This challenge is clearly focused on validating the state of the art and quantifying the strengths and weaknesses of the model. That is, it was not particularly focused on scientific return but rather on understanding the present collection of available modeling tools. Another statement of purpose is found in Liemohn (2006):

> “A few magnetic storms, known as the GEM Storms, were selected early in the campaign life for group-wide study. This list has been extended over the years to include almost a dozen events, and the ‘‘GEM Storms’’ sessions were becoming a potpourri of individual science results, rather than a concentrated effort on a few storm intervals. A climatic activity was therefore introduced to refocus the IM/S Campaign’s attention on a short but well-selected list of events: the Inner Magnetosphere/Storms Assessment Challenge (IMSAC). An additional outcome from the IMSAC is the definition of metrics for inner magnetospheric model results.”

This one is centered on scientific return, steering the community toward particular geomagnetic activity intervals to coordinate that knowledge advancement. Note that challenges do not have to be arranged around real events. The reconnection challenge was completely focused on model-model comparisons of an idealized initial and boundary condition scenario. The challenge had very specific hypotheses to test with these model-model comparisons. And it worked: the summary paper of Birn et al. (2001) is one of the most highly cited studies from any of the GEM challenges.

Arguably the most comprehensive paper about how to run a GEM challenge is that of Hietala et al. (2020). This paper is entirely about lessons they learned from running the dayside kinetic challenge. One issue they dealt with is keeping the research community engaged throughout the duration of the challenge:

> "The biggest struggle has been maintaining focus, interest, continuity, and communication from year to year. The organizers need to tirelessly reiterate the message of the value of Challenge activities. Most of the efforts of the organizers go into communication, that is, e-mails. People need to be individually persuaded to commit their time, essentially for free, for something that is most likely not directly in their interest, as they are presently funded to do something else."

It took substantial time and energy for the challenge leaders to foster engagement across several years of effort. A related point raised by Hietala et al. (2020) was about the timing of work, stating that "everything always seems to get done during the 2 weeks right before the Summer Workshop." This speaks to the need for shorter, rather than longer, timelines for a community-wide challenge. They argue that shorter is actually better for the scientific return; the challenge must then be narrowly defined with a central focus. Keeping the timeline short, and getting the community to agree to deadlines, reduces the need for sustained engagement. When a challenge continues past a year or so, it becomes a cumbersome burden to the organizers.

Hietala et al. (2020) continues with keen advice about metric selection. They urge each challenge group to think about the purpose of any data-model comparisons and make their own decision about what metrics best suit that need. Indeed, they warn, "A single number used as a metric for some previous Challenge may not be good for your purposes." When planning a data-model comparison effort as part of a challenge, organizers should carefully consider what elements of the available data sets are most critical for the assessment. They recommend the creation of a comprehensive listing of metrics options, including advice about what metrics work best in certain situations. Others, such as the International Space Weather Action Teams (https://www.iswat-cospar.org/), have also been conducting community discussions on best practices for metrics, such as for energetic particles (Zheng et al., 2019, 2024), auroral electrodynamics (Robinson et al., 2019), and ground magnetometer perturbations (Liemohn et al., 2018; Welling et al., 2018). A summary of common metrics is given in Appendix A.

The M&V RG leadership strongly encourages the adoption of open science best practices, in particular toward the FAIR (Findability, Accessibility, Interoperability, and Reusability) model (e.g., Wilkinson et al., 2016). A challenge activity that follows the FAIR principles more readily allows others to expand on the research with their own study later, including the use of other metrics and data-model comparison methods.

A critical concern for a number of challenges has been the availability and usability of relevant observations. In their summary of the first substorm challenge, Raeder and Maynard (2001) made a strong point about the sparseness of space science data: "In particular, magnetospheric data are restricted to statistical average pictures, point measurements, and ionospheric synoptic maps. Even the most comprehensive data sets have limited resolution." This situation still exists today; it seems that we will never have enough data to satisfy our need for comprehensive model comparison. Similarly, Dimmock et al. (2020), discussing the dayside kinetic challenge, raise the issue of being too specific in the data-model comparisons chosen for a challenge, a fact often driven by the sparseness of magnetospheric data:

> "It is shown that matching very specific upstream conditions can severely impact the statistical data if limits are imposed on several solar wind parameters. We suggest that future studies that wish to compare simulations and/or single events to statistical data should carefully consider at an early stage the availability of data in context with the upstream criteria. We also demonstrate the importance of how specific IMF conditions are defined, the chosen spacecraft, the region of interest, and how regions are identified automatically."

That is, modeling should be statistical, like the data compilations to which they are being compared. Because full magnetospheric models are often computationally expensive to run, it is usually the case that only a small handful of simulations are conducted, yet then compared with statistical compilations of observations. One way to approach this is to do a statistical analysis on simulations that are extended enough and driven by different solar wind conditions to be reasonably compared to observational statistics (e.g., El-Alaoui et al., 2013; García & Hughes, 2007; Gordeev et al., 2015; Guild et al., 2008; Haiducek et al., 2020; Katus et al., 2015; Liemohn & Jazowski, 2008; Morley, Brito, & Welling, 2018; Morley, Welling, & Woodroffe, 2018; Wiltberger et al., 2015).

A promising avenue for future studies is to determine how this type of approach can be done in a community challenge setting.

A related issue to this is data availability and ease of use by the challenge participants. Tu et al. (2019) make a good point about challenges as a place for collection of useful data, the one-stop location for all relevant files to conduct a robust data-model comparison:

> "A key part of the challenge activity was the collection—in publicly accessible cloud storage—of supporting data, especially of the type used for setting model boundary conditions or incorporating physical processes. Contributions to this effort included data products tailored to these events. Links to the community resources curated by the FG, including readme files and overview presentations for the challenge events."

This data collection and accessibility is a non-trivial task that takes time and effort by knowledgeable experts. Related to this, Öztürk et al. (2020), in their article on the conductance challenge, noted that a high priority of that leadership team was systematic quantification of the uncertainties in the data sets to which the models were being compared.

To conclude this section, we return to Hietala et al. (2020), who make a crucial point that is essential to repeat here. This is with respect to funding: "We would recommend the GEM community to consider writing collaborative proposals, one per Challenge." It seems to be a regular issue with challenges; we are asking the research community to conduct work and participate in a group-wide assessment, often without compensation for their time. Funding of participation would address the sustained engagement issue, allowing researchers to devote substantial time to the challenge effort. When asking the community to do it "for free," then each researcher must have their own funding source to participate. To accomplish full involvement, the objectives of the challenge must align with those of every researcher participating in the project. This is a difficult feat to accomplish.

A group-wide funding request could alleviate many of the problems with running a challenge. This could be partially done through a collaborative grant proposal to the National Science Foundation, or to some other federal agency, like the NASA Heliophysics Supporting Research program or Living With a Star program (assuming there is a Strategic Capability or Focus Science Topic team call solicited within the challenge timeline). However, there are logistical and budgetary challenges for funding agencies that make this pathway a generally non-viable solution for supporting larger groups participating in community challenge events, such as the fluid nature of the crew participating in the challenge activities from year to year. NSF has a potential model for this; the GEM Workshop coordinators receive funding to partially support a few dozen students to attend the workshop. That is, the grant includes travel funding for an unknown group of people, to be selected anew each spring, before the workshop. A similar method could be followed to facilitate participation in a community challenge.

Another potential pathway is to work with funding agencies, specifically the GEM program at NSF, to organize solicitations around community challenge events specifically. Multiple teams of smaller groups could then propose studies that all seek to address that particular community challenge. Organizing such a call requires early and active engagement with the funding agency, a budgetary environment that can support such a call within the general portfolio of that agency's program, and advanced planning and community buy-in for the challenge event that would be the focus of such a solicitation.

Finally, encouraging increased participation and engagement with already funded facilities and mission teams for challenge event contributions may also help with lowering barriers to community members. Needed funding support could then be distributed specifically to those that do not have directly relevant lines of funding for participation.

2.2. Organizational Framework Advice

In terms of organizing the challenge and conducting the task of comparison and analysis, two dominant methods have emerged. One option is to distribute the work effort across the entire community, allowing each researcher to analyze their own model results. Often, a summary paper is written synthesizing the results from the many individual studies emerging from the challenge. This method was applied to the first substorm challenge (e.g., Raeder & Maynard, 2001), the GEM reconnection challenge (e.g., Birn et al., 2001), the IM/S assessment challenge (e.g., Liemohn, 2006), the radiation belt modeling challenge (e.g., Tu et al., 2019), and the dayside kinetic challenge (e.g., Dimmock et al., 2020). An advantage of this choice is that studies can be published as they are ready, and provides a mechanism for recognition of the contribution of individual community members through publications. Another useful outcome is that the broad inclusion of many authors yields a wide variety of findings from the challenge, as each participant uses the challenge to build and expand their existing research projects and obligations. The drawback is that, with a few exceptions, the challenge rarely yields a definitive head-to-head comparison of models.

Another option is to collect the model results and have an independent group or person lead the analysis effort. The individual groups might publish follow-on papers using the same output, but such papers are often indirectly related to the goals of the challenge. The collection point has often, but not always, focused on the CCMC, with the staff there taking the lead on the papers resulting from the challenge. This method was used for the Grand Challenge (Lyons et al., 1996), the GEM 2008–2009 challenge (e.g., Pulkkinen et al., 2010, 2011), the SWPC selection challenge (e.g., Pulkkinen et al., 2013), the conductance challenge (e.g., Ozturk et al., 2020; Rastätter et al., 2016; Shim et al., 2017), and the surface charging challenge (e.g., Yu, Rastätter et al., 2019). The upside of this approach is that the output from the models are directly compared, not only against each other but also against observations. This is a powerful technique for discovering the strengths and limitations of the bevy of models included in the challenge, setting the stage for eventual transition to operations. Another advantage of this approach is that the demand on the research community is smaller than the other approach; each person contributes a data set or model output and then the CCMC staff person does the majority of the work as part of their NASA-defined and funded duties. The downsides of this option are that it usually yields fewer papers (perhaps counter to career-building first authorships for many researchers), as well as often being focused on assessing the model quality against observations (i.e., strict validation) and thus might have a smaller chance for a breakthrough scientific finding.

Neither of these approaches is inherently better or worse than the other. As noted above, they each have many strengths. The decision of which of these methods to apply for a new challenge depends on the objective: a more science-centric challenge might be better if choosing the distributed-work option while a more application-oriented challenge might be better using the centralized-work option. Moreover, they are not mutually exclusive, but any challenge could have both. For example, the Pulkkinen et al. (2013) challenge led to several follow-studies by individual research teams (e.g., Toth et al., 2014) and other CCMC inter-model comparisons (e.g., Glocer et al., 2016).

3. Suggestions From the GEM Community

We have had several discussions during sessions at past GEM Workshops and Mini-GEMs about best practices for community challenges. Here are some helpful highlights that have emerged as repeated themes in these discussions. In addition to the narrative below, specific website links suggested by the community as potentially helpful to future challenge leaders are listed in Appendix B.

3.1. Formation and Setup

When recruiting researchers to participate in the challenge, it is good to follow best practices of group formulation (NRC, 2015, and references therein). These include the following: (a) thinking about the breadth of skillsets and backgrounds of the group members, including non-traditional skills (e.g., story tellers, or code developers); (b) actively encouraging a broad mix of people to participate in the challenge; (c) creating a plan for keeping track of “membership” in the challenge activity; and (d) having a plan for how to bring in new researchers after the challenge is underway. Also, organizers should strive to make the community conducting the challenge as diverse as possible; it has been shown that such an approach will, on average, yield more creative and impactful science results (e.g., Guterl, 2014; Hunt et al., 2018; Medin et al., 2014; Rock & Grant, 2016; Stark et al., 2021). This has even been found to be true within the Earth and space sciences (e.g., Lerback et al., 2020; Moldwin & Liemohn, 2018).

A regularly occurring topic is that of how to interact as a geographically dispersed and diverse group of colleagues toward this common goal. One critical step is to agree on a code of conduct—how to act respectfully toward others during the discussions. There are several examples given in the Additional Resources (Appendix B). This plan should be not only formulated and enacted but also revisited—to remind the group and inform new members about this code—and revised, should any concerns about it arise.

Another important decision in group dynamics is choosing how to interact. There are many possible options for community communication channels, from simple methods like email groups to more formal project workspaces like Trello or Monday.com. Conversations can be conducted via informal group networks like Slack or Discord, or over a more structured process like a group Github website. However, care should be taken to ensure that all who want to participate can, as some institutions limit which tools can be used. A decision on a file management process needs to occur, such as using a shared drive as a repository or requiring "files on request" from the originator. If a single independent group is gathering model output and conducting the data-model comparisons, then sending files directly to them would most likely be easier than creating an online structure. Some groups find a formal "Rules of the Road" agreement to be useful in creating organizational and interactional structure, while other groups have found such agreements to be too strict and limiting on the scientific collaborations. This needs to be discussed early and a consensus reached so that all participants understand the expectations of involvement.

A related topic is the question of an open science policy. Following the FAIR principles is best; it ensures openness throughout the challenge process and enables follow-on studies that would build on the methods and findings of the challenge effort. This places a burden on community volunteers, so a discussion should take place and several questions need to be considered. Will observational data be made accessible and easily adoptable for model comparisons? Will numerical model output, and observational data be available to all, or kept within the control of the originating scientists or a select group of designated analysts? Will analysis code be freely shared between challenge participants, or will each researcher have to develop their own methods for their particular analysis? The discipline is moving rapidly toward a fully open science paradigm, but this takes time and effort to maintain. (Generally, NASA satellite mission teams are required to provide public access to science-quality data; however, higher-order or derived products have not always been made available.) If choices toward openness are made, then who will curate the repositories for the data, model output, code, or more general computational analysis environments? The research community embarking on a challenge should discuss this and make groupwide consensus decisions about the process.

3.2. Defining the Challenge

A major point that keeps arising in discussions about community challenges is the need to have a clear science or validation objective for it. These are not the same; a scientific objective might produce little in determining which model is best suited to possible operational requirements, and vice versa. The leaders—or, if they choose, then the community as a whole—needs to decide on the purpose of the challenge. Focusing on relevance to the particular Focus Group goals, and being timely with current interests within the research community, will most likely maximize involvement in the challenge.

The other crucial choice about a community challenge is whether to focus on one or a few selected event intervals or not. This entirely depends on the answer to the first major point on the purpose. Choosing times keeps the challenge centered, especially for a validation effort. If, however, the focus is on a particular science issue, then perhaps several events, or even leaving the event list open, perhaps all the way to the point of including large-scale statistical studies, allows that science topic to be investigated in the way that each participant thinks is best and is able to be studied with their model/set of observations. Events with diverse observations are, in general, more appealing to a broader cross section of the research community and might increase participation. There is no single, correct answer to the number of events to be assessed, not a set method recommended here to reach this decision, but a choice needs to be made on the event or event list early in the process.

If choosing real events for the challenge, then the question of which data-model comparison metrics to choose needs to be answered. Also discussed in Section 2.1, this is a recurring theme in GEM community discussions on challenges and an additional point can be made. If the challenge is focused on validation for possible future usage in an operational setting, then these end users should be included in conversations about data-model comparisons, not only of which aspects of the model output should be compared to which data sets, but also of which method of comparison (i.e., metric choice) should be used (e.g., Halford et al., 2019; Morley, 2020). If the challenge has a science objective, then data-model comparisons take on a different focus, and a different set of metrics might be of higher interest. See, for example, the study of Liemohn et al. (2006), who chose to focus on metrics of the large-scale morphology of inner magnetospheric plasma populations rather than fine-scale data-model comparisons (such as of particle fluxes in individual energy channels). It could be, for example, that in validation for operations the main concern might be only the extreme values, while for an example scientific objective the emphasis might be more directed at getting a good overall fit. One metric cannot do both. Summary lists of metrics are given in Appendix A to help with this selection. While choosing several metrics is usually a preferred method, it should be an appropriate set of metrics optimized for the purpose of the challenge. Again, this can take a lot of time, so conducting these discussions and decisions on metrics fairly early in the challenge timeline can optimize the chances for a successful challenge.

Of course, there is always the option of developing contrived scenarios for the challenge, as was done with the GEM Reconnection Challenge. Idealized but well-specified initial and boundary conditions greatly clarify model-model comparisons, allowing for similarities and differences to be identified and discussed. In this case, there are no data-model comparisons that need to be selected.

Once defined, a big question facing the challenge leadership is how to motivate people to participate in the challenge. In addition to following the suggestions above, another tactic is to broadly publicize the challenge through a variety of methods, and to do this well before any publication deadlines. In addition, challenges should be designed to be model agnostic, not a priori favoring one modeling style (or even one particular code) over others. Challenge leaders need to steer clear of the winner-loser mindset when defining challenges requiring modeling and simulations, as this is a huge disincentive to participation. (As was demonstrated especially well by the reconnection challenge, the physics represented by different modeling regimes are all useful depending on the different physical scenarios and spatiotemporal scales of interest, and much can be learned from differences in modeling outputs for these challenges.) Furthermore, a good strategy for a well-run challenge is to define clear deliverables and benefits. This encourages people to participate because they know what is expected of both them and of the challenge as a whole.

Regarding participation, NSF subsidizes student attendance and participation in the summer workshop. Students generally tend to have more time for sustained engagement, so it is natural for FG leaders to consider the inclusion of students in challenge activities. Students are often being trained on the latest data science techniques (e.g., AI/ML software stacks) and computational platforms (e.g., Colaboratory and Jupyter Lab computational notebooks, Cloud computing workflows). Some examples of this participation could be developing and maintaining a website or file repository, conducting model runs or analysis, or compiling a literature review of the topic for use in any publications arising from the challenge. Depending on their level of effort, they could even lead the presentation and publication of the study. Senior scientists involved in a challenge should take the group activity as an opportunity to interact with students and perhaps, if the situation is appropriate, offer to more formally mentor them in conjunction with their research advisor.

3.3. Additional Suggestions From the Community

In workshop discussions, the conversation occasionally turns toward "what if" scenarios of other structures, formats, or considerations regarding GEM challenges. There are three topics that particularly resonated with researchers.

One suggestion pertains to the way in which the challenge is conducted. Specifically, there was interest in using the method of a "research sprint" within a challenge, in which people come together for a dedicated multi-day work-focused (as opposed to presentation-focused) session on the particular topic. This is reminiscent of the Coordinated Data Analysis Workshops (CDAWs). Pre-internet, these were a primary means of bringing together people with expertise on relevant data sets and models for a concentrated effort on a single science objective. CDAWs were so successful that the community pushed federal agencies to fund one-stop repository for space physics data, eventually leading to the creation of NASA's CDAWeb (https://cdaweb.gsfc.nasa.gov/). The analogous entity on the modeling side is the CCMC. This focused gathering approach is essentially the point of an international team of the International Space Science Institute (ISSI), for which 8–12 researchers gather in either Bern, Switzerland or Beijing, China, for two or three weeklong work sessions to accomplish a focused science goal. While this type of multi-day effort does not fit within the structure of the GEM Workshop, the scientific productivity of the CDAWs and ISSI teams (e.g., https://www.issibern.ch/results/) shows that such activities are worthwhile and could be considered for future GEM challenges.

A second suggestion relates to the way in which we disseminate the results and findings from a challenge. Specifically, the lessons from investigative dead ends are often not captured in published literature. Publication length limits tend to hinder the inclusion of the trials and errors that didn't work toward the eventual successful experiment and analysis that did lead to a new scientific finding. With Open Access publishing, however, many journals are removing length limits, which allows for the possibility of including more of the unfruitful methods and analysis. There is value in others knowing about these attempted endeavors, if nothing else so that others do not invest serious time repeating a process that won't lead to new results. It is therefore encouraged that, when writing a paper from a challenge, to write up all paths taken during the research project. For the GEM workshop sessions devoted to results from a challenge, it is encouraged to include presentation of unsuccessful work. Explaining why the analysis failed is a large part of explaining why the successful path succeeded. In fact, it is encouraged for the GEM M&V RG to hold a regular session devoted to this—an honest conversation about investigative dead ends in science methodology.

A third topic of interest to many in the GEM community is the intersection of machine learning (ML) models with respect to challenges (and, more broadly, data-based empirical models developed by any method). Previous challenges have not specifically sought the inclusion of ML techniques within the model set, and only a few have intentionally included empirical models in the analysis (e.g., Pulkkinen et al., 2013). Development of such models require a large data set that is split into training, validation, and test subsets, and if the chosen event intervals fall within the first two, then that ML model has a distinct advantage in outperforming other codes. There are other ways in which models can be tuned to best match specific features or intervals, and this is true for not only ML codes but also empirical and physics-based models. There is now a resource group within the GEM program on this exact topic of machine learning codes for space physics. It is encouraged to develop strategies for best using ML models in research community challenges.

4. GEM Challenge Logistics

4.1. Proposing a GEM Challenge

Because of the structure of the GEM program, the act of initiating a GEM community challenge is the task of Focus Group leaders. Often, FG leaders will include a challenge within their proposal, as one of the key activities to be conducted during their 5-year term. A quick glance through the FG proposals reveals that most include a challenge. This means that, of these, less than half actually follow through with conducting a community challenge (at least one that resulted in one or more publications that specifically mention the existence of a challenge motivating the study). Not every FG requires a challenge to accomplish their proposed science goals, so it is not surprising that some exclude this activity. That some FGs propose a challenge but then do not carry through with one is often the result of the direction taken by the community during the breakout sessions for that FG. While the FG leaders enter the process with a plan, the execution of that plan is accomplished with the input and participation of the research community, and the implementation sometimes leads to other FG actions.

4.2. M&V Resource Group Interactions With Challenge Leaders

There are several strategies that seem to work well for beneficial interactions between the leadership of a GEM challenge effort and the M&V RG leadership. The first is that the M&V RG leadership should be aware of the proposed activities of all FGs, especially the new ones, and be aware of the timeline for each FG that proposes a challenge as part of their 5-year plan. The M&V RG leadership should then actively reach out to FG leaders and regularly work with them throughout the formulation of the challenge, when many of the best practices above need to be implemented. As stated above, the M&V RG leadership strongly encourages the adoption of open science best practices, toward the FAIR model (e.g., Wilkinson et al., 2016), as is being mandated by many funding agencies, journals, and in other research communities. Elements of this include providing data (scientifically useful observational or numerical number sets) on searchable platforms with readily understandable metadata that clearly explains the content and format of the larger number set files. Code sharing for greater use of available data is also a key element of the FAIR model, enabling the reuse of data in follow-on studies. When the challenge is ending, the M&V RG leadership should encourage the challenge leaders to produce a summary document (hopefully peer reviewed and published) that includes a lessons-learned section on their experience running the community challenge.

5. Key Recommendations

This paper reviewed the history of GEM challenges and detailed many of the lessons learned from those past challenges. It also distilled the conversations about the GEM challenge process that the community has had over

Table 2
Summary of Recommendations for Running a Community Challenge, Such as a GEM Challenge

Aspect	Recommendation	Comments
Scope/purpose	Three main options for focus: science topic (real events or idealized set up); science aspects of selected real events; validation of models for eventual operational implementation (statistical or case study)	Decide this very early
Organizational framework	Two main styles: distributed analysis by all; independent analysis (by, e.g., CCMC)	Decide this very early
Overall duration	Keep it to under 2 years	Motivation and participation wanes for longer duration activities
Metrics selection	Choose data-model comparison metrics based on the science/validation aspects of key importance	Plan on significant discussion time for this
File management	Decide on "openness of data" policy; then designate a person/group to manage data and model output availability for others	Decide this early to maximize participation
Participation	Actively encourage a broad swath of the specific disciplinary community to engage	This might require personalized invitations
Student participation	Actively encourage students to take on challenge-related tasks	Students tend to have more bandwidth for sustained engagement and more familiarity with new tools
Group interactions	Develop a code of conduct, regularly remind people of it, and periodically review/revise it	Establish this early
M&V RG coordination	Reach out to coordination and validation experts, such as, within GEM, the M&V RG leaders, for configuring the challenge, logistical support, cross-FG coordination, and guidance on difficulties that arise	You are not alone

the past several years. It has even made some suggestions for future elements of challenges that have not been tried yet.

Table 2 summarizes the key recommendations for running a community challenge, such as a GEM challenge. There is no single "right" way to run a community challenge. There are, however, guidelines for making it run smoothly, for optimizing its impact on the targeted focus, for maximizing participation, and for getting through the problems that inevitably arise. It is hoped that these guidelines for running a community challenge will be of use not only to the magnetospheric physics community but also in general for group research activities conducted by other research fields.

One key point to remember is that a community challenge is a large-scale group effort, and bringing together a broad scope of expertise will help optimize the definition of the challenge and maximize the intended return. Participation should not be ignored or left to passive publicity to garner involvement but rather proactively solicited, including from those with experience running challenges.

This study has focused on gleaning the best practices from published reports about community challenges and in-person discussions at the GEM workshops. One aspect of challenges that has not been documented, here or elsewhere, is a systematic, quantitative assessment of success criteria for the GEM community challenges. Some of this exists within the annual leadership reports archived in the GEMStone newsletters (https://gem.epss.ucla.edu/mediawiki/index.php/GEMstone_Newsletter), but these are anecdotal summaries of workshop proceedings and activities, not a comprehensive review. The published articles cited in earlier sections are a better source of impact assessment, but these are often published during or shortly after the conclusion of the challenge and therefore most likely miss those studies that took longer to develop and any follow-on work. Furthermore, a robust assessment of this nature would be a serious time investment; more than an typical volunteer effort, even of GEM program leadership.

It is recommended, therefore, that such an assessment of the efficacy of GEM challenges be commissioned and conducted. Such an endeavor would most likely encompass a large-scale literature and bibliometric review as well as interviews with key participants. This effort could include an analysis of metrics such as the number of related publications, number of researchers involved, citations to those publications, early career investigators trained through the challenge activities, and advancements in software tools (both data analysis methods and numerical approaches). Going beyond first-tier citations and analyzing the larger related-paper "tree" would provide a measure of the long-term impact of each of these challenges, especially when compared to a similar analysis of randomly selected non-challenge-related papers.

Another possible follow-on study to this one could be a systematic review of the social psychology research into group dynamics. One such collection is the National Research Council report "Enhancing the Effectiveness of Team Science" (NRC, 2015). While many elements of the nine summary recommendations from the NRC report overlap with the guidelines summarized in Table 2, that report is targeted at a single but large project team and some aspects of those recommendations do not readily transfer to this more specialized context. Therefore, it would be useful to commission the creation of a review of the social dynamics scholarship specifically for research community challenges.

We have seen how challenge events have had multiple benefits for the GEM community, and we hope would be beneficial to others as well. Specifically, the challenge events have (a) helped to achieve dedicated study of challenging physics issues, open questions, and event intervals which have pushed our fundamental understanding of the space environment further, (b) enabled cross comparisons of modeling techniques both validating model results as well as understanding which physics may be dominant among the processes during these challenge events, and (c) bring collaboration and collegiality to broader discussions within a research community, which otherwise could easily become competitive.

Appendix A: Metrics Options

The following is a quick guide to some prominent metrics that could be of use to GEM challenge organizers. It is a compact and updated version of what appeared in Liemohn et al. (2021) and Liemohn (2023). They are arranged in two major groupings—based on the style of the data-model comparison approach—and then by specific categories based on the facet of the data-model relationship that they assess.

The two groupings are fit performance metrics (also called continuous metrics) and event detection metrics (also called dichotomous metrics). The first grouping uses the exact values of the two number sets (the observed values and the model output) while the second grouping converts all of the values into a yes-or-no binary classification based on whether the value is within some range to be called an "event" (or not). More specific details of each of these groupings are given below.

The categories listed here follow the definitions of Murphy (1991) as expanded by Kubo et al. (2017) and Kubo (2019). For additional descriptions of the meaning of the metrics categories, see Chapter 8 of Wilks (2019). Potts (2021) is a thorough discussion of categories of metrics, including definitions for some, like the skill score. Murphy (1991) paper on weather forecast verification provides a good discussion of why multiple metrics are needed for robust assessments and discusses these metrics categories. It is good to take an expansive approach to data-model comparisons. Each metric only assesses a single aspect of the relationship, and therefore it is useful to use many metrics that tackle different perspectives. Recent reviews on this topic (e.g., Delzanno & Borovsky, 2022; Zheng et al., 2024) advocate for this system-wide approach to model assessments. This has also been the case for the review articles written by the International Space Weather Action Teams (e.g., Liemohn et al., 2018; Robinson et al., 2019; Welling et al., 2018; Zheng et al., 2019), as well as from terrestrial weather model assessments (e.g., Hoffman et al., 2017).

While there are several ways to cluster the wide array of available metrics, here we list five major categories: Accuracy, assessing the overall closeness of the model values to the data set values; Bias, examining the similarity in the centers of the two number sets; Precision, providing information on the likeness of the spread of the two number sets around their centers; Association, including measures of how well the model reproduces the timing of the up-and-down trends within the data; and Extremes, comparing the outliers of the number sets. There are two categories based on subsets of the full data-model paired set: Discrimination metrics, which are those that only use a portion of the full set as defined by a limited range of observational values; and Reliability metrics, which use only a portion of the numbers as defined by a limited range of modeled values. A final category is Skill, in which the metric score of the new model is compared to that of a reference model (for which there are some standard choices, but also could be any previous model or even guess at the data values).

A first step to any data-model comparison should be to make a set of plots—histograms, scatter plots, line plots, heat maps, quantile-quantile plots, or cumulative probability distributions. A nice discussion of introductory data visualization and data-model comparisons is given in the review by Kleiner and Graedel (1980). *Forecast Verification* (Jolliffe & Stephenson, 2011a) has useful information on data-model comparison techniques, regarding both visualization and metrics. The opening chapter (Jolliffe & Stephenson, 2011b) has a useful description of model goodness of fit and how to qualitatively assess it. The later chapters present quantitative aspects of data-model comparison. Liemohn (2023) also has a lengthy description of data visualization techniques and best practices for designing a good data-model comparison. The two subsections of this appendix provide a summary both fit performance and event detection metrics. Please see the cited works for additional details and usage advice.

A1. Fit Performance Metrics

Table A1 lists many common continuous metrics used in space physics data-model comparisons. In these formulas, the variables are as follows: M is the model output number set; O is the observed value number set; N is the total number of data-model pairs; M_i is an individual model output value (from the ith data-model pair); O_i is an individual observed value; σ is the standard deviation (of either the modeled or observed number set); γ is the skewness coefficient (of either the modeled or observed number set); k is the kurtosis coefficient (of either the modeled or observed number set); and d is the degrees of freedom in the model. This last term is the number of

Table A1
Summary of the Fit Performance Metrics

Metric	Formula	Range	Perfect score	Name and notes
Accuracy Metrics				
MSE	$\frac{1}{N-d}\sum_{i=1}^{N}(M_i - O_i)^2$	[0,∞)	0	Mean square error; only good for low variability number sets (LNVS); comparable to variances
RMSE	$\sqrt{\frac{1}{N-d}\sum_{i=1}^{N}(M_i - O_i)^2}$	[0,∞)	0	Root mean square error; only good for LVNS; comparable to standard deviations
MAE	$\frac{1}{N-d}\sum_{i=1}^{N}\lvert M_i - O_i\rvert$	[0,∞)	0	Mean absolute error; only good for LVNS; comparable to mean absolute differences
MAPE	$100\,\frac{1}{N-d}\sum_{i=1}^{N}\left\lvert\frac{M_i - O_i}{O_i}\right\rvert$	[0,∞)	0	Mean absolute percentage error; good for LVNS or HVNS but explodes if any O_i is close to zero; comparable to mean absolute differences converted to a percentage error
SMAPE	$100\,\frac{1}{N-d}\sum_{i=1}^{N}\left\lvert\frac{M_i - O_i}{(M_i + O_i)/2}\right\rvert$	[0,∞)	0	Symmetric mean absolute percentage error; good for LVNS or HVNS but not for number sets that straddle zero; comparable to mean absolute differences converted to perc. errors
MSA	$100\left(\exp\left[\text{Median}\left(\left\lvert\ln\left(\frac{M_i}{O_i}\right)\right\rvert\right)\right] - 1\right)$	[0,∞)	0	Median symmetric accuracy; good for LVNS or HVNS; only for positive-definite number sets; not influenced by outliers
Bias Metrics				
ME	$\frac{1}{N}\sum_{i=1}^{N}(M_i - O_i)$	(−∞,∞)	0	Mean error; only good for LVNS; comparable to mean absolute differences or standard deviations
ME_{med}	$\text{Median}(M_i) - \text{Median}(O_i)$	(−∞,∞)	0	Median error; good for LVNS or HVNS; not influenced by outliers; comparable to IQRs
ME_{geo}	$\exp\left(\frac{1}{N}\sum_{i=1}^{N}(\ln(M_i))\right) - \exp\left(\frac{1}{N}\sum_{i=1}^{N}(\ln(O_i))\right)$	(−∞,∞)	0	Geometric mean error; good for HVNS; comparable to IQRs calculated on log values
SSPB	$100\left(\text{sign}\left[\text{Median}\left(\ln\left(\frac{M_i}{O_i}\right)\right)\right]\right)\cdot\left(\exp\left[\left\lvert\text{Median}\left(\ln\left(\frac{M_i}{O_i}\right)\right)\right\rvert\right] - 1\right)$	(−∞,∞)	0	Signed symmetric percentage bias; good for LVNS or HVNS but not for number sets that straddle zero; comparable to the MSA metric

Table A1
Continued

Metric	Formula	Range	Perfect score	Name and notes
Precision Metrics				
YI	$\frac{\max(M)-\min(M)}{\max(O)-\min(O)}$	[0,∞)	1	Yield (or modeling yield); only good for LVNS; should be "close to 1"
$\mathrm{YI}_{\log}$	$\frac{\log[\max(M)]-\log[\min(M)]}{\log[\max(O)]-\log[\min(O)]}$	[0,∞)	1	Log yield; good for LVNS or HVNS but not for not for number sets straddling zero; should be "close to 1"
$P_{\sigma,\mathrm{ratio}}$	σ_M/σ_O	[0,∞)	1	Precision ratio; should be "close to 1"
$P_{\sigma,\mathrm{diff}}$	$\sigma_M - \sigma_O$	(−∞,∞)	0	Precision difference; comparable to standard deviations
Association Metrics				
R	$\frac{\sum(O_i-\overline{O})(M_i-\overline{M})}{\sqrt{\sum(O_i-\overline{O})^2\sum(M_i-\overline{M})^2}}$	[−1,1]	1	Pearson correlation coefficient; only good for LVNS; should not only be statistically significant but also be "good"
R^2	$\frac{\left[\sum(O_i-\overline{O})(M_i-\overline{M})\right]^2}{\sum(O_i-\overline{O})^2\sum(M_i-\overline{M})^2}$	[0,1]	1	Coefficient of determination; only good for LVNS
R_S	$\frac{\sum\left(\mathrm{rank}(O_i)-\overline{\mathrm{rank}}\right)\left(\mathrm{rank}(M_i)-\overline{\mathrm{rank}}\right)}{\sqrt{\sum\left(\mathrm{rank}(O_i)-\overline{\mathrm{rank}}\right)^2\sum\left(\mathrm{rank}(M_i)-\overline{\mathrm{rank}}\right)^2}}$	[−1,1]	1	Spearman rank-order correlation coefficient (also ROCC); good for LVNS or HVNS; not influenced by outliers; evaluated like R
$R_{\log}$	$\frac{\sum\left(\log\,O_i-\overline{\log\,O}\right)\left(\log\,M_i-\overline{\log\,M}\right)}{\sqrt{\sum\left(\log\,O_i-\overline{\log\,O}\right)^2\sum\left(\log\,M_i-\overline{\log\,M}\right)^2}}$	[−1,1]	1	Log correlation coefficient; good for LVNS or HVNS but only positive-definite number sets; evaluated like R
Extremes Metrics				
$\mathrm{LEDE}_\varepsilon$	$M_\varepsilon - O_\varepsilon$	(−∞,∞)	0	Low-end distribution error; ε is the quantile at which the value is extracted from each number set; usually ε is [0.01, 0.1]; good for LVNS and HVNS
$\mathrm{HEDE}_\varepsilon$	$M_{1-\varepsilon} - O_{1-\varepsilon}$	(−∞,∞)	0	High-end distribution error; "1−ε" is the quantile at which the value is extracted from each number set; good for LVNS and HVNS
γ_Δ	$\gamma_M - \gamma_O$	(−∞,∞)	0	Skew difference; only good for LVNS; measures goodness of tail directionality; comparable to skew coefficients
k_Δ	$k_M - k_O$	(−∞,∞)	0	Kurtosis difference; only good for LVNS; measures heaviness of distribution tails; comparable to kurtosis coefficients
Skill Metrics				
PE	$1-\frac{\sum(M_i-O_i)^2}{\sum(O_i-\overline{O})^2}$	(−∞,1]	1	Prediction efficiency; only good for LVNS; anything higher than 0 is "better than the observational average"
$\mathrm{SS}_{\mathrm{MSE}}$	$1-\frac{\sum(M_i-O_i)^2}{\sum(M_i^{old}-O_i)^2}$	(−∞,1]	1	MSE-based skill score like PE but compared against previous model; only good for LVNS; anything higher than 0 indicates that the new model is better
SS	$\frac{\mathrm{Score}_{\mathrm{New\ \ Model}}-\mathrm{Score}_{\mathrm{Ref\ \ Model}}}{\mathrm{Score}_{\mathrm{Perfect}}-\mathrm{Score}_{\mathrm{Ref\ \ Model}}}$	(−∞,1]	1	Generic skill score formula; may be used with any metric from the categories above; anything higher than 0 indicates that the new model is better than the reference model

data-based values or free parameters used in the definition of the model. While this could be difficult to accurately define, especially for complicated, coupled models, it can often be ignored when $N >> d$.

There are several good sources for more information on these metrics. Armstrong (1985) includes several chapters devoted to fit performance metrics, including a good discussion of when to use certain metrics based on the variability of the number sets being inspected. Another good source for a discussion of different fit performance metrics, with a space weather focus on data sets that span multiple orders of magnitude, is given by Morley, Brito, and Welling (2018). Murphy (1988) offers a robust introduction to prediction efficiency and other fit performance skill scores. The book by Wilks (2019) contains discussion of a few of these metrics, and the book by Jolliffe and Stephenson (2011a) also covers aspects of fit performance metrics; of particular use for fit performance metrics is Chapter 5 by Déqué (2011). A couple of other excellent reviews on fit performance metrics are Goodwin and Lawton (1999) and Hyndman and Koehler (2006).

Although never used in a GEM challenge thus far, another approach to this type of assessment is probabilistic forecasting, in which a range of model outcomes are compared against an observed outcome. Smith et al. (2020) use the Brier skill score for assessing storm sudden commencements, and Meredith et al. (2023) similarly used probabilistic forecasting for extreme cases of radiation belt fluxes. A related approach is defining the characteristics of a "once in a decade" event (or some other, often longer, time interval), as has been done by Love (2012), Love et al. (2015), Riley (2018), and Hapgood et al. (2021). A related technique is uncertainty quantification of numerical models, such as the study by Morley, Welling, and Woodroffe (2018), who conducted an ensemble of geospace simulations by varying the upstream conditions and assessing the variability of the inner magnetospheric predictions.

Note that Table A1 does not contain any listings for metrics in the discrimination and reliability categories. That is because there are no unique fit performance metrics for these categories; any of the listed metrics can be used on a portion of the full paired number set; as subset by a range of observed values (discrimination) or modeled values (reliability). Accuracy metrics are the most common to use with these methods, but using a variety of metrics from different categories is useful for a robust assessment. Calculating metrics based on such subsetting is encouraged because it could be highly informative: a discrimination procedure reveals the portions of the data range for which the model works well, while a reliability test reveals the portions of the model range for which the model works well.

A piece of information listed in the notes column of Table A1 is the type of number set for which that metric works well. There are some metrics that are only good with number sets that span only one or two orders of magnitude. We will call such number sets low variability number sets, or LVNS, in the table. Other metrics are good for number sets spanning two or more orders of magnitude, what we call high variability number sets, or HVNS. If either the observed values or the model output is an HVNS, then the metrics that only work for LVNS should be used with caution, if at all.

Also listed in the notes column of Table A1 are suggestions about how to assess the goodness of the metric score. When a number set parameter is suggested, like standard deviation, it should be remembered that the parameters for both the observed and modeled number sets should be considered in this goodness assessment. For example, for mean square error (MSE), it might look really good against the observed variance, but could be bad against the model variance. Also, keep in mind that all of these goodness guidelines are merely possibilities and should not be considered definitive cutoffs.

Finally, the generic skill score formula is given at the end of the Skill list of metrics (labeled as SS). The placeholder "Score" within its formula can be any of the metrics listed earlier in the table and that "ref model" can be any previous model to which you want to compare.

A2. Event Detection Metrics

Table A2 lists common dichotomous metrics used across the disciplines within space physics. These are all based on a conversion of the exact values into yes-or-no binary values. Two event identification thresholds must be defined—one of the observed number set and one for the modeled number set—to which the values are compared. These two thresholds are often the same but do not have to be identical. From each data-model pairing, four options are possible based on the event status of each value, and the counts of these options populate the cells within the contingency table (sometimes called a confusion matrix). While several naming conventions exist for the cells of a contingency table, we adopt this one: Hits (H) are when both the data and the model values are in event state (yes-yes); Misses (M) are when the data value is in event state but the model value is not (yes-no); False Alarms (F) are when the model value is in event state and the data value is not (no-yes); and Correct Negatives (C) are when both the data and model values are not in event state (no-no). Note that because the exact values are only used to define the cell counts in the contingency table and then not used again, all of the metrics

Table A2
Summary of the Event Detection Metrics

Metric	Formula	Range	Perfect score	Name and notes
Accuracy Metrics				
PC	$(H + C)/N$	[0, 1]	1	Proportion correct; includes correct negatives in both numerator and denominator
CSI	$H/(H + M + F)$	[0, 1]	1	Critical success index; like PC but removes correct negatives from the equation
F_1	$2H/(2H + M + F)$	[0, 1]	1	$F - 1$ score
Bias Metrics				
FB	$(H + F)/(H + M)$	[0, ∞)	1	Frequency bias; modeled events divided by observed events
Precision Metrics				
Precision	$H/(H + F)$	[0, 1]	1	Precision; equivalent to the reliability metric positive predictive value (below)
Recall	$H/(H + M)$	[0, 1]	1	Recall; equivalent to the discrimination metric probability of detection (below)
Association Metrics				
θ	$(H \cdot C)/(M \cdot F)$	[0, ∞)	1	Odds ratio; is 0 if either H or C is 0, is undefined if either F or M is 0
ORSS	$\frac{(H \cdot C) - (M \cdot F)}{(H \cdot C) + (M \cdot F)}$	[−1, 1]	1	Odds ratio skill score; defined as $(\theta - 1)/(\theta + 1)$ to convert θ to a skill-score-like range
MCC	$\frac{(H \cdot C) - (F \cdot M)}{\sqrt{(H+F)(H+M)(F+C)(M+C)}}$	[1–, 1]	1	Matthews correlation coefficient
Extremes Metrics				
EDS	$\frac{2 \cdot \ln(N) - 2 \cdot \ln(H+M)}{\ln(N) - \ln(H)} - 1$	[−1, 1]	1	Extreme dependency score; focused on a comparison of H relative to observed events
SEDS	$\frac{2 \cdot \ln(N) - \ln(H+M) - \ln(H+F)}{\ln(N) - \ln(H)} - 1$	[−1, 1]	1	Symmetric extreme dependency score; symmetric in its usage of observed events and modeled events
Discrimination Metrics				
POD	$H/(H + M)$	[0, 1]	1	Probability of detection; focused on observed events subset
POFD	$F/(F + C)$	[0, 1]	0	Probability of false detection; focused on observed non-events subset
FNR	$M/(H + M)$	[0, 1]	0	False negative rate; focused on observed events subset
Specificity	$C/(F + C)$	[0, 1]	1	Specificity; focused on observed non-events subset
Reliability Metrics				
PPV	$H/(H + F)$	[0, 1]	1	Positive predictive value; focused on modeled events subset
MR	$M/(M + C)$	[0, 1]	0	Miss ratio; focused on modeled non-events subset
FAR	$F/(H + F)$	[0, 1]	0	False alarm ratio; focused on modeled events subset
NPV	$C/(M + C)$	[0, 1]	1	Negative predictive value; focused on modeled non-events subset
FR	H/F	[0, ∞)	1	Forecast ratio; focused on modeled events subset
Skill Metrics				
HSS	$\frac{2[(H \cdot C) - (F \cdot M)]}{(H+M)(M+C) + (H+F)(F+C)}$	[−1, 1]	1	Heidke skill score; symmetric use of table cells; please do not use anymore (see Liemohn et al., 2025)
PSS	$\frac{(H \cdot C) - (M \cdot F)}{(H+M)(F+C)}$	[−1, 1]	1	Peirce skill score; discrimination-focused skill; please do not use anymore
CSS	$\frac{(H \cdot C) - (M \cdot F)}{(H+F)(M+C)}$	[−1, 1]	1	Clayton skill score; reliability-focused skill; please do not use anymore
GSS	$\frac{H - H_{\mathrm{ref}}}{H + F + M - H_{\mathrm{ref}}}$	[−0.33, 1]	1	Gilbert skill score; $H_{\mathrm{ref}} = (H + F)\,(H + M)/N$; deemphasizes correct negatives, which only appears in the denominator of H_{ref}; please do not use anymore
SS	$\frac{\mathrm{Score}_{\mathrm{New\ Model}} - \mathrm{Score}_{\mathrm{Ref\ Model}}}{\mathrm{Score}_{\mathrm{Perfect}} - \mathrm{Score}_{\mathrm{Ref\ Model}}}$	(−∞,1]	1	Generic skill score formula; may be used with any metric from the categories above; anything higher than 0 indicates that the new model is better than the reference model

listed in Table A2 work well regardless of the variability of either number set (i.e., for both LVNS and HVNS), as well as for number sets straddling zero.

Binary event metrics are thoroughly discussed by Hogen and Mason (2011). Wilks (2019) also has extensive discussion of nearly every metric mentioned below. Kubo (2019) calls out a few "best choice" event detection metrics for the category definitions used here. Matthews (1975) popularized the contingency-table-based correlation coefficient (labeled MCC below). A nice description and application of the extreme dependency score for atmospheric science is given in Stephenson et al. (2008).

Although not used previously for GEM challenges, multilevel event detection methods are often used in terrestrial weather forecasting. Further information on this can be found in the books by Jolliffe and Stephenson (2011a) and Wilks (2019). An example of a space weather usage of multilevel event detection analysis is that of Kahler and Darsey (2021), who used it for assessing a solar energetic particle prediction model.

For event detection, metrics for the discrimination and reliability categories are easily defined because there is a natural split of the number sets at the event definition threshold. Therefore, these two categories have metrics of their own that only use two of the four cells of the contingency table.

In Table A2, most skill scores are listed in the final category block (Skill), but some are listed elsewhere in the table if they measure some facet of the data-model comparison within that more specific category. For example, the odds ratio skill score is listed in Association because it is defined from the odds ratio, θ, as $(\theta - 1)/(\theta + 1)$. For most (but not all) of these predefined skill scores in the event detection grouping, the "reference model" to which the new model is compared is the "random reshuffling of events" matrix defined by Gilbert (1884), for which the events in both the observed and modeled number sets are rearranged along the timeline. As argued by Liemohn et al. (2025), the Heidke skill score, Peirce skill score (also known as the True Skill Statistic), the Clayton skill score, and the Gilbert skill score (also known at the Equitable Threat Score) are flawed and should not be used. They are listed below for completeness (because they have been widely used in GEM challenges), but Liemohn et al. (2025) present new skill scores with independent reference models that should be used instead.

Appendix B: Additional Resources for GEM Challenges

Here is a list of websites with additional resources that could be of value to GEM challenge organizers and participants.

GEM-specific pages:

- GEM Program main webpage: https://gem.epss.ucla.edu/mediawiki/index.php/Main_Page.
- Focus Group listing, including links to their original proposals: https://gem.epss.ucla.edu/mediawiki/index.php/GEM_Focus_Groups.
- GEM M&V Resource Group website: https://gem.epss.ucla.edu/mediawiki/index.php/RG:_Modeling_Methods_and_Validation.
- GEMStone newsletters: https://gem.epss.ucla.edu/mediawiki/index.php/GEMstone_Newsletter.

Modeling and data-model comparison tools:

- CCMC Model Catalog https://ccmc.gsfc.nasa.gov/models/?statuses=Production&statuses=Result+Only.
- CAMEL (Comprehensive Assessment of Models and Events using Library Tools (CAMEL) Framework (following open science practice: validation codes and data are open to the community), including a list of validation campaigns using this toolkit https://ccmc.gsfc.nasa.gov/tools/CAMEL/.
- Kamodo https://ccmc.gsfc.nasa.gov/tools/kamodo/.
- PyHC tools https://heliopython.org/.
- Application Usability Levels https://www.swsc-journal.org/articles/swsc/full_html/2019/01/swsc190028/swsc190028.html.

Open science resources:

- The GO FAIR website: https://www.go-fair.org/fair-principles/.
- The AGU site on FAIR principles and data leadership: https://www.agu.org/learn-about-agu/about-agu/data-leadership.
- The AGU news and posts on open data and software: https://www.agu.org/Channel/Open-and-FAIR-Data-and-Software.
- Committee on Publication Ethics: https://onlinelibrary.wiley.com/.

Team-building resources:

- NAS team science best practices https://nap.nationalacademies.org/catalog/19007/enhancing-the-effectiveness-of-team-science.
- Write-up of a community challenge from space apps: https://science.nasa.gov/science-news/2022-space-apps-challenge.

Codes of conduct:

- Geospace Environment Modeling:
- GEM Workshop: https://gemworkshop.org/gem-code-of-conduct/.
- American Geophysical Union: https://www.agu.org/plan-for-a-meeting/agumeetings/meetings-resources/meetings-code-of-conduct.
- Center for Geospace Storms: https://cgs.jhuapl.edu/News-and-Events/Code%20of%20Conduct.pdf.

Data Availability Statement

No new data was used for this study. The background image for Figure 2 was taken from the model output movie available at the NASA Visualization Studio, https://svs.gsfc.nasa.gov/5193/, and was produced by A.J. Christensen and Slava Merkin. More about the MAGE model can be found here: https://cgs.jhuapl.edu/Models/mage.php.

Acknowledgments
The M&V Resource Group leadership would like to thank NSF, the GEM Program, and the entire GEM research community for their support and engagement with this compilation of challenge best practices. We thank all of those that have participated in the GEM workshop sessions discussing community challenges, as well as those that have led or participated in any of the many GEM challenges. The authors would like to especially thank Howard J. Singer at the NOAA Space Weather Prediction Center and E. Joshua Rigler at the US Geological Survey for their contributions to this project. While this paper coalesces our current understanding of how to run an effective challenge, we expect this to be a living document on the GEM program website (https://gem.epss.ucla.edu/mediawiki/index.php/Main_Page) at the M&V RG page (https://gem.epss.ucla.edu/mediawiki/index.php/RG:_Modeling_Methods_and_Validation), allowing for additional suggestions and continual updates. Any use of trade, firm, or product names in this article is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

Albert, J. M., Selesnick, R. S., Morley, S. K., Henderson, M. G., & Kellerman, A. C. (2018). Calculation of last closed drift shells for the 2013 GEM radiation belt challenge events. *Journal of Geophysical Research: Space Physics*, *123*(11), 9597–9611. https://doi.org/10.1029/2018JA025991

Anonymous. (1990). GEM launches funded campaign. *Eos, Transactions American Geophysical Union*, *71*(33), 1027–1034. https://doi.org/10.1029/EO071i033p01027-03

Armstrong, S. J. (1985). *Long range forecasting* (2nd ed.). Wiley. Retrieved from https://repository.upenn.edu/marketing_papers/211

Birn, J., Drake, J. F., Shay, M. A., Rogers, B. N., Denton, R. E., Hesse, M., et al. (2001). Geospace Environmental Modeling (GEM) magnetic reconnection challenge. *Journal of Geophysical Research*, *106*(A3), 3715–3719. https://doi.org/10.1029/1999JA900449

Birn, J., Galsgaard, K., Hesse, M., Hoshino, M., Huba, J., Lapenta, G., et al. (2005). Forced magnetic reconnection. *Geophysical Research Letters*, *32*(6), L06105. https://doi.org/10.1029/2004GL022058

Birn, J., & Hesse, M. (2001). Geospace Environment Modeling (GEM) magnetic reconnection challenge: Resistive tearing, anisotropic pressure and Hall effects. *Journal of Geophysical Research*, *106*(A3), 3737–3750. https://doi.org/10.1029/1999JA001001

Brito, T. V., & Morley, S. K. (2017). Improving empirical magnetic field models by fitting to in situ data using an optimized parameter approach. *Space Weather*, *15*(12), 1628–1648. https://doi.org/10.1002/2017SW001702

Chen, L., Jordanova, V. K., Spasojević, M., Thorne, R. M., & Horne, R. B. (2014). Electromagnetic ion cyclotron wave modeling during the geospace environment modeling challenge event. *Journal of Geophysical Research: Space Physics*, *119*(4), 2963–2977. https://doi.org/10.1002/2013JA019595

Chen, Y., Friedel, R. H. W., & Reeves, G. D. (2006). Phase space density distributions of energetic electrons in the outer radiation belt during two geospace environment modeling inner magnetosphere/storms selected storms. *Journal of Geophysical Research*, *111*(A11), A11S04. https://doi.org/10.1029/2006JA011703

Chen, Y., Tóth, G., Hietala, H., Vines, S. K., Zou, Y., Nishimura, Y. T., et al. (2020). Magnetohydrodynamic with embedded particle-in-cell simulation of the geospace environment modeling dayside kinetic processes challenge event. *Earth and Space Science*, *7*(11), e2020EA001331. https://doi.org/10.1029/2020EA001331

Collado-Vega, Y. M., Dredger, P., Lopez, R. E., Khurana, S., Rastaetter, L., Sibeck, D., & Anastopulos, M. (2023). Magnetopause standoff position changes and geosynchronous orbit crossings: Models and observations. *Space Weather*, *21*(6), e2022SW003212. https://doi.org/10.1029/2022SW003212

Delzanno, G. L., & Borovsky, J. E. (2022). The need for a system science approach to global magnetospheric models. *Frontiers in Astronomy and Space Sciences*, *9*, 9808629. https://doi.org/10.3389/fspas.2022.808629

Déqué, M. (2011). Deterministic forecasts of continuous variables. In I. T. Jolliffe & D. B. Stephenson (Eds.), *Forecast Verification*. https://doi.org/10.1002/9781119960003.ch5

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 and Space Science*, *7*(6), e2020EA001095. https://doi.org/10.1029/2020EA001095

Dusenbery, P. B., & Siscoe, G. L. (1992). Geospace environment modeling program. *Eos, Transactions American Geophysical Union*, *73*(7), 83–84. https://doi.org/10.1029/91EO00065

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. *Journal of Geophysical Research: Space Physics*, *118*(12), 7752–7761. https://doi.org/10.1002/2013JA019322

Engebretson, M. J., Posch, J. L., Braun, D. J., Li, W., Ma, Q., Kellerman, A. C., et al. (2018). EMIC wave events during the four GEM QARBM challenge intervals. *Journal of Geophysical Research: Space Physics*, *123*(8), 6394–6423. https://doi.org/10.1029/2018JA025505

Fedder, J. A., Slinker, S. P., & Lyon, J. G. (1998). A comparison of global numerical simulation results to data for the January 27–28, 1992, Geospace Environment Modeling challenge event. *Journal of Geophysical Research*, *103*(A7), 14799–14810. https://doi.org/10.1029/97JA03664

Ganushkina, N. Y., Pulkkinen, T. I., Milillo, A., & Liemohn, M. W. (2006). Evolution of the proton ring current energy distribution during 21–25 April 2001 storm. *Journal of Geophysical Research*, *111*(A11), A11S08. https://doi.org/10.1029/2006JA011609

García, K. S., & Hughes, W. J. (2007). Finding the Lyon-Fedder-Mobarry magnetopause: A statistical perspective. *Journal of Geophysical Research*, *112*(A6), A06229. https://doi.org/10.1029/2006JA012039

Gilbert, G. K. (1884). Finley's tornado predictions. *American Meteorological Journal*, *1*, 166–172.

Glocer, A., Rastätter, L., Kuznetsova, M., Pulkkinen, A., Singer, H. J., Balch, C., et al. (2016). Community-wide validation of geospace model local K-index predictions to support model transition to operations. *Space Weather*, *14*(7), 469–480. https://doi.org/10.1002/2016SW001387

Goldstein, J., Sandel, B. R., Forrester, W. T., Thomsen, M. F., & Hairston, M. R. (2005). Global plasmasphere evolution 22–23 April 2001. *Journal of Geophysical Research*, *110*(A12), A12218. https://doi.org/10.1029/2005JA011282

Goodwin, P., & Lawton, R. (1999). On the asymmetry of the symmetric MAPE. *International Journal of Forecasting*, *15*, 405–408. https://doi.org/10.1016/S0169-2070(99)00007-2

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

Guild, T. B., Spence, H. E., Kepko, E. L., Merkin, V., Lyon, J. G., Wiltberger, M., & Goodrich, C. C. (2008). Geotail and LFM comparisons of plasma sheet climatology: 1. Average values. *Journal of Geophysical Research*, *113*(A4), A04216. https://doi.org/10.1029/2007JA012611

Guo, Z., Lin, Y., Wang, X., Vines, S. K., Lee, S. H., & Chen, Y. (2020). Magnetopause reconnection as influenced by the dipole tilt under southward IMF conditions: Hybrid simulation and MMS observation. *Journal of Geophysical Research: Space Physics*, *125*(9), e2020JA027795. https://doi.org/10.1029/2020JA027795

Guterl, F. (2014). Diversity in science: Why it is essential for excellence. *Scientific American*, *311*(4), 38–40. https://doi.org/10.1038/scientificamerican1014-38

Haiducek, J. D., Welling, D. T., Morley, S. K., Ganushkina, N. Y., & Chu, X. (2020). Using multiple signatures to improve accuracy of substorm identification. *Journal of Geophysical Research: Space Physics*, *125*(4), e2019JA027559. https://doi.org/10.1029/2019ja027559

Halford, A., Kellerman, A., Garcia-Sage, K., Klenzing, J., Carter, B., McGranaghan, R., et al. (2019). Application usability levels: A framework for tracking project product progress. *Journal of Space Weather and Space Climate*, *9*, A34. https://doi.org/10.1051/swsc/2019030

Hapgood, M. M. J. A., Attrill, G., Bisi, M., Cannon, P. S., Dyer, C., Eastwood, J. P., et al. (2021). Development of space weather reasonable worst-case scenarios for the UK national risk assessment. *Space Weather*, *19*, 4. https://doi.org/10.1029/2020SW002593

Hesse, M., Birn, J., & Kuznetsova, M. (2001). Collisionless magnetic reconnection: Electron processes and transport modeling. *Journal of Geophysical Research*, *106*(A3), 3721–3735. https://doi.org/10.1029/1999JA001002

Hietala, H., Dimmock, A. P., Zou, Y., & Garcia-Sage, K. (2020). The challenges and rewards of running a geospace environment modeling challenge. *Journal of Geophysical Research: Space Physics*, *125*(3), e2019JA027642. https://doi.org/10.1029/2019JA027642

Hill, T. W., & Toffoletto, F. R. (1998). Comparison of empirical and theoretical polar cap convection patterns for the January 1992 GEM interval. *Journal of Geophysical Research*, *103*(A7), 14811–14817. https://doi.org/10.1029/97JA03525

Hoffman, R. N., Boukabara, S.-A., Kumar, V. K., Garrett, K., Casey, S. P. F., & Atlas, R. (2017). An empirical cumulative density function approach to defining summary NWP forecast assessment metrics. *Monthly Weather Review*, *145*(4), 1427–1435. https://doi.org/10.1175/MWR-D-16-0271.1

Hogen, R. J., & Mason, I. B. (2011). Deterministic forecasts of binary events. In I. T. Jolliffe & D. B. Stephenson (Eds.), *Forecast Verification: A Practitioner's Guide in Atmospheric Science* (2nd ed., pp. 31–60). John Wiley, Ltd. https://doi.org/10.1002/9781119960003.ch3

Honkonen, I., Rastätter, L., Grocott, A., Pulkkinen, A., Palmroth, M., Raeder, J., et al. (2013). On the performance of global magnetohydrodynamic models in the Earth's magnetosphere. *Space Weather*, *11*(5), 313–326. https://doi.org/10.1002/swe.20055

Huang, C.-L., Spence, H. E., Lyon, J. G., Toffoletto, F. R., Singer, H. J., & Sazykin, S. (2006). Storm-time configuration of the inner magnetosphere: Lyon-Fedder-Mobarry MHD code, Tsyganenko model, and GOES observations. *Journal of Geophysical Research*, *111*(A11), A11S16. https://doi.org/10.1029/2006JA011626

Hunt, V., Lee, L., Prince, S., & Dixon-Fyle, S. (2018). *Delivering through diversity* (Technical Report). McKinsey and Company. Retrieved from https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights/delivering-through-diversity

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–699. https://doi.org/10.1016/j.ijforecast.2006.03.001

Jolliffe, I. T., & Stephenson, D. B. (2011a). *Forecast verification: A practioner's guide in atmospheric science* (2nd ed.). John Wiley and Sons. https://doi.org/10.1002/9781119960003

Jolliffe, I. T., & Stephenson, D. B. (2011b). Introduction. In I. T. Jolliffe & D. B. Stephenson (Eds.), *Forecast Verification*. https://doi.org/10.1002/9781119960003.ch1

Jordanova, V. K., Miyoshi, Y. S., Zaharia, S., Thomsen, M. F., Reeves, G. D., Evans, D. S., et al. (2006). Kinetic simulations of ring current evolution during the geospace environment modeling challenge events. *Journal of Geophysical Research*, *111*(A11), A11S10. https://doi.org/10.1029/2006JA011644

Kahler, S. W., & Darsey, H. (2021). Exploring contingency skill scores based on event sizes. *Space Weather*, *19*(5), e2020SW002604. https://doi.org/10.1029/2020SW002604

Kamaletdinov, S. R., Artemyev, A. V., Runov, A., & Angelopoulos, V. (2024). Thin current sheets in the magnetotail at lunar distances: Statistics of ARTEMIS observations. *Journal of Geophysical Research: Space Physics*, *129*(3), e2023JA032130. https://doi.org/10.1029/2023JA032130

Katus, R. M., Liemohn, M. W., Ionides, E., Ilie, R., Welling, D. T., & Sarno-Smith, L. K. (2015). Statistical analysis of the geomagnetic response to different solar wind drivers and the dependence on storm intensity. *Journal of Geophysical Research: Space Physics*, *120*(1), 310–327. https://doi.org/10.1002/2014JA020712

Khazanov, G. V., Chen, M. W., Lemon, C. L., & Sibeck, D. G. (2019). The magnetosphere-ionosphere electron precipitation dynamics and their geospace consequences during the 17 March 2013 storm. *Journal of Geophysical Research: Space Physics*, *124*(8), 6504–6523. https://doi.org/10.1029/2019JA026589

Kleiner, B., & Graedel, T. E. (1980). Exploratory data analysis in the geophysical sciences. *Reviews of Geophysics*, *18*(3), 699–717. https://doi.org/10.1029/RG018i003p00699

Kozyra, J. U., & Liemohn, M. W. (2003). Ring current energy input and decay. *Space Science Reviews*, *109*(1–4), 105–131. https://doi.org/10.1023/b:spac.0000007516.10433.ad

Kubo, Y. (2019). Why do some probabilistic forecasts lack reliability? *Journal of Space Weather & Space Climate*, *9*, A17. https://doi.org/10.1051/swsc/2019016

Kubo, Y., Den, M., & Ishii, M. (2017). Verification of operational solar flare fore cast: Case of regional warning center Japan. *Journal of Space Weather and Space Climate*, *7*, A20. https://doi.org/10.1051/swsc/2017018

Kuznetsova, M. M., Hesse, M., & Winske, D. (2001). Collisionless reconnection supported by nongyrotropic pressure effects in hybrid and particle simulations. *Journal of Geophysical Research*, *106*(A3), 3799–3810. https://doi.org/10.1029/1999JA001003

Lerback, J. C., Hanson, B., & Wooden, P. (2020). Association between author diversity and acceptance rates and citations in peer-reviewed Earth science manuscripts. *Earth and Space Science*, *7*(5), e2019EA000946. https://doi.org/10.1029/2019ea000946

Liemohn, M. W. (2006). Introduction to special section on "results of the national science foundation geospace environment modeling inner magnetosphere/storms assessment challenge". *Journal of Geophysical Research*, *111*(A11), A11S01. https://doi.org/10.1029/2006JA011970

Liemohn, M. W. (2023). *Data Analysis for the geosciences: Essentials of uncertainty, comparison, and visualization*. John Wiley and Sons. Textbook Series. ISBN: 978-1-119-74787-1 (paperback), 978-1-119-74789-5 (e-pub).

Liemohn, M. W., Ganushkina, N. Y., Welling, D. T., & Azari, A. R. (2025). Defining an independent reference model for event detection skill scores. *AGU Advances*. Submitted to 25 February 2025, manuscript # 2025AV001710. Available on ESSOar here:. https://doi.org/10.22541/essoar.174129298.81850235/v1

Liemohn, M. W., & Jazowski, M. (2008). Ring current simulations of the 90 intense storms during solar cycle 23. *Journal of Geophysical Research*, *113*(A3), A00A17. https://doi.org/10.1029/2008JA013466

Liemohn, M. W., Kozyra, J. U., Jordanova, V. K., Khazanov, G. V., Thomsen, M. F., & Cayton, T. E. (1999). Analysis of early phase ring current recovery mechanisms during geomagnetic storms. *Geophysical Research Letters*, *25*(18), 2845–2848. https://doi.org/10.1029/1999gl900611

Liemohn, M. W., Kozyra, J. U., Thomsen, M. F., Roeder, J. L., Lu, G., Borovsky, J. E., & Cayton, T. E. (2001). Dominant role of the asymmetric ring current in producing the stormtime Dst*. *Journal of Geophysical Research*, *106*(A6), 10883–10904. https://doi.org/10.1029/2000ja000326

Liemohn, M. W., McCollough, J. P., Jordanova, V. K., Ngwira, C. M., Morley, S. K., Cid, C., et al. (2018). Model evaluation guidelines for geomagnetic index predictions. *Space Weather*, *16*(12), 2079–2102. https://doi.org/10.1029/2018SW002067

Liemohn, M. W., Ridley, A. J., Kozyra, J. U., Gallagher, D. L., Thomsen, M. F., Henderson, M. G., et al. (2006). Analyzing electric field morphology through data-model comparisons of the geospace environment modeling inner magnetosphere/storm assessment challenge events. *Journal of Geophysical Research*, *111*(A11), A11S11. https://doi.org/10.1029/2006JA011700

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. *Journal of Atmospheric and Solar-Terrestrial Physics*, *218*, 105624. https://doi.org/10.1016/j.jastp.2021.105624

Lin, D., Wang, W., Garcia-Sage, K., Yue, J., Merkin, V., McInerney, J. M., et al. (2022). Thermospheric neutral density variation during the "SpaceX" storm: Implications from physics-based whole geospace modeling. *Space Weather*, *20*(12), e2022SW003254. https://doi.org/10.1029/2022SW003254

Lorentzen, K. R., Cooper, M. D., & Blake, J. B. (2001). Relativistic electron microbursts during the GEM storms. *Geophysical Research Letters*, *28*(13), 2573–2576. https://doi.org/10.1029/2001GL012926

Lotko, W. (1993). Milestones in geospace environment modeling. *Eos, Transactions American Geophysical Union*, *74*(52), 618–622. https://doi.org/10.1029/93EO00565

Love, J. J. (2012). Credible occurrence probabilities for extreme geophysical events: Earthquakes, volcanic eruptions, magnetic storms. *Geophysical Research Letters*, *39*(10), L10301. https://doi.org/10.1029/2012GL051431

Love, J. J., Joshua Rigler, E., Pulkkinen, A., & Riley, P. (2015). On the lognormality of historical magnetic storm intensity statistics: Implications for extreme-event probabilities. *Geophysical Research Letters*, *42*(16), 6544–6553. https://doi.org/10.1002/2015GL064842

Lyons, L. R. (1995). The ionosphere as a screen for magnetospheric processes. *Reviews of Geophysics*, *33*(S1), 715–720. https://doi.org/10.1029/95RG00286

Lyons, L. R. (1998). The geospace modeling program grand challenge. *Journal of Geophysical Research*, *103*(A7), 14781–14785. https://doi.org/10.1029/98JA00015

Lyons, L. R., & de la Beaujardiére, O. (1994). Program probes the magnetosphere. *Eos, Transactions American Geophysical Union*, *75*(9), 97–109. https://doi.org/10.1029/94EO00722

Lyons, L. R., Lu, G., de la Beaujardière, O., & Rich, F. J. (1996). Synoptic maps of polar caps for stable interplanetary magnetic field intervals during January 1992 geospace environment modeling campaign. *Journal of Geophysical Research*, *101*(A12), 27283–27298. https://doi.org/10.1029/96JA02457

Lyons, L. R., McPherron, R. L., Zesta, E., Reeves, G. D., Sigwarth, J. B., & Frank, L. A. (2001). Timing of substorm signatures during the November 24, 1996, geospace environment modeling event. *Journal of Geophysical Research*, *106*(A1), 349–359. https://doi.org/10.1029/1999JA000601

Lyons, L. R., Ruohoniemi, J. M., & Lu, G. (2001). Substorm-associated changes in large-scale convection during the November 24, 1996, geospace environment modeling event. *Journal of Geophysical Research*, *106*(A1), 397–405. https://doi.org/10.1029/1999JA000602

Ma, Q., Li, W., Bortnik, J., Thorne, R. M., Chu, X., Ozeke, L. G., et al. (2018). Quantitative evaluation of radial diffusion and local acceleration processes during GEM challenge events. *Journal of Geophysical Research: Space Physics*, *123*(3), 1938–1952. https://doi.org/10.1002/2017JA025114

Ma, Z. W., & Bhattacharjee, A. (2001). Hall magnetohydrodynamic reconnection: The geospace environment modeling challenge. *Journal of Geophysical Research*, *106*(A3), 3773–3782. https://doi.org/10.1029/1999JA001004

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, *405*(2), 442–451. https://doi.org/10.1016/0005-2795(75)90109-9

McAdams, K. L., Reeves, G. D., Friedel, R. H. W., & Cayton, T. E. (2001). Multisatellite comparisons of the radiation belt response to the Geospace Environment Modeling (GEM) magnetic storms. *Journal of Geophysical Research*, *106*(A6), 10869–10882. https://doi.org/10.1029/2000JA000248

Medin, D., Lee, C. D., & Bang, M. (2014). Point of view affects how science is done. *Scientific American*, *1*.

Meredith, N. P., Cayton, T. E., Cayton, M. D., & Horne, R. B. (2023). Extreme relativistic electron fluxes in GPS orbit: Analysis of NS41 BDD-IIR data. *Space Weather*, *21*(6), e2023SW003436. https://doi.org/10.1029/2023SW003436

Milillo, A., Orsini, S., Massetti, S., & Mura, A. (2006). Geomagnetic activity dependence of the inner magnetospheric proton distribution: An empirical approach for the 21–25 April 2001 storm. *Journal of Geophysical Research*, *111*(A11), A11S13. https://doi.org/10.1029/2006JA011956

Miyoshi, Y. S., Jordanova, V. K., Morioka, A., Thomsen, M. F., Reeves, G. D., Evans, D. S., & Green, J. C. (2006). Observations and modeling of energetic electron dynamics during the October 2001 storm. *Journal of Geophysical Research*, *111*(A11), A11S02. https://doi.org/10.1029/2005JA011351

Moldwin, M. B., & Liemohn, M. W. (2018). High impact papers in space physics: Examination of gender, country and paper characteristics. *Journal of Geophysical Research: Space Physics*, *123*(4), 2557–2565. https://doi.org/10.1002/2018JA025291

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

Morley, S. K., Brito, T. V., & Welling, D. T. (2018). Measures of model performance based on the log accuracy ratio. *Space Weather*, *16*(1), 69–88. https://doi.org/10.1002/2017SW001669

Morley, S. K., Welling, D. T., & Woodroffe, J. R. (2018). Perturbed input ensemble modeling with the space weather modeling framework. *Space Weather*, *16*(9), 1330–1347. https://doi.org/10.1029/2018SW002000

Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, *116*(12), 2417–2424. https://doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2

Murphy, A. H. (1991). Forecast verification: Its complexity and dimensionality. *Monthly Weather Review*, *119*(7), 1590–1601. https://doi.org/10.1175/1520-0493(1991)119<1590:fvicad>2.0.co;2

National Research Council. (2015). Enhancing the effectiveness of team science. Committee on the science of team science. In N. J. Cooke & M. L. Hilton (Eds.), *Board on Behavioral, Cognitive, and Sensory Sciences, Division of Behavioral and Social Sciences and Education*. The National Academies Press. https://doi.org/10.17226/19007

NSF, FY 1988. (1987). *Global geosciences program, Publ. Dir. Geosci*. National Science Foundation.

O'Brien, T. P., Looper, M. D., & Blake, J. B. (2004). Quantification of relativistic electron microburst losses during the GEM storms. *Geophysical Research Letters*, *31*(4), L04802. https://doi.org/10.1029/2003GL018621

Otto, A. (2001). Geospace Environment Modeling (GEM) magnetic reconnection challenge: MHD and Hall MHD—Constant and current dependent resistivity models. *Journal of Geophysical Research*, *106*(A3), 3751–3757. https://doi.org/10.1029/1999JA001005

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

Potts, J. M. (2021). Basic concepts. In I. T. Jolliffe & D. B. Stephenson (Eds.), *Forecast Verification*. https://doi.org/10.1002/9781119960003.ch2

Pritchett, P. L. (2001). Geospace environment modeling magnetic reconnection challenge: Simulations with a full particle electromagnetic code. *Journal of Geophysical Research*, *106*(A3), 3783–3798. https://doi.org/10.1029/1999JA001006

Pulkkinen, A., Kuznetsova, M., Ridley, A., Raeder, J., Vapirev, A., Weimer, D., et al. (2011). Geospace environment modeling 2008–2009 challenge: Ground magnetic field perturbations. *Space Weather*, *9*(2), S02004. https://doi.org/10.1029/2010SW000600

Pulkkinen, A., Rastätter, L., Kuznetsova, M., Hesse, M., Ridley, A., Raeder, J., et al. (2010). Systematic evaluation of ground and geostationary magnetic field predictions generated by global magnetohydrodynamic models. *Journal of Geophysical Research*, *115*(A3), A03206. https://doi.org/10.1029/2009JA014537

Pulkkinen, A., Rastätter, L., Kuznetsova, M., Singer, H., Balch, C., Weimer, D., et al. (2013). Community-wide validation of geospace model ground magnetic field perturbation predictions to support model transition to operations. *Space Weather*, *11*(6), 369–385. https://doi.org/10.1002/swe.20056

Raeder, J., Berchem, J., & Ashour-Abdalla, M. (1998). The geospace environment modeling grand challenge: Results from a global geospace circulation model. *Journal of Geophysical Research*, *103*(A7), 14787–14797. https://doi.org/10.1029/98JA00014

Raeder, J., & Maynard, N. C. (2001). Foreword [to special section on proton precipitation into the atmosphere]. *Journal of Geophysical Research*, *106*(A1), 345–348. https://doi.org/10.1029/2000JA000600

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

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

Rastätter, L., Kuznetsova, M. M., Vapirev, A., Ridley, A., Wiltberger, M., Pulkkinen, A., et al. (2011). Geospace environment modeling 2008–2009 challenge: Geosynchronous magnetic field. *Space Weather*, *9*(4), S04005. https://doi.org/10.1029/2010SW000617

Rastätter, L., Shim, J. S., Kuznetsova, M. M., Kilcommons, L. M., Knipp, D. J., Codrescu, M., et al. (2016). GEM-CEDAR challenge: Poynting flux at DMSP and modeled Joule heat. *Space Weather*, *14*(2), 113–135. https://doi.org/10.1002/2015SW001238

Ridley, A. J., Hansen, K. C., Tóth, G., De Zeeuw, D. L., Gombosi, T. I., & Powell, K. G. (2002). University of Michigan MHD results of the geospace global circulation model metrics challenge. *Journal of Geophysical Research*, *107*(A10), 1290. https://doi.org/10.1029/2001JA000253

Riley, P. (2018). Statistics of extreme space weather events. In *Extreme Events in Geospace* (pp. 115–138). https://doi.org/10.1016/B978-0-12-812700-1.00005-4

Robinson, R., Zhang, Y., Garcia-Sage, K., Fang, X., Verkhoglyadova, O., Ngwira, C., et al. (2019). Space weather modeling capabilities assessment: Auroral precipitation and high latitude ionospheric electrodynamics. *Space Weather*, *17*(2), 212–215. https://doi.org/10.1029/2018SW002127

Rock, D., & Grant, H. (2016). Why diverse teams are smarter. *Harvard Business Review*. https://hbr.org/2016/11/why-diverse-teams-are-smarter

Roederer, J. G. (1988). GEM: Geospace environment modeling. *Eos, Transactions American Geophysical Union*, *69*(33), 786–787. https://doi.org/10.1029/88EO01064

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

Shay, M. A., Drake, J. F., Rogers, B. N., & Denton, R. E. (2001). Alfvénic collisionless magnetic reconnection and the Hall term. *Journal of Geophysical Research*, *106*(A3), 3759–3772. https://doi.org/10.1029/1999JA001007

Shim, J. S., Rastätter, L., Kuznetsova, M., Bilitza, D., Codrescu, M., Coster, A. J., et al. (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., Song, I.-S., Jee, G., Kwak, Y.-S., Tsagouri, I., Goncharenko, L., et al. (2023). Validation of ionospheric specifications during geomagnetic storms: TEC and foF2 during the 2013 March storm event-II. *Space Weather*, *21*(5), e2022SW003388. https://doi.org/10.1029/2022SW003388

Shim, J. S., Tsagouri, I., Goncharenko, L., Rastaetter, L., Kuznetsova, M., Bilitza, D., et al. (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

Sigsbee, K., Cattell, C. A., Mozer, F. S., Tsuruda, K., & Kokubun, S. (2001). Geotail observations of low-frequency waves from 0.001 to 16 Hz during the November 24, 1996, Geospace Environment Modeling substorm challenge event. *Journal of Geophysical Research*, *106*(A1), 435–445. https://doi.org/10.1029/2000JA900090

Siscoe, G., & Fedder, J. (1994). "Daunting" task of assembling data for a geospace global circulation model begins. *Eos, Transactions American Geophysical Union*, *75*(29), 330–331. https://doi.org/10.1029/94EO00983

Smith, A. W., Rae, I. J., Forsyth, C., Oliveira, D. M., Freeman, M. P., & Jackson, D. R. (2020). Probabilistic forecasts of storm sudden commencements from interplanetary shocks using machine learning. *Space Weather*, *18*(11), e2020SW002603. https://doi.org/10.1029/2020SW002603

Spence, H. E. (1996). Geospace environment modeling program flourishes. *Eos, Transactions American Geophysical Union*, *77*(25), 237. https://doi.org/10.1029/96EO00167

Starck, J. G., Sinclair, S., & Shelton, J. N. (2021). How University diversity rationales inform student preferences and outcomes. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(16), e2013833118. https://doi.org/10.1073/pnas.2013833118

Stephenson, D. B., Casati, B., Ferro, C. A. T., & Wilson, C. A. (2008). The extreme dependency score: A non-vanishing measure for forecasts of rare events. *Meteorological Applications*, *15*(1), 41–50. https://doi.org/10.1002/met.53

Tóth, G., Meng, X., Gombosi, T. I., & Rastätter, L. (2014). Predicting the time derivative of local magnetic perturbations. *Journal of Geophysical Research: Space Physics*, *119*(1), 310–321. https://doi.org/10.1002/2013JA019456

Trattner, K. J., Burch, J. L., Fuselier, S. A., Petrinec, S. M., & Vines, S. K. (2020). The 18 November 2015 magnetopause crossing: The GEM dayside kinetic challenge event observed by MMS/HPCA. *Journal of Geophysical Research: Space Physics*, *125*(7), e2019JA027617. https://doi.org/10.1029/2019JA027617

Tu, W., Cunningham, G. S., Chen, Y., Henderson, M. G., Camporeale, E., & Reeves, G. D. (2013). Modeling radiation belt electron dynamics during GEM challenge intervals with the DREAM3D diffusion model. *Journal of Geophysical Research: Space Physics*, *118*(10), 6197–6211. https://doi.org/10.1002/jgra.50560

Tu, W., Li, W., Albert, J. M., & Morley, S. K. (2019). Quantitative assessment of radiation belt modeling. *Journal of Geophysical Research: Space Physics*, *124*(2), 898–904. https://doi.org/10.1029/2018JA026414

Weimer, D. R. (2001). An improved model of ionospheric electric potentials including substorm perturbations and application to the Geospace Environment Modeling November 24, 1996, event. *Journal of Geophysical Research*, *106*(A1), 407–416. https://doi.org/10.1029/2000JA000604

Weiss, L. A., Reiff, P. H., Weber, E. J., Carlson, H. C., Lockwood, M., & Peterson, W. K. (1995). Flow-aligned jets in the magnetospheric cusp: Results from the geospace environment modeling pilot program. *Journal of Geophysical Research*, *100*(A5), 7649–7659. https://doi.org/10.1029/94JA03360

Welling, D. T., Ngwira, C. M., Opgenoorth, H., Haiducek, J. D., Savani, N. P., Morley, S. K., et al. (2018). Recommendations for next-generation ground magnetic perturbation validation. *Space Weather*, *16*(12), 1912–1920. https://doi.org/10.1029/2018SW002064

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 3. https://doi.org/10.1038/sdata.2016.18

Wilks, D. S. (2019). *Statistical methods in the atmospheric sciences* (4th ed.). Academic Press.

Wiltberger, M., Merkin, V., Lyon, J. G., & Ohtani, S. (2015). High-resolution global magnetohydrodynamic simulation of bursty bulk flows. *Journal of Geophysical Research: Space Physics*, *120*(6), 4555–4566. https://doi.org/10.1002/2015JA021080

Winglee, R. M., Papitashvili, V. O., & Weimar, D. R. (1997). Comparison of the high-latitude ionospheric electrodynamics inferred from global simulations and semiempirical models for the January 1992 GEM campaign. *Journal of Geophysical Research*, *102*(A12), 26961–26977. https://doi.org/10.1029/97JA02461

Yu, Y., Liemohn, M. W., Jordanova, V. K., Lemon, C., & Zhang, J. (2019). Recent advancements and remaining challenges associated with inner magnetosphere cross-energy/population interactions (IMCEPI). *Journal of Geophysical Research: Space Physics*, *124*(2), 886–897. https://doi.org/10.1029/2018JA026282

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*(2), 299–312. https://doi.org/10.1029/2018SW002031

Zhang, H., & Zong, Q. (2020). Transient phenomena at the magnetopause and bow shock and their ground signatures. In Q. Zong, P. Escoubet, D. Sibeck, G. Le, & H. Zhang (Eds.), *Dayside Magnetosphere Interactions*. https://doi.org/10.1002/9781119509592.ch2

Zheng, Y., Ganushkina, N. Y., Jiggens, P., Jun, I., Meier, M., Minow, J. I., et al. (2019). Space radiation and plasma effects on satellites and aviation: Quantities and metrics for tracking performance of space weather environment models. *Space Weather*, *17*(10), 1384–1403. https://doi.org/10.1029/2018SW002042

Zheng, Y., Jun, I., Tu, W., Shprits, Y. Y., Kim, W., Matthiä, D., et al. (2024). Overview, progress and next steps for our understanding of the near-Earth space radiation and plasma environment: Science and applications. *Advances in Space Research*. https://doi.org/10.1016/j.asr.2024.05.017