Topical Discussion Meeting report - TDM2

Name of the TDM: <u>Through Validation: Building Confidence in Space Weather Services for a Resilient Operational Future</u>
Conveners: Daria Shukhobodskaia (STCE/SIDC-ROB, Belgium), Veronique Delouille (ROB, Belgium), Sophie Murray (DIAS, Ireland), Suzy Bingham (Met Office, UK) - Secretary
Date: 1330-1430 Tue 28th Oct 2025
Number of attendees *(approximate):* ~90 people (~70 in the room, ~50 online)
Panellists: KD Leka (NorthWest Research Associates, US), Martin Reiss (NASA CCMC), Christian Möstl (Austrian Space Weather Office, Austria), Kasper van Dam (KNMI, Netherlands)
Form of TDM: Panel Forum
Description: see Annex 1.

Objective of the TDM

To focus on the role of validation in operational space weather services - covering metrics, frameworks, user feedback, and communication of uncertainty - aiming to bring together researchers, forecasters, and end-users across sectors.
Panellists were invited to cover particular areas:

- Statistical Validation Specialist – KD Leka
- Model Developer / Research-to-Operations (R2O) Expert - Martin Reiss
- End-User Representative – Kasper van Dam
- Service Developer – Christian Möstl

Discussion Highlights

The TDM opened with remarks from the lead convener, Daria Shukhobodskaia, who introduced the panellists. Each panellist then delivered a brief presentation offering their perspective on the topic (refer to Annex 2). Discussion then followed.

Christian Mostl, the Austrian Space Weather Office (ASWO), emphasised the importance of model robustness, especially for real-time operations. Early integration of end-user needs into the development process was highlighted as essential for operational relevance. AI/ML models were recognised in particular for their potential in event detection. Sub-L1 monitoring (e.g., Bz, CME arrival times) and scoreboards were identified as critical tools to improve lead time and forecast accuracy. Europe was encouraged to pursue independent real-time data capabilities, with missions like Vigil, SHIELD, and L3 seen as strategic assets.

Kasper van Dam, KNMI, presented findings from PECASUS showing inconsistencies in ICAO space weather advisories depending on which of the four centres is on duty. While technically understandable, these inconsistencies were confusing for end-users, underscoring the need for consistent policy and service delivery. The dissemination system for warnings was deemed as important as the warnings themselves, with a call for better coordination and transparency.

KD Leka discussed flare forecasting and validation frameworks, highlighting a particular benchmarking activity comparing operational flare forecasting methods (Leka et al., 2019). KD emphasised the use of a number of skill metrics for validating flare forecasts – noting that not one forecasting method came top across all metrics.

Unfortunately, due to a technical issue with the online meeting platform, Martin Reiss appeared on screen but his audio was unavailable. Although he had submitted his presentation slides in advance, we regret that he was unable to participate in the discussion.

The panel agreed on the need for robust, standardised validation frameworks, especially for comparing models across domains and centres. Human-in-the-loop forecasting was valued for operational environments, though automation (e.g., for Mars missions) is increasingly pursued.

There was a strong call for:

- Benchmark datasets for flares and CMEs.
- Consistent evaluation frameworks across operational centres.
- User engagement throughout the R2O2R pipeline—from model development to service delivery.

Building long-term relationships, training, and understanding user needs were seen as essential for trust and effectiveness between researchers/service providers and end-users.

- The session concluded with a discussion on model maturity and forecast goals:
  - Model maturity tracking using AULs and TRLs.
  - Translating user requirements into measurable forecast goals.

Main Conclusion of the Meeting

The TDM included discussion on lessons from Austria's (ASWO's) R2O2R efforts, emphasising model robustness, early integration of end-user needs, and the importance of AI/ML for event detection. Sub-L1 data and scoreboards (e.g., for Bz and CME arrival times) were highlighted as critical for improving lead times and forecast accuracy.
Findings from analysing PECASUS advisories were highlighted, showing inconsistencies in the ICAO space weather advisories depending on which of the four centres is on duty—highlighting the need for consistent policy and service delivery.
Flare forecasting comparisons were discussed, noting that multiple metrics were required as there was no *universal* skill metric. The need for robust, standardised validation frameworks was noted.
The panel emphasised the value of human-in-the-loop forecasting, especially in operational environments, while acknowledging the push toward automation (e.g., for Mars missions).
The group called for better benchmarking datasets, consistent evaluation frameworks (e.g., for flares and CMEs), and stronger engagement with end-users. Building long-term relationships, training, and understanding user needs were seen as essential for effective space weather services.
The session closed with a discussion on model maturity tracking (e.g., AULs, TRLs) and the importance of translating user requirements into measurable forecast goals.

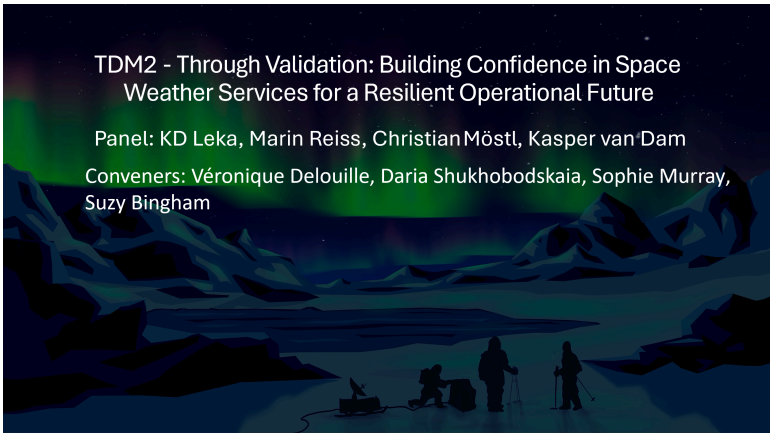
Annex 1: TDM Description

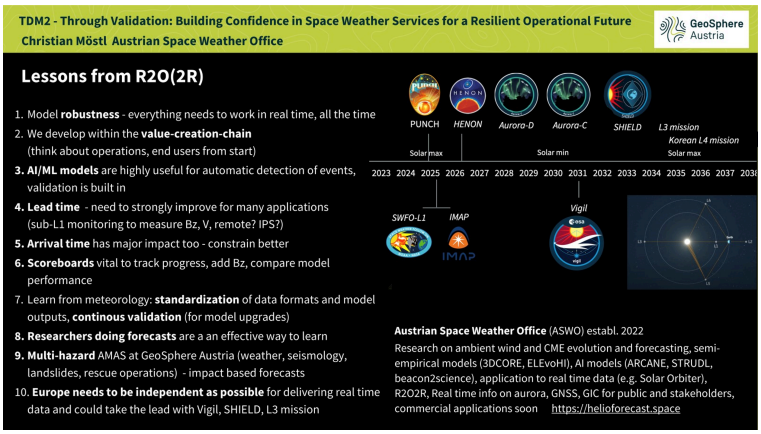
Space weather services are essential for sectors like satellite operations, GNSS, aviation, energy, and crewed spaceflight, where accurate forecasts help prevent costly disruptions. As their use grows, these services must be validated for real-world operations, transparent in their methods, and trusted by users. This session explores the full validation process, from research to operational deployment, emphasizing how end users benefit every step of the way.
We will examine practical questions such as which frameworks and metrics best measure the quality and reliability of space weather services. We'll look at ways to compare models and products across different operational centers and how to clearly communicate uncertainty and confidence to users.
The session will cover key performance tools - like scoreboards, maturity indices, and Key Performance Indicators - that track forecast accuracy, latency, and false alarms. We will also focus on evaluating resilience during significant storm events and the role of user feedback in confirming a

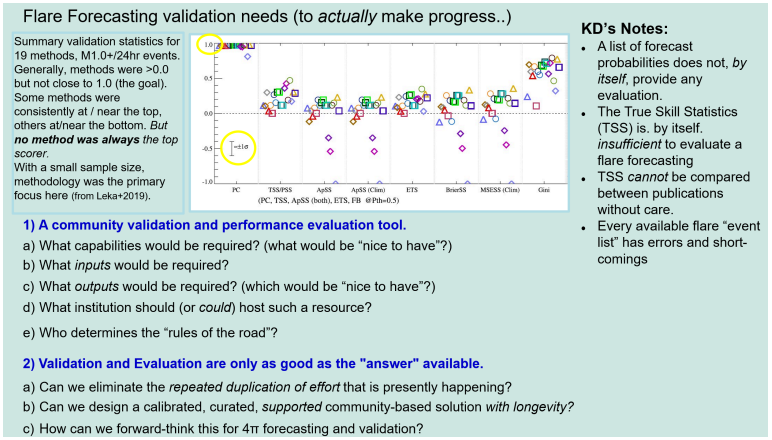
product’s effectiveness.
By uniting researchers, operators, and end users, this session aims to show how thorough validation builds trust, supports informed decision-making, and drives the development of flexible, user-driven space weather capabilities.

Annex 2: Materials Presented

At the start of the TDM, each panellist presented brief material, introducing their perspectives of the topic.







Assessment of Space Weather Modeling Capabilities

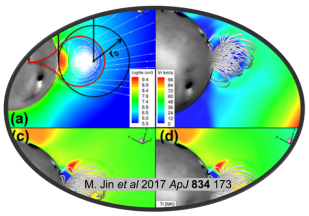
European Space Weather Week 2025

Martin Reiss, Maria Kuznetsova, Edmund Henley, and many more

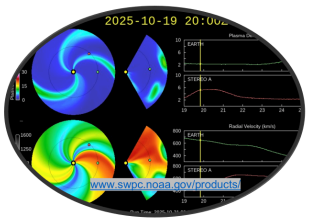
Tuesday, October 28th, 13:30 (CEST)

Two Core Challenges in Advancing Space Weather Forecasting

1. Develop Predictive Capabilities

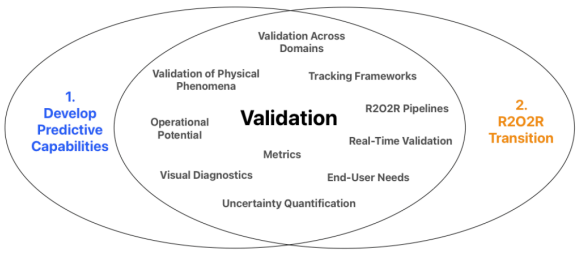


2. R2O2R Transition



To deliver reliable space weather forecasts, we must both develop new predictive capabilities and ensure their smooth transition to operations.

Validation is essential to make progress in both challenges

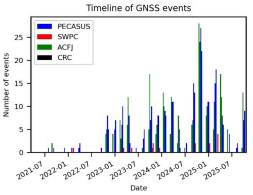


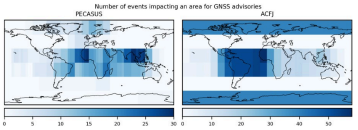
Model validation is **multi-faceted** and involves many **interconnected components**.

Kaspervan Dam, RoyalNetherlandsMeteorologicalInstitute

Statistical Analysis of ICAO Space Weather Advisories

- Fourcentresalternatelyprovidinga global spaceweatherservicefor aviation
- Servicediffersevery2 weeks
  - Understandablefrom technicalperspective (complicatedoperationinvolving technical/political/culturalchallenges, differentinfrastructures,etc. etc.)
  - Not understandablefrom an end-user perspective
- Moregeneralpoint: Thewarning disseminationsystemis asimportantasthe warningsthemselves!
  - Moreinstruments, data, science,won't solvethis byitself. Good policyis justas essential.





Annex 3: Minutes of meeting

Daria Shukhobodskaia opened the TDM, introducing each panellist who then presented an introductory slide.
Christian Mostl provided an overview of Lessons from R2O(2R). ASWO are focusing on accelerating R2O2R. They see model robustness as important. When transitioning from R2O they consider the value-creation-chain, thinking about operations and end-users from the start. ASWO mostly uses AI and ML methods for auto-detection of events, noting that validation is built-in. ASWO recognises that they need to improve lead-time for many applications, for example, with sub-L1 observations. Arrival time scoreboards are important. There were plans for a Bz Scoreboard. If there were sub-L1 data then there could be Bz predictions – but at the moment there aren't so many predictions of Bz. We can learn from meteorology – for example, using standardised data formats and continuous validation. Researchers doing the role of forecaster is a good way for them to learn. Multi-hazards are necessary to consider and impact forecasts too. Europe needs to be as independent as possible – we need more real-time data and we can take the lead through Vigil, SHIELD and an L3 mission.

Kasper van Dam explained that KNMI is part of PECASUS – one of the four centres that present warnings and advisories for ICAO. Kasper has conducted statistical analyses of ICAO space weather advisories. It's important to have data, science and instruments but also to have policy in place. GNSS is one advisory. Kasper showed a plot of GNSS advisories per week – every week that PECASUS is on duty, there are many more advisories than when another of the three centres is on duty. Depending on the duty centre, the distribution of advisories is different – effectively, the service available to aviation differs every two weeks – this is understandable from the centres' perspective as there are a lot of infrastructure and policies involved but this is not understandable for the end-user. The end-user needs a consistent service. The warning dissemination policy is just as important as the science itself.

KD Leka, who focus on research, has done a lot of work in flare forecasting comparisons – a necessary part of R2O – comparisons and benchmarking are needed to identify what works best for operations. How can we tell when new flare forecasting models/facilities are going to work in operations? This is particularly difficult as there have been so many papers published on new methods, particularly using ML models.
KD showed a plot comparing 19 flare forecasting methods (Leka et al., 2019) – there's no single best skill score or metric because of the need of end-users – none of the methods compared were always the top scorer for all metrics. TSS is a popular metric but insufficient to evaluate a flare forecast. TSS cannot be compared between publications without care. Every available flare event list has errors and short-comings.
The CCMC Flare Scoreboard doesn't currently have a verification system. Many resources are being reused – can we eliminate repeated duplication of effort? KD posed a handful of questions (refer to slide), encouraging discussion.
Martin Reiss joined online but it was very unfortunate that the room facilities were not working, so we were unable to hear Martin.

Discussion between audience and panel followed the panellist slides.

A question from the audience: when validating flare models, do you only care about the model or do you care about the environment – do you assess with unreliable inputs for example, as is seen in operations? KD replied, in the Leka et al., 2019 paper, flare 'facilities' were compared – they were penalised if they did miss a forecast. KD suggested that it was time to compare research models – in a little more relaxed way perhaps than is done for operational models – through comparisons at different levels. In comparing research models, you shouldn't be able to use data after the fact but robustness could be considered by using, for example, tiers of capability – ultimately testing with unreliable (operational) data streams, i.e. robustness of the model. It should be possible to build a tool quickly, get a large set of validation statistics, then compare to operational centres to see how well the research model performs in comparison. It's a lot of work to bring a research tool into operations so doing staged validation is useful.

A second question from the audience: SWPC flare forecasts have humans-in-the-loop, isn't a forecast without an operator what is really needed? How an auto-system performs is a vital question. KD noted that human-in-the loop forecasts performed better in the 2019 paper but that it's valuable to evaluate tools before they are in front of the operator. Kasper said that we can draw from meteorology, at KNMI we have a forecaster in the loop. Christian also noted that at ASWO there was a human-in-the-loop, otherwise it is too risky.
But shouldn't operational tools include the human in the comparisons? KD noted that there was a comparison in ~2013 that looked at the experience of forecasters and forecast performance – experience was a variable affecting forecast performance. Forecaster training includes simulating large events but some new forecasters don't have the experience on the actual bench. Kasper noted that it may be cheaper to do auto-forecasts without a human but it's a lesson from meteorology that having a human-in-the-loop works best. KD also noted that auto-forecasts need really good data so may not be so cheap.
In the US, long delays in communicating between Earth and Mars means they're working on auto-forecast systems – for example, for astronauts to understand when they should shelter on Mars – it's too much to train astronauts to forecast, with all their other tasks.

Everyone likes to say their model is the best – is there a consistent evaluation framework? Christian commented that for CME arrival times, there's a CCMC CME Scoreboard and papers have compared predictions (e.g. Riley et al., 2018) – for example, reporting that predictions haven't improved over ~5 years. This CME comparison is a standard.
KD, notes that the flare event prediction Scoreboard needs additional aspect, statistical analysis. Providing an algorithm for the CCMC probability Scoreboard means a lot of overhead – it would be useful to the community to be able to easily upload output to the Scoreboard and then have comparative skill scores – with a range of metrics. A framework doesn't yet exist. Christian also noted that for the CME Scoreboard, there are also improvements needed. It was suggested that if the research and operations community requested this, it may help to further progress.
Daria suggested that different catalogues and the CME Scoreboard have different CME arrival times – it would be nice to have some benchmarks for comparisons. How do we create and use benchmarking catalogues? Christian noted that for CME arrival times this is a notorious problem – we have done some studies but are only now getting large enough datasets for comparisons – ASWO are working on multi-point datasets too. For the CME Scoreboard, it's often not clear how the arrival time is derived. A benchmark dataset would be useful and agreement on standard scores. KD said we also need benchmarking datasets

for flares – it's costly in terms of resource to set up a dataset at the start of every student project.

Communication and understanding who the users are is important. For aviation for example, pilots don't need to know which statistical method was used – they just need top level information.
There's a difference in how a researcher and a service provider look at a model and understand performance – the service provider looks at performance from an end-user perspective - how can a researcher talk to end-users? Kasper explained that there were some hurdles, e.g. sometimes needing security clearance – highlighting that building long-term relationships with end-users like KLM proves valuable. Talking to end-users about their procedures really helps to understand their needs – e.g. finding that dispatchers read advisories so if service providers make advisories available to upload to their specific software tools then products are useful. Building personal relationships with end-users is important. Christian noted that you can have a long email list of end-users but it can be rare to get feedback – you have to build up connections otherwise you don't get that feedback.
A relationship was built between satellite operators and NASA researchers – regular meetings were held, there were annual conferences between forecasters and end-users – these were essential so that both researchers and end-users were all on the same page.
How many end-users or stakeholders are here at ESWW? And forecasters? Not many of either. Training courses have helped for building relationships with end-users – so if an end-user leaves who understands the system – the replacement learns quickly to fill the gap.
Training courses have worked well for building relationships with end-users. Christian noted that in Austria we understand this is a good idea. Veronique noted that we do this in aviation and with other end-users but maybe we need to consider from other perspectives – for example, having discussions with end-users from the start of developing models/applications. When developing models, ASWO have some feedback from end-users – e.g. aurora chasers – and ASWO are looking to build these relationships with other end-users in the future. KD highlighted that if researcher focuses on research then it's someone else's role to understand the end-user – that everyone doesn't have to do all jobs.
At ASWO, they do try to understand the whole process. The '2R' is important – if you get feedback from an end-user like, improve X metric by a certain %, then this is really useful feedback from an end-user for how they want a model/forecast evaluated – but in this request, it's also important to be able to access the relevant data.
Do end-users need to understand statistical metrics? KD suggested that if a user says they don't want to miss any events, then that is translatable – this type of specific request is useful. Christian suggests that this is also valid for power grid end-users – we shouldn't miss a Carrington-type event but then they can receive many false alarms.
Application Usability Levels, AULs, are an R2O tracking framework, does anyone use these? Met Office have used. ASWO suggested that AULs could be useful to compare model levels. NASA uses TRLs, similar to AULs, to identify model maturity.

Daria thanked panellists and the audience, and the session closed at 1435.