Metrics, Verification and Validation

ESWW11, Liège, Belgium, 20 Nov 2014

Minutes of Meeting (v1)

Agenda

16:30 Introduction

16:35 Towards a verification framework for forecasts of different centres, Andy Devos, J. Andries, C. Verbeeck, D. Berghmans (STCE-ROB), Belgium.

16:45 Verification of extreme event forecasts, Peter Wintoft, Swedish Institute of Space Physics, Sweden.

16:50 SWENET Index Quality Statistics & Database Assessment, Alexi Glover, ESA.

16:55 "Performance Verification of Solar Flare Prediction Techniques: Present Status, Caveats, and Expectations", Manolis Georgoulis, Academy of Athens, Greece.

17:05 On the use of modified Taylor diagrams to compare ionospheric models, Sean Elvidge, University of Birmingham, UK.

17:15 Translating verification experience from Meteorology to Space Weather, Suzy Bingham, Met Office, UK

17:25 Lessons learned from CCMC-led community-wide model validation challenges. Outlook on international coordination of M&V activities, Maria Kuznetsova, CCMC , USA.

17:35 Discussion

18:00 End of the session

Attendees (-incomplete)

AG – Alexi Glover
SB – Suzy Bingham
PW – Peter Wintoft
AD – Andy Devos
MK – Maria Kuznetsova
PJ – Piers Jiggens
SE – Sean Elvidge
MA – Matthew Angling
MG – Manoulis Geogoulis
TO – Terry Onsager

Andy Devos – reflections on STCE forecasting

AD presented work at the Belgian Solar Terrestrial Centre of Excellence (STCE) on evaluation of forecasts for Solar flare probability, Kp Index, 10.7 cm radio flux and solar proton events.

AD expressed that there were many ISES & other forecast centres & so it was difficult for end users to understand the best data to use.

Looking at Kp, questions for validation concern whether to compare local or global forecasts, max or average values and how different forecast windows and lead times will have an impact on forecast accuracy.

The STCE have agreed terminology for alerts/warnings but how subjective are they, and for comparisons between both different forecasters and different forecast centres, what should constitute 'unsettled' for Kp and what is 'eruptive flaring'?

AD stated that forecasts should be: simple, clear, well-defined, consistent, use unambiguous terminology, relevant, well-structured, flexible and customisable for the user/purpose, easy to access, use appropriate time scales and provide appropriate detail.

For comparison one ideally needs to: fix timescales, decide on common parameters, common format, common terminology, provide data access, and use adequate metrics which are easy to interpret, easily reproducible and not hedgeable. For forecast validation it would be useful when one has multiple lead times to be able to combine the scores via some weighting of different lead times.

Suggested metrics and verification analysis included ROC curves, error analysis & reliability diagrams.

For reliability diagrams of forecasted verses observed probability, outputs should be within confidence boundaries. In areas where forecasts exceed these boundaries, forecasts should not be used.

AD stressed that there should be more transparency in our forecasts, communication between centres and users is important, that different user domains required different tailoring/approaches and that forecast centre groups should coordinate for the definition of validation parameters.

Peter Wintoft – Extreme events

Speaking about extreme events PW highlighted that there are three different perspectives: the science perspective (e.g., what can we learn from other solar like stars), statistics perspective (based

on a timeline of what has already observed) and user perspectives (what do they see as extreme). In the last case, from a Swedish point of view, only the strongest geomagnetic storms are considered to be of relevance for electrical grid operation.

Regarding forecast verification in general PW highlighted the series of papers by Murphy (see, for example, [RD 1]). as giving a framework for forecast verification.

Forecasting extremes raise big difficulties for many skill scores because these tend to degenerate to trivial values when we look at rare events (see, for example, [RD 2]). To address this problem the Extremal Dependence Index (EDI) has been developed [RD 3]. The EDI has the suitable properties of being independent of the rarity of events, difficult to hedge, asymptotically equitable, and lies in the range [-1,1]. The EDI has been explored by applying it to the ground dB/dt forecasts provided at the Swedish Space Weather Center (http://src.irf.se).

Alexi Glover – SWENET index quality statistics and database assessment

The ESA SWENET system (http://swe.ssa.esa.int) set up 10+ years ago has now been collecting data for a solar cycle meaning that this is a good time to assess performance of models and forecasts initially developed as part of the ESA space weather applications pilot project. AG referenced initial work on performance assessment presented elsewhere during the ESWW [RD 4] which addresses Dst forecast and Ap nowcast service developments.

Simple metrics exist on the system and are automated and applied to complete time series. Equitable skill scores with weighting for rare events are being investigated and could be added as part of a foreseen list of new scores forming part of a redeveloped data browsing and analysis interface.

In upcoming SSA SWE Expert Service Centre activities, a harmonised approach to validation and estimation of accuracy will be investigated and an assessment of accuracy/performance will ultimately be provided along with all SSA SWE service products.

Manoulis Georgoulis – Validation efforts applied to solar flare prediction

There is a need to understand the physical parameter (e.g. solar flare) before we look into how to predict it. Asking a statistics colleague MG was informed that solar flare time series from a single active region appeared as pink noise (meaning that there is a high stocasticity).

When using metrics for validation one should assess the value of different metrics, i.e. some quantitative assessment of the quality of the metric.

Reliability diagrams are very useful for this type of forecast. There are Brier and other Skill Scores (SS) which can be adapted for this type of probabilistic forecast.

Training sets and testing sets are important but then the SSs vary depending on which part of the data is used for training and which for testing. These differences could come from using different solar cycles or different parts of the solar cycle for training/testing. It is therefore necessary to alternate which part of the data is used for testing, perform a full set of training and testing procedures such that each data segment is part of the testing set once and then to average the resulting SSs.

Binary scores are created through application of thresholds. In response to the question of how do we set thresholds (YES/NO) for probabilistic forecasts MG proposes setting the threshold to maximise the skill scores. One question on this is point is how does that match with what users need? Especially as skill would increase with reducing threshold as we look at rarer events.

This will be further addressed in the A-EFFort SSA development activity.

Sean Elvidge– On the use of Modified Taylor diagrams to compare ionospheric models

SE highlighted the benefits of plots which use normalised statistics by taking the std of the model and dividing it by the std of the observations. One can then plot different model outputs in a meaningful way on the same plots using modified Taylor Diagrams. These provide an easy way of visualising and comparing statistical information about a number of models, and for multiple parameters, simultaneously. In the presented approach, 5 statistical parameters are presented in a single diagram. SE and MA have a paper published in radio science including detail [RD 5].

Suzy Bingham – translating verification experience to space weather

Regarding validation the Met Office are planning work with the CCMC for TEC model outputs. Regarding verification the Met Office plans to use experience with terrestrial weather and likewise for applications metrics and KPIs.

The models currently run operationally at the Met Office are Enlil & Relativistic Electron Forecast Model (REFM). The Met Office are working on implementation of other models including Bath University's Multi-Instrument Data Analysis System (MIDAS).

The Met Office would like to be able to view all the outputs from the 4 regional warning centres who run Enlil, plotted against ACE, to provide a poor man's ensemble & to understand any differences between the models.

It's important at the Met Office to think about real-time verification so that forecasters are aware of model accuracy in order to make confident decisions about warnings. For example, REFM is plotted against actual GOES data for forecasters to view.

There are 9 metrics, or performance indicators, which are used as verification for the Public Weather Service, some of which may be applicable for space weather. These are: (1) severe weather warning accuracy, (2) forecast accuracy, (3) public value, (4) public reach, (5) service quality, (6) emergency responder value, (7) responder reach, (8) national capability & (9) milestone achievement.

The Met Office issue twice daily space weather guidance which includes probabilistic forecasts of geomagnetic storms, X-ray flares, high energy electrons and protons. To verify these forecasts, the Met Office are planning to adapt a flexible weather verification system which is currently used for rainfall warnings, etc.

Building on weather forecast experience, forecast assessment builds on simple contingency tables to add categories to give thresholds and timing information. Meaning that a forecaster would get some credit for predicting something a bit too small or not quite at the right time rather than this being flagged completely incorrect.

ROC (Relative Operating Characteristic) curves of FAR (False Alarm Rate) against POD (Probability of Detection) form a cornerstone of model evaluation and one can use the ROC area (area under ROC curve) to give a measure of quality combining FAR and POD measures.

Application metrics can be used for picking up on model run completion time, availability.

Business Performance Measures (BPMs) are targets set by government for the Met Office. The 'Forecast Accuracy' BPM uses verification of different terrestrial models. In the future, Enlil may be included in this BPM so verification to show improvement in Enlil accuracy will be required.

The Met Office is taking data requirement inputs from WMO OSCAR database (http://www.wmo-sat.info/oscar/applicationareas/view/25).

Maria Kuznetsova: Lessons learned from CCMC-led community-wide model validation challenges

Validation is one of CCMCs main activities & is undertaken when a new model arrives at CCMC.

When one is doing a validation it is natural to assume the errors come from the model but actually there can be errors in the validation data or in the algorithm used to transform the model outputs into the form for performing the validation (post-processing). As a result one must be very careful that you don't misrepresent the model inaccuracies during the validation process.

The CCMC has undertaken the following challenges: GEM (2008) – Magnetosphere; CEDAR (2009) – Ionosphere; SHINE (2011) – Solar.

Physical parameters from GEM-CEDAR challenges include: (1) magnetic perturbations at geosynchronous orbits, (2) joule heating/Poynting flux along DMSP, (3) auroral boundaries, (4) neutral densities at CHAMP, (5) electron density parameters at CHMAP, ISRs, COSMIC, (6) TEC from ground-based GPS, (7) Dst index, (8) magnetic perturbations at ground stations & regional K.

Plotting of results and automatic calculation of skill scores is possible on the CCMC system.

At present there is a CME arrival time prediction scoreboard (http://kauai.ccmc.gsfc.nasa.gov/SWScoreBoard/) with 17 registered methods and in the future there will be a flare forecast scoreboard which is being planned in a collaboration between the CCMC and Met Office.

MK stated that on the basis of their work with validation metrics, models have been chosen to be made operational (by NOAA for example).

MA pointed out that this can have the opposite effect if researchers fear a loss of funding when they don't do well in such challenges. This can discourage participation. It's important to frame the validation challenge in a representative and unbiased way.

Discussion Points

1. What metrics and validation techniques are required in the current space weather landscape?

Due to the varied nature of space weather forecasts including probabilistic forecasts, binary forecasts, time series predictions, 2-d 3-d matrices of predictions as well as single value predictions a variety of techniques and metrics are needed.

For probabilistic forecasts one question is how to set thresholds (YES/NO)? One idea is to set the threshold to maximise the skill scores but it Is not clear how that works with inter-model comparisons and how it matches up with what users want. This is especially true because skill scores often deteriorate as we look at rarer events but these are often the thresholds most interesting to operators. Equitable skill scores were mentioned, but not discussed extensively during this session (see [RD 6] for more details).

Taylor diagrams seem like a great way to combine different synoptic metrics such as correlation, bias, etc. in a single plot. This is potentially very useful for inter-model comparisons, but not likely to be the sort of metric you could easily present to a non-specialist service end-user.

One clear target should be an agreed list of events to be used for validation exercises. This should not be dictated by modellers but they must have input because they know what is sufficient and what selections might bias the process.

2. What are the key challenges currently in model and forecast benchmarking?

It is important that we are careful in identifying where errors are coming from whether they be model errors, data errors or interpolation errors for getting results that may be compared. Data is used to drive models and for validation so if there is an issue at either end of the process it could result in an incorrect assessment of model validity.

In cases where forecasts are generated manually there may also be human errors or differences in interpretation.

3. What direction should the space weather community be taking?

It was agreed that additional work in this area is needed and a coordinated approach would be very beneficial. Modelling groups and forecast centres have complementary requirements for validation and metrics. It is important that information flow between these communities such that forecast centres can present results with appropriate metrics for the service user and models are tested against metrics which can be translated into user requirements.

4. What actions can agencies and organisations take in order to support a wider space weather validation effort?

Agencies can provide an unbiased platform and encourage participation in community wide initiatives. One thing that agencies and organisations can provide is computing resources and manpower for carrying out independent tests, as is done at the CCMC. Organisation of dedicated workshops and campaigns involving scientific and application development/forecast communities.

5. How to establish agreed realistic model/service targets to encourage targeted development and prototyping?

There are a range of parameters which space weather models presently forecast including: Dst, solar wind at L1, TEC, electron fluxes at GEO, proton fluxes in the SAA (at a range of energies), Probability of a significant solar flare, K-index, local ground geomagnetic fields, 10.7 cm radio flux, proton events, foF2.

Much work has already been done and is ongoing in order to identify the main parameters and timescales relevant for end users. These differ per user community but programmes such as ESA SSA, The WMO's ICTSW and NOAA's SWPC have established lists of products requested by various user communities (e.g http://www.wmo-sat.info/oscar/applicationareas/view/25, http://swe.ssa.esa.int/DOCS/SSA-SWE/SSA-SWE-RS-SSD-0001_i1r3.pdf)

The lead times or forecast horizons should ideally be fixed or else we would need to find a fair way to weight forecasts such that those delivered further in advance were given higher grading. Again, these forecast windows should be driven by end user needs.

One thing seems relatively clear which is that given the wide range of space weather domains it would initially be simpler to organise an assessment of this type by domain involving entities who really understand the wide range of models and forecast techniques per domain, along with their strengths and weaknesses. This doesn't mean there is not collaboration, in fact, it is critical that different groups learn from others regarding validation metrics and their correct use. Furthermore, as efforts continue to build cross-domain tools such as the VSWMC and SWMF, involvement of experts from multiple domains will become increasingly important in validating coupled systems.

6. What targeted actions would encourage groups not currently involved to further participate in space weather validation activities?

MK's presentation made the point that a subset of models involved in the challenge workshops coordinated by the CCMC have been transitioned to operations following successful validation campaigns. However, organisations embarking on such campaigns need to exercise caution as this can have the opposite negative effect if researchers fear that a poor performance will lead to a loss of further development funding. It's important to frame the validation challenge in a representative and unbiased way and include the modellers in the loop when finalising challenges both in terms of the events/time period selected for study and means by which the results will be presented to the wider community and potential funding bodies. However, an unbiased organisation/agency must finalise the selection in order to ensure an unbiased result.

General Conclusions

The session was deemed very interesting by many participants. It was felt that it was probably not long enough and that not enough time was given over to the discussion points. Some way of co-ordinating efforts in this area would be welcomed and participants would rather not wait a year until this could be discussed again because there would be a great loss of momentum.

It was agreed to propose a further splinter meeting at the next ESWW and to include this topic as a more regular agenda item.

More targeted verification actions are needed as we look to transition models from the research domain into operations.

Upcoming meetings

A preliminary list of meetings where a validation & verification session is planned or could be associated to is as follows:

1. Space Weather Workshop, Boulder, Colorado, April 2015
2. European Space Weather Week 12, Belgium November 2015
3. ILWS/COSPAR workshop focussing on COSPAR SW Roadmap, India January 2016 [metrics and validation session in planning]
4. COSPAR PSW Event: *Metrics and Validation Needs for Space Weather Models and Services*, COSPAR Assembly, Istanbul, August 2016

References

[RD 1] Murphy, A. H., *"A general framework for forecast verification,"* Monthly Weather Rev., 1987.

[RD 2] Murphy, A.H., *"The Finley Affair: A Signal Event in the History of Forecast Verification,"* Weather and Forecasting, 11, 3--20, 1996.

[RD 3] Ferro, A.T. and Stephensson, D.B., *"Determinstic forecasts of extreme events and warnings, Forecast Verification,"* Ed. I.T Jolliffe and D.B. Stephensson, Wiley-Blackwell, 2012.

[RD 4] Laurens, H. et al, *"The SWENET Online Archive: 10 Years of a European Space Weather Community Resource,"* 11th European Space Weather Week - Session 10, 2014.

[RD 5] Elvidge, S., Angling, M. & Nava, B., *"On the use of modified Taylor diagrams to compare ionospheric assimilation models,"* Radio Science, Volume 49, Issue 9, pp. 737-745, 2014.

[RD 6] Clarke, E. & Thomson, A., *"Forecast evaluation as applied to geomagnetic activity categories,"* 10th European Space Weather Week – Session 12, 2013.