TECHNICAL NOTE

Guidelines for common validation in the SSA SWE Network

Prepared by	**I. Tsagouri, C. Borries, C. Perry, M. Dierckxsens, J. de Patoul, C. Cid, T. Moretto-Jorgenson,**

Reference	**ssa-swe-escdef-tn-5401**
Issue/Revision	**2.2**
Date of Issue	**08/09/2020**
Status	**approved**

European Space Agency
Agence spatiale européenne

APPROVAL

Title Guidelines for common validation in the SSA SWE Network	
Issue Number 2	**Revision Number** 2
Author I. Tsagouri, C. Borries, C. Perry, M. Dierckxsens, J. de Patoul, C. Cid, T. Moretto-Jorgenson, D.S. Bloomfield	**Date** 08/09/2020
Approved By	**Date of Approval**
I-ESC coordinator	i.A. C.B Digital signiert von Claudia Borries DN: C=DE, O=Deutsches Zentrum fuer Luft- und Raumfahrt e.V. (DLR), CN=Claudia Borries Grund: Ich bin der Verfasser dieses Dokuments Ort: Neustrelitz Datum: 08-09-2020 08:46:51 Foxit PhantomPDF Version: 9.7.1 i.A. J. Berdem Digital signiert von Jens Berdermann DN: C=DE, OU=SO-WWE, O=DLR, CN=Jens Berdermann, E=jens.berdermann@dlr.de Grund: Ich genehmige dieses Dokument Ort: Neustrelitz Datum: 08-09-2020 10:38:51 Foxit PhantomPDF Version: 9.7.1
ESA	Alexi Glover Digitally signed by Alexi Glover Date: 2020.09.11 13:51:26 +02'00'

CHANGE LOG

Reason for change	Issue Nr.	Revision Number	Date
Version based on best practice review. This version was discussed in the VWS	0	0	12/10/2018
Revision based on the input received during the VWS discussions	0	1	20/12/2018
Update based on input from WG5 members	0	2	05/03/2019
First issue being part of WG5 P3-SWE-V delivery	1	0	12/03/2019
Second issue including feedback from validation campaigns and new recommendations for continuous validation	2	0	14/05/2020
Revision based on ESA feedback	2	1	11/08/2020
Small changes on last ESA feedback, Sec. 8.5	2	2	08/09/2020

CHANGE RECORD

Issue Number 2	**Revision Number 2**		
Reason for change	**Date**	**Pages**	**Paragraph(s)**
Minor editing recommended by ESA	08/09/2020	43	8.5

DISTRIBUTION

Name/Organisational Unit

Table of contents:

1 INTRODUCTION

1.1 Background

The “Inter ESC Working Group 5 (P3-SWE-WG5)” of the ESA Space Situational Awareness Programme - Period 3, aims to support the development of harmonized validation across the different Expert Service Centre (ESCs) of the SSA Space Weather Service Network, helping the users to judge the quality of the products available through those ESCs. To this effect, WG5 needs to compile well-defined guidelines for coordinated product validation at ESC level.

A first issue of this document has been released based on: i) review of best practices in terms of product validation in order to suggest a common approach for the validation campaigns in each ESC; and ii) organization of a validation workshop (VWS).

This is the second issue of this document, which updates the guidelines for validation campaigns based the feedback from validation campaigns executed by all ESCs in 2019; and extends the guidelines to cover also recommendations for continuous validation.

1.2 Purpose and scope of the document

This document has been prepared in the frame of the P3-SWE-WG5 activities and is an output from the task described above.

The scope of this document is to describe a common approach for the execution of validation campaigns and continuous validation in each ESC. This document provides guidelines for the generation of the plan and the report of validation campaigns and recommendations for the presentation of continuous validation on the SSA SWE Portal.

For the sake of clarification, the definition of validation is included here and how it compares to verification. These definitions are adopted from [RD-1] and complemented with a description on continuous validation:

Validation

Validation is a **process** which demonstrates that the **product** is able to accomplish its intended use in the intended **environment**. The status of the product following validation is “validated”. Verification is a pre-requisite for validation. Continuous validation or revalidation is used to check that the product continues to accomplish its intended use.

Footnote: The definition applies also to tool and processes.

Verification

Verification is a **process** which demonstrates through the provision of objective evidence that the **product** is designed and produced according to its **specifications** and the agreed **deviations** and **waivers**, and is free of **defects**. A waiver can arise as an output of the verification process. Verification can be accomplished by one or more of the following methods:

- analysis (including similarity),
- test,

- inspection,
- review of design.

The status of the product following verification is “verified”.

Footnotes: a) The term specification is intended as Product Specification Document (PSD) requirement [AD-SWEPSD]; b) The definition applies also to tool and processes.

This document revision has been prepared by I. Tsagouri (NOA) in the frame of ESA Contract No. 4000113184/15/D/MRP, with contribution from C. Borries (DLR), C. Perry (RAL), M. Dierckxsens (BIRA-IASB), J. de Patoul (ROB), C. Cid (UAH), T. Moretto-Jorgenson (UiB).

1.3 Document life cycle

This document was generated by P3-SWE-WG5 members. In its initial (draft) version, the document was prepared through review of best practices in terms of product validation to serve as basis for assessment, application and discussion in a validation workshop (VWS) held in October 2018, with participation by all ESCs (except from G-ESC, which was not yet contracted). Based on the initial recommendations, ESCs generated a validation test plan during the workshop and provided feedback about the applicability of the draft version to P3-SWE-WG5, as well as recommendations for revision. The first issue of the document included the revised guidelines, based on which the ESCs executed their validation campaigns in 2019. This second issue of this document incorporates the feedback about the applicability of the first issue of this document, which was provided by each ESCs to P3-SWE-WG5. In addition, a first version of recommendations for continuous validation has been incorporated in this second issue.

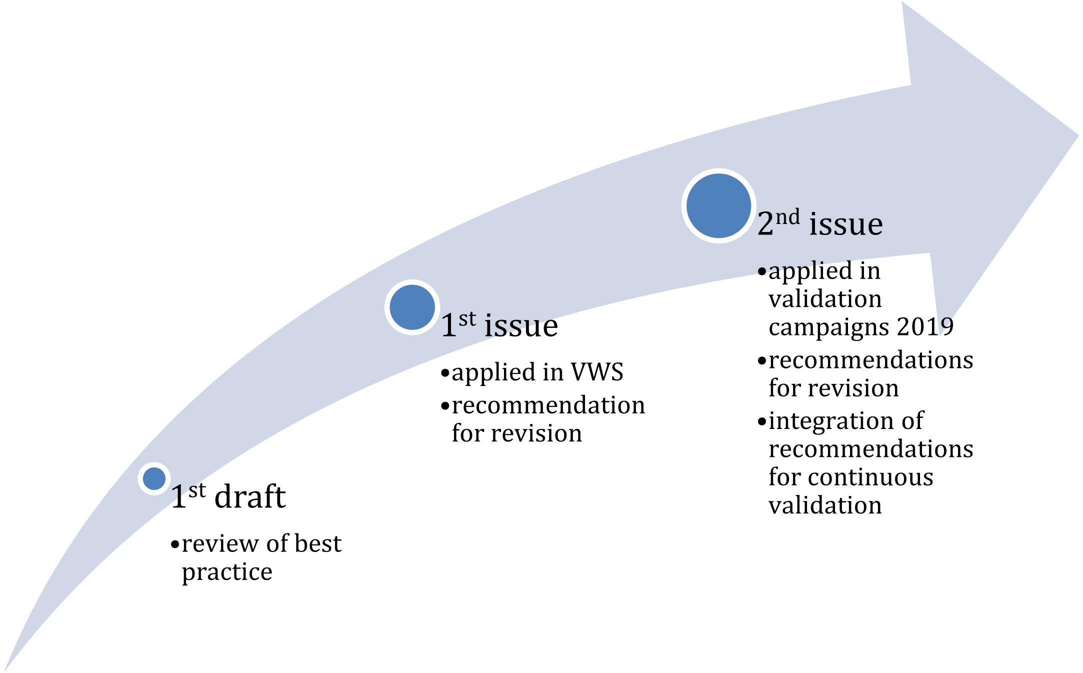

1st draft
•review of best practice
1st issue
•applied in VWS
•recommendation for revision
2nd issue
•applied in validation campaigns 2019
•recommendations for revision
•integration of recommendations for continuous validation

Figure 1-1 Evolution of this document

1.4 Applicable documents

ID	Document Title	Reference, Issue, Date
[AD- SWESET]	SWE Service Template (SWE-SeT)	SSA-SWE-ESCDEF-DRD-0100, i1r0, 12/02/2016
[AD-SWERD]	SWE Roadmaps	
[AD-SWEPSD]	SWE Product Specification Document	SSA-SWE-RS-SSD-0001,i1r3, 08/07/2013

1.5 Reference documents

ID	Document Title	Reference, Issue, Date
[RD-1]	SSA SWE GLOSSARY P3-SWE-WG4: WORKING GROUP 4 ON TERMINOLOGY	ssa-swe-escdef-tn-w401, i2r5, 05/02/2020
[RD-2]	http://www.cawcr.gov.au/projects/verification/	
[RD-3]	Forecast Verification: A Practitioner's Guide in Atmospheric Science (Second Edition)	Edited by I.T. Jolliffe & D.B. Stephenson, John Wiley & Sons Ltd, ISBN: 978-0-470-66071-3,2012.

ID	Document Title	Reference, Issue, Date
[RD-4]	Solar Flare Prediction Using Time Series of SDO/HMI Vector Magnetic Field Data and Machine Learning Methods	Bobra, M. G. & Couvidat, S., 2015, Astrophys. J., 798, 135, DOI: 10.1088/0004-637X/798/2/135
[RD-5]	Toward Reliable Benchmarking of Solar Flare Forecasting Efforts	Bloomfield, D. S., Higgins, P. A., McAteer, R. T. J. & Gallagher, P. T., , 2012, Astrophys. J., 747, L41, DOI: 10.1088/2041-8205/747/2/L41
[RD-6]	Feature Ranking of Active Region Source Properties and the Uncompromised Stochasticity of Flare Occurrence	Campi, C., Benvenuto, F., Massone, A. M., Bloomfield, D. S., Georgoulis, M. K. &Piana, M., , 2019, Astrophys. J., 883, 150, DOI: 10.3847/1538-4357/ab3c26
[RD-7]	Gene Selection for Cancer Classification using Support Vector Machines	Guyon, I., Weston, J., Barnhill, S. *et al.*. *Machine Learning* **46**, 389–422 (2002). https://doi.org/10.1023/A:1012487302797
[RD-8]	Evaluation of the performance of DIAS ionospheric forecasting models	Tsagouri, I., J. Space Weather and Space Clim., 1, A02, 2011, DOI: 10.1051/swsc/2011110003.
[RD-9]	Space Weather Service Network Preliminary Product Validation for the Period of Heightened Activity Observed in September 2017	Burley, S et al., , *16th European Space Weather Week*, Nov. 2019, Liege, Belgium, https://register-as.oma.be/esww16/contributions/public/S16-P1/S16-P1-04-BurleySophie/Session16_Validation_Poster.pdf
[RD-10]	Developing the LDi and LCi geomagnetic indices, an example of application of the AULs framework	Cid, C., Guerrero, A., Saiz, E., Halford, A. J., & Kellerman, A. C.,. *Space Weather*, 18, e2019SW002171, 2020. https://doi.org/10.1029/2019SW002171
[RD-11]	Validation of IMPC beta TEC maps	V. Wilken and M. Kriegel, 2020, ssa-swe-escion-tn-4412, issue 1, revision 1, 16 July 2020, ssa-swe-escion-tn-4412_i1r1b-7_2020_signed_CB-VW-MK.pdf
[RD-12]	Development and integration report of integrated UAH products	SSA-SWE-P2-SWE-2.0-TN08

1.6 Acronyms and abbreviations

AE	Auroral Electrojet index
AUC	Area Under Curve

AVIDOS	AVIationDOSimetry (ESA SSA SWE Network)
BIRA-IASB	Royal Belgian Institute for Space Aeronomy
BS	Brier Score
BSS	Brier Skill Score
CME	Coronal Mass Ejection
CRD	Customer Requirements Document
DIAS	European DIgital upper Atmosphere Server
DLR	German Aerospace Center
EIS	European Ionosonde Service (ESA SSA SWE Network)
ESA	European Space Agency
ESC	Expert Service Center
FAR	False Alarm Ratio
FN	False Negative
FP	False Positive
G-ESC	Geomagnetic Conditions – ESC
GFZ	German Research Centre for Geosciences
GOES	Geostationary Operational Environmental Satellite system
HSS	Heidke Skill Score
I-ESC	Ionospheric Weather ESC
IQD	International Quiet Days
IRF	Swedish Institute of Space Physics
MAE	Mean Absolute Error
ME	Mean Error
MRE	Mean Relative Error
MSE	Mean Squared Error
NOA	National Observatory of Athens
POD	Probability Of Detection
POFD	Probability Of False Detection
PSD	Product Specifications Document
RCAAM	Research Center for Astronomy and Applied Mathematics
RMSE	Root Mean Squared Error
ROC	Relative Operating Characteristic curves
ROTI	Rate Of TEC (Total Electron Content) Index
SR	Success Ratio
SRD	System Requirements Document
SS	Skill Score
SSA	Space Situational Awareness
SSCC	SSA Space Weather Coordination Centre
SWE	Space Weather
SWIF	Solar Wind driven autoregression model for Ionospheric short-term Forecast
TEC	Total Electron Content
TN	True Negative
TP	True Positive
TS	Threat Score
UAH	University of Alcala
UIO	University of Bergen
VWS	Validation Workshop

WG5	Working Group 5

2 PRODUCT OVERVIEW

The SSA SWE Network delivers numerous products to pass to the user of the SWE Network a wide range of space weather information including forecasts[1], nowcasts[2], alarms, models, indices and measurements (raw or processed). Typically, the products are delivered in real time, but a posteriori products and product archives are also foreseen in several occasions.

The number of the SSA SWE Network products is continuously increasing to presently count more than 150 products, while they are delivered in many different formats to cover a variety of spatial scales (e.g., single-site, regional and global) and temporal scales (e.g., point-in-time, short-term and long-term forecasts). All above outline a complex scene that requests the elaboration of an effective product validation concept, able to cover the needs across all products/ESCs.

2.1 Classification of products for validation purposes

To meet the goals, it is suggested to see the SSA SWE Network products into two general types (independent on their timeliness): **predictions** and **measurements (raw or processed)**.

Predictions

In this context, predictions are considered in a broad frame to include descriptions of the space environment provided for ***past, present or future*** dates as the output of a process or model (i.e., forecasts, nowcasts, alarms and models).

Indices may also fit in this grouping: they may be treated either as predictions - when they are provided as predictions of standard indices (e.g., Kp, AE), or as measurements of an observable quantity (e.g., ROTI).

For validation purposes the products could be further classified according to the nature, specificity and space-time domain they support ([1], [2]). A suggested way of distinguishing SSA SWE products is discussed in the following section.

[1] Forecast: Description of the space environment at a future date based on actual data, proxies and models (SSA-SWE-RS-RD-0001 definition).

[2] Nowcast: Reconstruction in near real-time of one or several parameters based on actual data, proxies and models (SSA-SWE-RS-RD-0001 definition).

2.1.1 *Predictions*

Table 2.1 below lists one way of distinguishing the predictions, followed by supportive definitions (see also [1], [2]).

	Nature of prediction	
	Non-probabilistic	**Probabilistic**
Specificity of the prediction	Continuous Multi-categorical	
Space-time domain	Time series Spatial distribution A combination of the above	

Table2-1: Suggested way of distinguishing predictions within the SSA SWE Network.

Nature of the prediction: A prediction can be:

- *Non-probabilistic* in the case where a single value of a predictand quantity (i.e. the observable quantity that is to be predicted) is predicted.
- *Probabilistic* in case a probability (with a value between 0 and 1 or 0 and 100%) is assigned to the occurrence of the predictand quantity or category (see below).

Specificity of the prediction

- *Continuous:* A continuous predictand is one for which, within the limits over which the variable ranges, any value is possible (e.g. frequency, velocity, magnetic field, density, temperature etc). Predictions are verified against the observed predictand quantity.
- *Multi - categorical:* A prediction in which a discrete number of K categories of separate event definitions each receive individual predictions. The predictions in each of these K categories are verified against their own dichotomous event-definition outcomes (i.e. that specific category event definition did/didn't occur). Predictions issued for these K categories can be either probabilistic (e.g. 0.3 probability of event definition occurring) or dichotomous (i.e. event definition is/isn't expected) in nature, although dichotomous yes/no values are interpreted as probability 1/0.

  It is worth noting that single-category predictions (i.e. K=1) are still considered here (e.g. storm occurrence). For truly multi-category predictions (i.e. K>1), the verification strategies that may be employed depend on whether observations can satisfy the event definitions of multiple categories (i.e. different GOES class flares occurring in the forecast window with prediction categories of "1 or more C-class flares", "1 or more M-class flares" and "1 or more X-class flares"; categories can only be verified individually) or observations can exclusively satisfy the event definition of one of the K categories (i.e. different GOES class flares occurring in the forecast window with prediction categories of “largest flare will be C-class”, “largest flare will be M-class” and “largest flare will be X-class”; categories can be verified either individually or combined into one cross-category verification).

Space-time domain
Time series: a series of prediction points listed in time order.
Spatial distribution: predictions with spatial distribution involving the same parameter over a range of geographic locations (e.g. a map). Then, the product values could be function of both space and time (e.g. series of maps).

In each case, a full set of methods is available to cover a wide variety of particular validation needs (e.g. accuracy, skills, bias, etc; see also http://www.cawcr.gov.au/projects/verification/). Recommended methods for indicative cases are discussed in Section 3.2.

It may be important to note that a prediction may be treated differently in case specific thresholds are defined. For instance:

- A non-probabilistic prediction may be considered as a special case of a probabilistic prediction when a probability of unity is assigned to one of the categories and zero to the others.
- A non-probabilistic multi-categorical prediction can be treated as a set of non-probabilistic binary (dichotomous) predictions by considering each category separately as a binary event.
- A probabilistic binary prediction can be converted into an infinite sequence of non-probabilistic binary predictions by using a sequence of probability decision thresholds. A non-probabilistic binary event is defined to occur when the prediction probability exceeds the threshold probability.

In this respect, it is up to each ESC to elaborate the most representative plan in order to effectively address the users' needs (see also Section 3.4).

2.1.2 Measurements

Measurements are considered here to be of non-probabilistic nature. In this respect, validation methods for non-probabilistic predictions are also valid for the measurements.

3 RECOMMENDED VALIDATION METHODOLOGY

3.1 General concepts

The validation plan is recommended to be established on:

1. Comparison with reference/ground-truth data

A measure of reference/ground-truth data is available and the *discrepancy between predictions* (*or measurements* in case of the validation of measurements) *and reference/ground-truth data can be estimated*. This comparison is a strong requirement in any validation plan.

For definition purposes, one may consider the following:

- Reference data are data that define the set of permissible values to be used by other data fields. Reference data gain in value when they are widely re-used and widely referenced. Typically, they do not change overly much in terms of definition, apart from occasional revisions (https://en.wikipedia.org/wiki/Reference_data). An indicative example of reference data may be the International Sunspot Numbers.
- Ground truth is a term used in various fields to refer to information provided by direct observation (i.e. empirical evidence) as opposed to information provided by inference (https://en.wikipedia.org/wiki/Ground_truth)). In the context of the present document, as inference one may consider a process or model.

The reference/ground-truth data generally stem from observational data (reference/ground-truth data are also reported as observations in the text). Whenever applicable (e.g. in an event-oriented prediction), independent official reports and/or catalogues may be considered as "reference/ground truth".

In many cases there are uncertainties or errors in the observations. Sources of uncertainty include random and bias errors in the measurements themselves, sampling errors and other errors of representativeness, as well as analysis errors, when the observational data are analyzed or processed before being compared to predictions. In any case, it is necessary to discuss the limitations and uncertainties of the reference/ground-truth data during the evaluation of the results.

The difference between prediction (or measurement in case of the validation of measurements) and reference/ground-truth data is assessed by a score[3] (see Section 3.2). The scores should be determined through the suggested classification scheme provided in Section 2.1 and the users'

[3] A measure of the prediction quality.

requirements/needs. This task may be supported by review of users' requirements documents (CRD, SRD, PSD) to receive any useful input regarding desirable specifications per product (e.g. accuracy, prediction horizon, relevance of hits or false alarms).

In case no reference/ground truth data are available for the comparison tests, then cross-comparison between relevant products could be invoked to address the needs. In this case, the results should be communicated to the users in terms of consistency between the compared products.

2. Comparison with reference predictions or model

This part applies mainly to the predictions. This comparison provides information about the value or worth of a prediction relative to a reference prediction or model. The reference prediction is generally an unskilled prediction based on e.g. random chance, persistence (defined as the most recent set of observations, "persistence" implies no change in condition) or climatology (e.g. monthly means or medians)[4]. The reference model is often based on climatology or can be a community-wide agreed standard model.

The relative value of the prediction (or the measurement in case of the validation of measurement) over the reference is assessed by a Skill Score (SS). This is a single number resulting from comparative analysis of related scores (e.g. prediction score vs. reference score). In a generalized formulation, the SS is established as:

$$SS = \frac{score_{prediction} - score_{reference}}{score_{perfect\ prediction} - score_{reference}}$$

Here: SS< 0 means predictions/measurements are worse than reference
SS = 0 means predictions/measurements are as good as reference
0 < SS< 1 means predictions/measurements are better than reference
SS = 1 is a perfect skill score

Notice that the skill score can be unstable for small sample sizes.
The skill score may support also cross-comparison purposes.

[4] It is important to note that the relevant to climatology time periods can be different depending on the product.

3.2 Methods and scores

The literature on Verification/Validation/Uncertainty Quantification of (statistical) models is huge and continuously growing. This section does not have the purpose to give a complete review or cold claim on the best methods. Still, a few common methods and scores will be introduced here to get a first impression. It is strongly recommended to **consult the references suggested in Sec. 3.2.8.**

The sections below are provided for predictions, but as it is already mentioned the methods for non-probabilistic predictions can be applied to measurements as well.

3.2.1 Methods for non-probabilistic dichotomous or binary (yes/no) predictions

A dichotomous or binary prediction says, "yes, an event will happen", or "no, the event will not happen". For certain applications a threshold may be specified to separate "yes" and "no" – e.g. for the case of the occurrence of geomagnetic storms, min Dst less than -30 nT.

To validate this kind of predictions it is recommended to start with a contingency table.

Contingency table

A 2 x 2 contingency table is used in statistics as the simplest way to summarize the relationship between several categorical variables and reference/ground truth data. The table shows the frequency of "yes" and "no" predictions and their corresponding outcomes. The four combinations of predictions (yes or no) and observations (yes or no), called the joint distribution, are:

True Positive (TP)/ Hit: event predicted to occur, and did occur

False Negative (FN)/ Miss: event was not predicted, but did occur

False Positive (FP)/ False alarm: event predicted to occur, but did not occur

True Negative (TN)/ Correct negative: event was not predicted, and did not occur

The total numbers of observed and predicted occurrences and non-occurrences are given on the lower and right sides of the contingency table and are called the *marginal distributions*.

Table 3-1: Contingency table (c.f. [RD-5])

		Observed		**Totals**
		Yes	**No**	
Prediction	**Yes**	*True Positive (TP)*	*False Positive (FP)*	*Prediction yes*
	No	*False Negative (FN)*	*True Negative (TN)*	*Prediction no*
Totals		*Observed yes*	*Observed no*	*Grand Total*

The contingency table is a straightforward way to see what types of errors are being made. A perfect prediction system would produce only hits (TP) and correct negatives (TN), with no misses (FN) or false alarms (FP). Special attention to particular error types (i.e. false alarms (FP) or misses (FN)) should be given based on the users' needs. For instance, considering large scale ionization depletions, HF-propagation users note that missed events are more critical for the reliable performance of their operations than false alarms.

Applicable scores:

A large variety of categorical metrics can be computed from the four elements of the 2 x 2 contingency table above to describe particular aspects of the prediction performance, such as:

Accuracy: Accuracy = (TP + TN)/ total = (TP + TN)/(TP+TN+FN+FP)
Bias score: BIAS = (TP + FP) / (TP + TN)

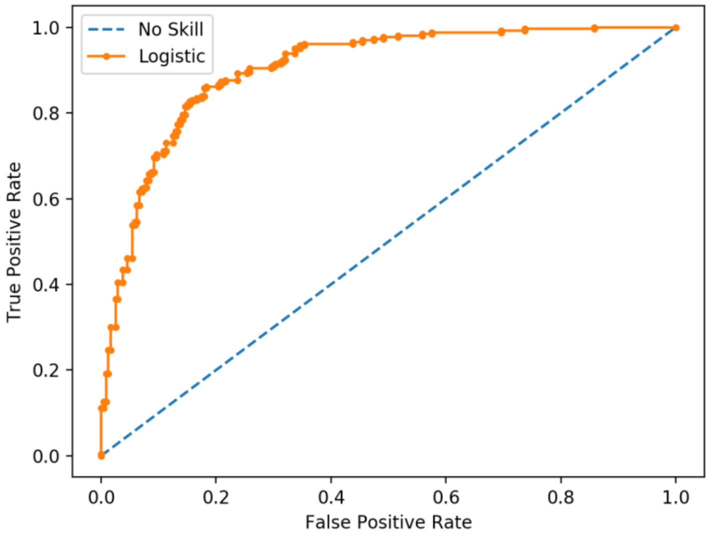
Accuracy and Bias are usually the common skill scores. However, in case of class imbalance (if there are significantly more or fewer examples for one class than for the other(s), c.f. [RD-4]), other scores should be considered, too.

Probability Of Detection (POD, *also known as* ***Recall, Sensitivity*** *or* ***True Positive Rate*****):** POD = TP / (TP + FN)
False Alarm Ratio (FAR, *also known as* ***False Positive Rate*****):** FAR =FP / (TP + FP)
Success Ratio (SR, *also known as* ***Precision*****)**: SR = TP / (TP + FP)
Threat Score (TS): TS = TP / (TP + TN + FP)

Precision and recall are usually anti-correlated: the recall will decrease when the precision increases, and vice versa. Therefore, a useful quantity to compute is their harmonic mean, the f1 score [RD-4]:

$$f1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Important Note: Besides the score estimates used to characterize the quality of the prediction, the values in the contingency table should be kept and provided in the validation report to facilitate future comparisons.

Applicable plots	**Receiver Operating Characteristic (ROC)** curve [RD-2]**:** Plot FAR (false positive rate) vs. POD (true positive rate) using a set of increasing probability thresholds (for example, 0.05, 0.15, 0.25, etc.) to make the yes/no decision. The area under the ROC curve is frequently used as a score.  **Figure 2: ROC Curve of a Logistic Regression Model and a No Skill Classifier (https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/)** ROC graphs are widely used to evaluate classifiers (classifiers are dichotomous or multi-categorical predictands) under presence of class imbalance. However, in case the imbalance is associated to the presence of a low sample size of minority instances, the estimates can be unreliable. **Precision-Recall (PR) Curve**: Plot *Precision* vs. *Recall.* Precision-recall curves (PR curves) are recommended for highly skewed domains where ROC curves may provide an excessively optimistic view of the performance.

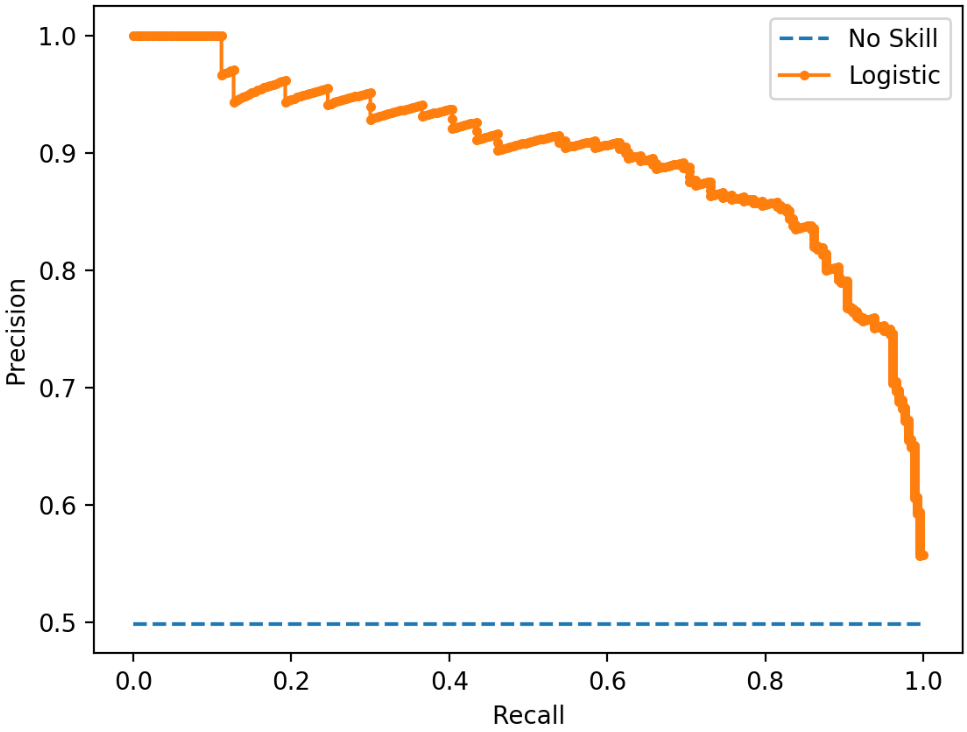


Figure 3: Precision-Recall Curve of a Logistic Regression Model and a No Skill Classifier. (https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/)

Area Under the Curve (AUC): For evaluation, AUC of ROC and PR Curve is going to be interpreted.

- ROC Curves and Precision-Recall Curves provide a diagnostic tool for binary classification models.
- ROC AUC and Precision-Recall AUC provide scores that summarize the curves and can be used to compare classifiers.

3.2.2 Methods for non-probabilistic continuous variables

The validation of products related to continuous variables should aim to measure how the product values (predictions or measurements) differ from the reference/ground truth data. Validation methods for these products may include exploratory plots, such as scatter plots or box plots, as well as various summary scores.

Exploratory plots

Such plots aim to provide a first look at correspondence between product values (predictions or measurements) and the reference/ground truth data and/or similarities between location, spread, and skewness in the corresponding distributions.

Applicable plots/scores:

Scatter plots: Plots the product values against reference/ground truth data. The correlation coefficient and the coefficient of determination may be calculated to measure the degree of the linear association between product values and observations (reference/ground truth data).

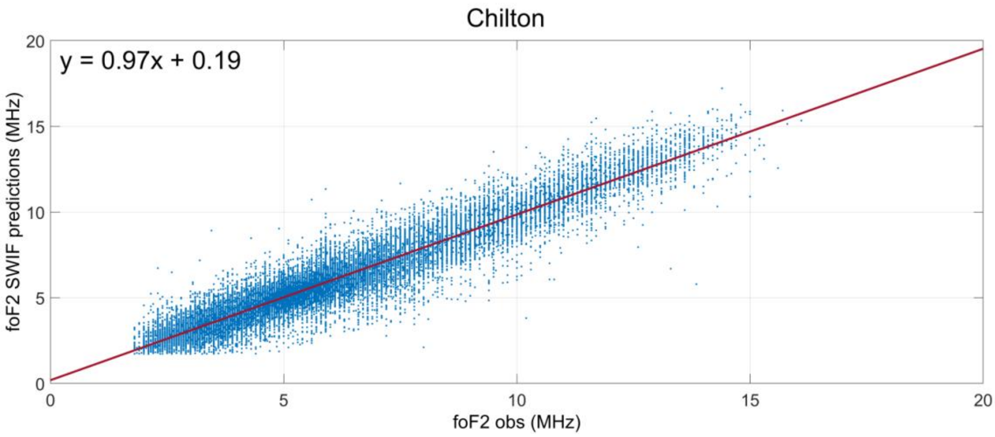


Figure 4: Example for a scatter plot that compares predictions of the foF2 critical frequency provided by the Solar Wind driven autoregression model for Ionospheric short-term Forecast (SWIF) model with foF2 observations provided by Chilton Digisonde (see also [RD-8]). The linear regression line and equation are also given in the plot.

Box plot: Plot boxes to show the range of product values falling between the 25th and 75th percentiles, horizontal line inside the box showing the median value, and the whiskers showing the complete range of the data.

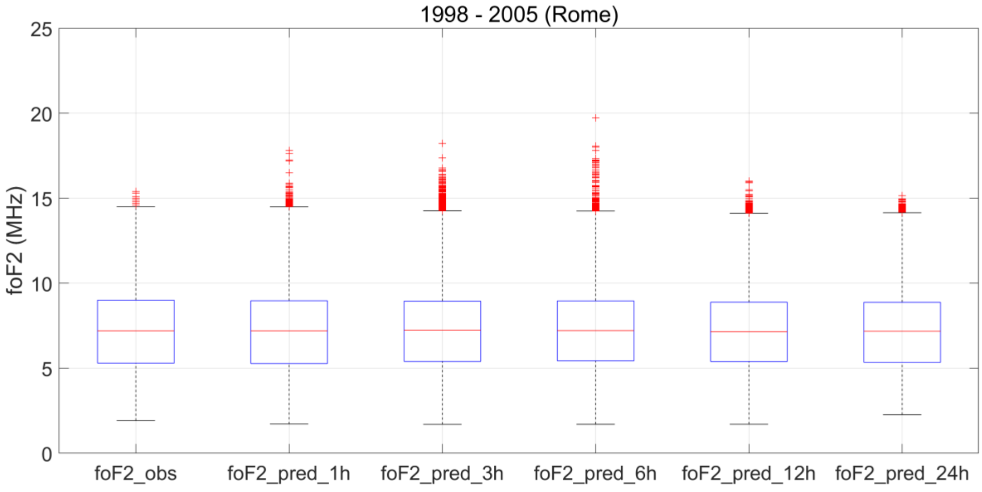


Figure 5: Example of the application of box plots to demonstrate foF2 prediction abilities by comparing the distributions of the observed (foF2_obs) and predicted values (for prediction step from 1h to 24h ahead). The foF2 predictions are obtained by SWIF model and foF2 observations are obtained from Rome Digisonde. The box has lines at the lower quartile, median (red line) and upper quartile values. Whiskers extend from each end of

the box to the adjacent values in the data - in this case to the most extreme values within 1.5 times the interquartile range from the ends of the box. Red crosses represent the outliers (e.g. data with values beyond the ends of the whiskers).

Violin Plot: Violin plot allows to visualize the distribution of a numeric variable for one or several groups. Each 'violin' represents a group or a variable. The shape represents the density estimate of the variable: the more data points in a specific range, the larger the violin is for that range. It is really close to a boxplot, but allows a deeper understanding of the distribution.

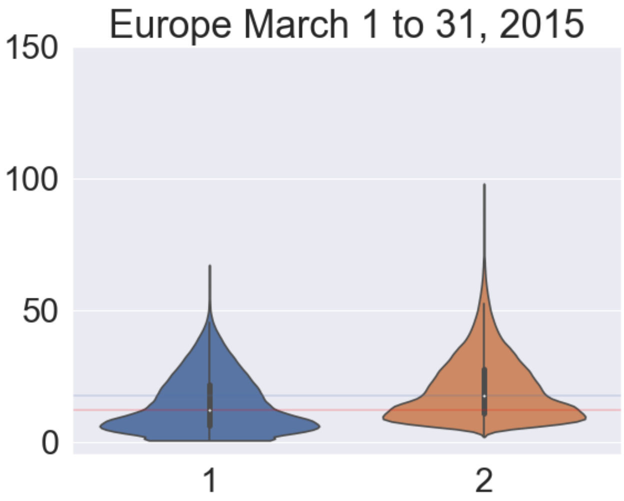


Figure 6 Violin plots of two kind of TEC maps: IMPC beta (1) and IGS (2).

Summary scores

Summary scores listed below aim mainly to provide an estimate of the accuracy of the product values –i.e. the level of agreement between the product values and the reference/ground-truth data (as represented by observations). The difference between the prediction and the observation is the prediction error. The lower the errors, the greater the accuracy.

Applicable scores:

Mean Error(ME):

$$ME = \frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)$$

Mean Absolute Error (MAE):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|P_i - O_i|$$

Root Mean Squared Error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2}$$

Normalized Root Mean Square Error (NRMSE):

$$NRMSE = RMSE/\bar{O} = \sqrt{N\frac{\sum_{i=1}^{N}(P_i - O_i)^2}{\sum_{i=1}^{N} O_i}}$$

There are various ways to normalize. In this case, the mean is used (other option are for instance dividing with the standard deviation, the difference between max and min observed, or difference between 75% and 25% quartile)

Mean Squared Error (MSE):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2$$

Mean Relative Error (MRE):

$$MRE = \frac{1}{N}\sum_{i=1}^{N}\frac{(P_i - O_i)}{O_i}$$

Prediction Efficiency index(PE):

$$PE = 1 - \frac{\langle (P_i - O_i)^2 \rangle}{\sigma_O^2} = 1 - \frac{\langle (P_i - O_i)^2 \rangle}{\langle (O_i - \langle O_i \rangle)^2 \rangle}$$

While the correlation coefficient quantifies the consistence of variations, without measuring the agreement in absolute values, the prediction efficiency index is sensitive to both variations and absolute prediction error.

In all above formulas, P_i and O_i stand for predicted and observed instances, respectively.<...> denotes the arithmetic mean

	Relative improvement: This measures the improvement of the predictions relative to a reference prediction (usually the long-term or sample climatology). It follows the generalized formulation of the SS given in Section 3.1, where the score may be any of the quantities listed above.

Uncertainty propagation	
The propagation of uncertainty (or propagation of error) is the effect of variables' uncertainties (or errors, more specifically random errors) on the uncertainty of a function based on them. When the variables are the values of experimental measurements they have uncertainties due to measurement limitations (e.g. instrument precision) which propagate due to the combination of variables in the function (https://en.wikipedia.org/wiki/Propagation_of_uncertainty). In this context, the uncertainty of the measurements is propagated to the products (e.g. predictions provided through empirical expressions or processed measurements).	
Applicable scores:	The uncertainty can be expressed in a number of ways. It may be defined by the absolute error, the relative error (usually written as a percentage) or the standard deviation. The calculation method of the propagated uncertainty depends on the combination of the variables in the function's formulation (examples on the calculation methods can be found in https://en.wikipedia.org/wiki/Propagation_of_uncertainty.)

3.2.3 Methods for probabilistic predictions

A probabilistic prediction provides a probability of an event occurring, with a value between 0 and 1 or 0 and 100%. In general, it is not straightforward to validate a single probabilistic prediction. Instead, a set of probabilistic predictions is usually validated using observations that those events either occurred or did not occur.

A probability prediction system is evaluated in terms of:

Reliability: agreement between prediction probability and mean observed frequency
Sharpness: tendency to predict probabilities near 0 or 1, as opposed to values clustered around the mean
Resolution: ability of the prediction to resolve the set of sample events into subsets with characteristically different outcomes

Exploratory plots	
Plots to visualize the performance of the prediction method	
Applicable plots/scores:	**Reliability diagrams** plot the observed frequency against the prediction probability, where the range of prediction probabilities is divided into bins (e.g. 0-5%, 5-15%, etc.) with observed frequency calculated separately from each bin of predictions. In practice,

	reliability diagrams indicate differences between probabilities predicted and their resulting average event outcomes (i.e. *observed frequencies*). **Relative/Receiver Operating Characteristic (ROC)** curves plot hit rate (POD) against false rate (Probability of False Detection-POFD), using a set of increasing probability thresholds to convert prediction probabilities into yes/no binary predictions. The Area Under the Curve (AUC) is frequently used as a ROC-derived score. (Example see Sec. 3.2.2) **Discrimination diagrams** plot the likelihood of each prediction probability when the event occurred and when it did not occur. A summary score can be computed as the absolute value of the difference between the mean values of these two distributions.

Summary scores	
Summary scores listed below aim to quantify the performance of a probabilistic prediction.	
Applicable scores:	**Brier Score (BS):** In its simplest form, BS is equivalent to the mean-squared error between the issued prediction probability, *f*(i.e., 0–1), and the observed binary outcome for that prediction, *o* (i.e., 0 or 1), for a total of N prediction – observation pairs $BS = \frac{1}{N}\sum_{I=1}^{N}(f_i - o_i)^2$ If the issued predictions can be identified as *K* groups of unique prediction probabilities, the BS can be decomposed into three components, $\mathrm{BS} = \frac{1}{N}\sum_{k=1}^{K} n_k(f_k - \bar{o}_k)^2 - \frac{1}{N}\sum_{k=1}^{K} n_k(\bar{o}_k - \bar{o})^2 + \bar{o}(1-\bar{o}),$ $= \text{reliability} - \text{resolution} + \text{uncertainty},$ **Brier Skill Score (BSS):** This measures the improvement of the probabilistic prediction relative to a reference prediction (usually the long-term or sample climatology), thus taking climatological frequency into account. It follows the generalized formulation of the SS: $BSS = 1 - \frac{BS}{BS_{ref}}$

3.2.4 *Methods for multi-categorical predictions*

Methods for validating multi-categorical predictions are also based on a generalized contingency table showing the correlations between predictions and observations in the various category bins. It is analogous to a scatter plot for categories.

Contingency table

Multi-category contingency table

			Observed Category			Total
	i,j	1	2	...	K	
	1	$n(F_1, O_1)$	$n(F_1, O_2)$	...	$n(F_1, O_K)$	$N(F_1)$
Prediction	2	$n(F_2, O_1)$	$n(F_2, O_2)$	...	$n(F_2, O_K)$	$N(F_2)$
Category	...	...	...	...	...	...
	K	$n(F_K, O_1)$	$n(F_K, O_2)$	...	$n(F_K, O_K)$	$N(F_K)$
Total		$N(O_1)$	$N(O_2)$	...	$N(O_K)$	N

In this contingency table, $n(F_i,O_j)$ denotes the number of predictions in category i that had observations in category j, $N(F_i)$ denotes the total number of predictions in category i, $N(O_j)$ denotes the total number of observations in category j, and N is the total number of predictions.

Applicable scores:	The *distributions approach* examines the relationship among the elements in the multi-category contingency table. For a perfect prediction system, non-zero elements would be appeared only along the diagonal, while all entries off the diagonal would have values of 0 would. The off-diagonal elements give information about the specific nature of the prediction errors. The *marginal distributions* (*N*'s at right and bottom of table) show whether the predictions produces the correct distribution of categorical values when compared to the reference/ground truth data. **Accuracy:** Accuracy measures the fraction of the predictions that were in the correct category, $Accuracy = \frac{1}{N}\sum_{i=1}^{K} n(F_i, O_i)$ **Heidke Skill Score (HSS):** HSS measures the fraction of correct predictions after eliminating those which would be correct due purely to random chance,

	$HSS = \dfrac{\frac{1}{N}\sum_{i=1}^{K} n(F_i,O_i) - \frac{1}{N^2}\sum_{i=1}^{K} N(F_i)N(O_i)}{1 - \frac{1}{N^2}\sum_{i=1}^{K} N(F_i)N(O_i)}$ More skill scores, such as the Appleman skill score evaluating the performance with respect to climatology for various event categories, could also apply here. There are two variations of HSS discussed in [3], one which varies in the range (-∞, 1] and the other is better normalized, varying in the range [-1, 1]. Other methods, which have been used in the network are e.g.: **Recursive Feature Elimination (RFE)**: RFE is an iterative ranking procedure described in [RD-6] and [RD-7]

3.2.5 Further considerations

3.2.5.1 Validation of spatially distributed samples

In the case of predictions with spatial distribution, the prediction quality may be assessed in a number of ways as the various scores described in the previous sections can be estimated also in a number of ways by partitioning the full data set into various subsets. Once again, the chosen scores and methods should depend on the nature of the prediction and the scope of the validation in combination with users' requirements.

- In the simpler approach, one may ignore the temporal and spatial dimensions and have the entire set of the prediction-observation pairs as a combined ensemble over both space and time, i.e. pool everything into one data set. This approach comes with a big disadvantage: the loss of information on the spatial and temporal variability in the quality of the predictions. For this reason, this approach is only suggested for predictions of relatively rare events, as for instance extreme storm events to ensure sufficient number of events in the test sample.

- Typically, validation approaches aim at providing information on the spatial and temporal variability in the quality of the predictions. To this effect, the full data set is divided into subsets. Two possible ways to deal with spatial and temporal dimensions are given below.
  i. *Spatial averaging:* grouping together all predictions for the entire spatial array at a given time to calculate a score over all spatial points. In this way, a spatial distribution product can be promptly compared against the reference/ground truth at a given time to support validation tests in both "validation campaign" and "continuous validation" modes.
  ii. *Temporal averaging:* Grouping together all predictions at different times at the same spatial location, so that each location (station or grid point or area) is treated as a different variable. In this way, the emphasis in the validation tests could be given to

specific locations that may be of special interest for the users (e.g. high or equatorial latitudes).

The difference between a spatial distribution (e.g. a map) and a combination of time/space data (series of maps in time) might be negligible. During a validation campaign, the user should be more inclined to compare spatial distributions over an interval of time. In a special case of time/space combination, the validation of a spatial distribution product can be seen as the single comparison of it against the reference/ground truth at a specific moment in time. This could be considered when producing continuous validation results (i.e. Real-time or Near Real-Time validations) e.g. to inform the user of the accuracy of the product outcomes. In case a product has clearly the characteristics of spatial distribution (e.g. the I-ESC DIAS/EIS Current Ionospheric conditions) the information presented could be treated as separate time-series as above. In this case the ESC should decide how to treat the product according also to its scope (on the interest of the user).

3.2.5.2 Limited number of data available for statistical validation analyses

It often happens that the number of samples available for validation is rather limited and restricts the statistical reliability of the results (see [RD-9]). For similar cases, the following points are recommended:

1. To ensure sufficient numbers of predictions and observed events in the validation data one may include tests that ignore any difference between the temporal and spatial dimensions of a product. This way the set of predictions and observed values are treated as a combined ensemble over both space and time to give one overall score or skill measure. This is also the simplest approach to address the needs for rare events.
2. For multi-categorical predictions, one may consider the possibility to group the category bins and provide more general results if applicable.
3. Include error bounds on the validation results themselves. This is highly recommended for all cases, but it becomes extremely important in case of small test samples (e.g. using bootstrapping methods).
4. In case a sufficient/satisfactory number of samples cannot be reached, an individual assessment of the events can be performed without a quantitative evaluation based on scores.
5. In case scores are used on small sample sizes, they have to be used cautiously as some may introduce misleading information. For instance:
    - BSS is unstable when applied to small data sets; the rarer the event, the larger the number of samples needed.
    - Accuracy can be misleading since it is heavily influenced by the most common category in the sample (usually the "no event" category). The same holds for TS.
    - ROC estimates can be unreliable in case of imbalance in the sample that is associated to the presence of a low sample size of minority instances.

3.2.5.3 Predictions/ measurements without valid ground-truth

Use independent prediction/ measurement of the same entity

Often, if there is no ground truth available, there are still predictions/ measurements of the same physical entity with independent instruments, methods or data. It will be ideal to use not only one alternative dataset but more (if available). In many cases, it is not clear, which of the different predictions/ measurements is closer to ground-truth. Therefore, they have to be treated at the same level. However if possible, the quality of each dataset should be taken into account and any caveat should be described. The quantities should be comparable. In this case, we should speak of "cross-comparison" instead of "validation". Methods for non-probabilistic continuous variables are applicable for this cross-validation. Basically, Figure 7 shows an example of a cross-comparison of many different maps of Total Electron Content (TEC). All TEC maps estimate the TEC with independent methods that rely on different bias estimations and different background TEC models. Until now, there no measurements of TEC, which can be considered as ground truth. All TEC products have their justification. Thus, no validation against ground truth is possible, but a cross-comparison.

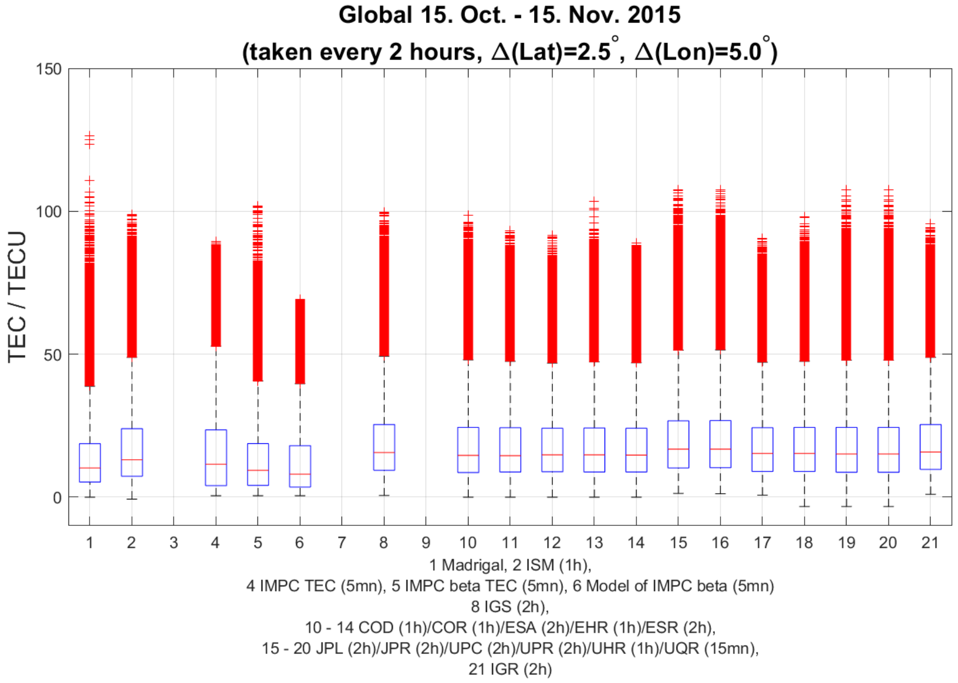


Figure 7: Boxplot of the TEC maps generated by different providers. This figure has been generated in the TEC maps validation campaign 2019 within P3-SWE-V activity [RD-11].

Another example for this cross comparison is the approach implemented to "validate" the ESA SWE product G.126, LDiñ. This index provides the local magnetic disturbance at Iberian Peninsula with one-minute resolution. There is not a similar local geomagnetic index to compare to as local indices are lower temporal resolution.

In one of the validation approaches for the LDiñ, the official three-hour resolution K index from San Pablo-Toledo Observatory (SPT) was used as ground-truth. For cross comparison a K index was computed from LDiñ by taking the maximum value of LDiñ for three hours and translating the value to logarithmic scale.

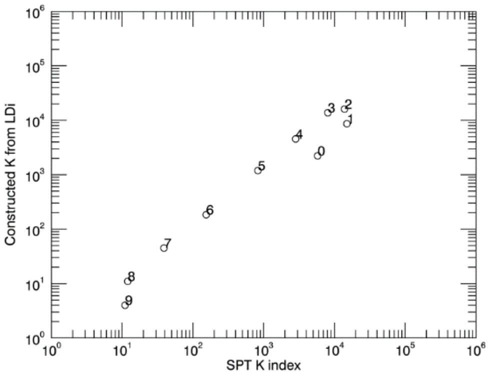


Figure 8: Probability density function of K official values from SPT observatory vs K values reconstructed from LDiñ. This figure has been generated during the development activity P2-SWE-1.5 [RD-12].

Cross comparison of different entities reflecting equivalent parameters

In some cases without valid ground-truth there is also no possibility to find a comparable measurement of the same entity. Thus, a cross comparison process is not applicable. In this case, a recommendation is to find any item involved in the process of getting the prediction/measurement which have valid ground-truth. This is considered as cross comparison of equivalent items. Usually, it is necessary to define assumptions to the comparison between the different data sets. These should be clearly listed (e.g. comparing fluxes between instruments, the different ranges of the energy channels require an assumption of the energy spectrum). It is rather a validation of the procedure than a validation of the parameter.

An example of cross comparison of equivalent items is one of the validation procedures applied to the index LDiñ. Besides there is no valid ground-truth for the LDiñ, as explained above, International Quiet Days (IQDs) can be used as the proper reference data to check one of the most sensible steps in the procedure to compute the index: to discriminate, in real time, if the day under analysis is a quiet day or not. A cross-comparison between the IQDs and the quiet days obtained in the LDiñ analysis provides a validation of, at least, this step in the procedure [RD-10].

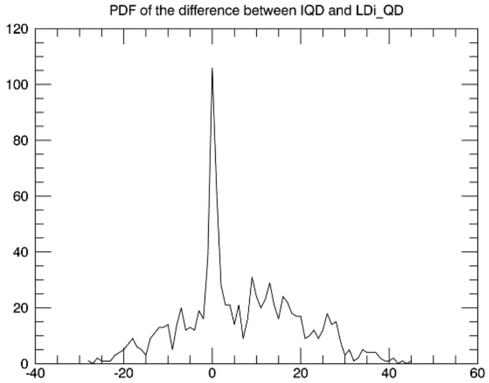


Figure 9: The probability density function of the difference in days between IQDs and the quietest days found with LDiñ procedure. This figure has been generated during the development activity P2-SWE-1.5 [RD-12].

4 TEST TIME INTERVALS

To ensure availability of relevant resources, the validation tests may be applied to time intervals included in the pre-operational phase of each product within the ESA/SSA/SWE. Nevertheless, the validation plan in each ESC should support completeness to the maximum possible extent. To this purpose, it is recommended that a validation plan includes:

i. Tests regarding any possible dependence in the products performance, as for instance:
    - Solar cycle dependence
    - Seasonal dependence
    - Local time dependence
    - Location dependence

ii. Both quiet and disturbed periods at the correct balance, to fulfill climatology;

iii. Intervals different than the ones used for the development of the product.

Ideally, a complete validation plan should anticipate tests for a whole solar cycle and if possible for more than one solar cycles. In this respect, the validation plan may be established as complementary to previous efforts with an eye to future developments.

5 VALIDITY OF THE VALIDATION RESULTS

At first, the evaluation of the validation results needs to discuss the limitations and uncertainties of the reference/ground-truth data.
Then, validation typically involves acceptance and suitability with external customers. In this respect, it is highly recommended that the validation results be discussed against users' requirements/needs (e.g. product accuracy determined by validation tests with respect to the desirable accuracy defined by the users).

6 BEST PRACTICE EXAMPLES

In the following, we provide indicative integrations of space weather products into the suggested predictions' classification scheme. A list of references is given for each case, to provide further support and guidance.

1. Dichotomous (yes/no) predictions (binary predictions)	
Examples	Indicative examples of this class of products include alerts and warnings for: • Arrival of CME at Earth (e.g., Dumbović et al. 2017) • Prediction of solar flare events (e.g., Devos et al. 2014) • Ionospheric storm time disturbances (e.g., Tsagouri and Belehaki, 2015) • Events of enhanced solar wind properties, as for instance solar wind speed (e.g., Reiss et al. 2016)
References	- Dumbović, M., Srivastava, N., Rao, Y.K. et al., Validation of the CME Geomagnetic Forecast Alerts Under the COMESEP Alert System, Sol Phys 292: 96, 2017, https://doi.org/10.1007/s11207-017-1120-5 - Devos, A., C. Verbeeck, and E. Robbrecht, Verification of space weather forecasting at the Regional Warning Center in Belgium, J. Space Weather Space Clim., 4(27), A29, 2014, DOI:10.1051/swsc/2014025. - Tsagouri, I., and A. Belehaki, Ionospheric forecasts for the European region for space weather applications. J. Space Weather Space Clim., 5, A09, 2015, DOI: 10.1051/swsc/2015010. - Reiss, M. A., M. Temmer, A. M. Veronig, L. Nikolic, S. Vennerstrom, F. Schöngassner, and S. J. Hofmeister (2016), Verification of high-speed solar wind stream forecasts using operational solar wind models, Space Weather, 14, 495–510, doi:10.1002/2016SW001390.

2. Predictions of continuous variables	
Examples	Indicative examples of this class of products include measurements, nowcasts and forecasts of: • Geomagnetic or solar indices (e.g., Devos et al. 2014) • Neutral atmosphere densities (e.g., Bruinsma 2015; 2017) • Ionospheric characteristics (e.g., Tsagouri 2011) • Solar wind properties as for instance solar wind speed (e.g., Reiss et al. 2016)
References	- Devos, A., C. Verbeeck, and E. Robbrecht, Verification of space weather forecasting at the Regional Warning Center in Belgium, J. Space Weather Space Clim., 4(27), A29, 2014, DOI:10.1051/swsc/2014025. - Bruinsma S., The DTM-2013 thermosphere model, J. Space Weather Space Clim., 5 (2015) A1 DOI: https://doi.org/10.1051/swsc/2015001

	- Bruinsma S., Daniel Arnold, Adrian Jäggi and Noelia Sánchez-Ortiz, Semi-empirical thermosphere model evaluation at low altitude with GOCE densities, J. Space Weather Space Clim., 7 (2017) A4, DOI: https://doi.org/10.1051/swsc/2017003 - Tsagouri, I. Evaluation of the performance of DIAS ionospheric forecasting models. J. Space Weather and Space Clim., 1, A02, 2011, DOI: 10.1051/swsc/2011110003. - Reiss, M. A., M. Temmer, A. M. Veronig, L. Nikolic, S. Vennerstrom, F. Schöngassner, and S. J. Hofmeister (2016), Verification of high-speed solar wind stream forecasts using operational solar wind models, Space Weather, 14, 495–510, doi:10.1002/2016SW001390.

3. Probabilistic predictions	
Examples	Indicative examples of this class of products include: • Probabilistic flare forecasting (e.g., McCloskey et al. 2018; Murray et al. 2017)
References	- McCloskey AE, Gallagher PT, Bloomfield DS, Flare forecasting using the evolution of McIntosh sunspot classifications. J. Space Weather Space Clim. 8: A34, 2018. - Murray S.A., S. Bingham, M. Sharpe, D. R. Jackson, Flare forecasting at the Met Office Space Weather Operations Centre, Space Weather, 15, 4, 2017.

4. Multi-categorical predictions	
Examples	Indicative examples of this type/class of predictions include: • Multi-categorical solar flare forecast (e.g., Kubo et al. 2017) • Caveats / pitfalls in regards to the generation and time coverage of training and testing samples and potential remedies of the (often severe) class imbalance between the positive and negative samples (Georgoulis and Bloomfield, 2019)
References	Kubo Y., M. Den and M. Ishii, Verification of operational solar flare forecast: Case of Regional Warning Center Japan, J. Space Weather Space Clim., 7, A20, 2017 DOI: https://doi.org/10.1051/swsc/2017018 M. Georgoulis, D.S. Bloomfield, Validation practices and caveats of recent solar flare forecasting studies, SSA-SWE-P3SWEII-TN-1500, issue 1, revision 0, 08 Nov. 2019

7 GUIDELINES FOR THE VALIDATIONS CAMPAIGNS

7.1 Purpose of the guidelines

These guidelines shall help the test manager generate an appropriate validation plan, execute the tests and generate the validation report. The validation report will be the documentation of all information discussed in Sec. 7.2 to 7.4.

7.2 Validation plan

7.2.1 Introduction and scope

- Describe the purpose of this validation campaign
- Which product(s) is/are going to be validated
- What is/are the main use case(s) of this product?[5]
- How is the user supposed to apply this product?
- What are the performance requirements for this application/use case?[6]
- What information should be communicated to the users as result of the campaign?

7.2.2 Test product assessment

- Provide a short technical description of the product(s).

7.2.3 Assessment of available reference/ground-truth data

- The ESC should provide information about recommended reference/ground-truth data for the considered test product:
    - Provide description about the reference/ground-truth data;
    - Describe data availability;
    - Describe any necessary information about uncertainties, biases and limitations of the reference/ground-truth data.
- If no recommendation for reference/ground-truth data is provided by the ESC, make an assessment of potential reference/ground-truth data (description, availability,

[5] Awareness of how product will be used may influence the selection of validation approach. The demonstrated link to high priority use cases will strengthen the case. This work may be supported by the advantage of the SWE Network in that the use cases can be built based on service structure and proposed product linking coupled with user feedback.

[6] SWE requirements baseline accuracy info must be referenced where available. If found to be incomplete or caveats (such as critical data availability) mean these targets aren't currently achievable, limitations can and should be highlighted. Results of SWE Network validation work will provide important input for next review of these values.

limitations, etc.) and identify the most suitable data set that should be the recommended reference/ground-truth data for the considered test product.

- In case of cross-comparison tests, provide the relevant information about the products to be used in the comparisons.

7.2.4 Selection of validation methods

- What type of product is tested (i.e., prediction or measurement)?
- What class of product is tested (see Table 2.1)?
- What validation methods are applicable for the type and class of product tested in this campaign?[7]
- Select validation methods that are most suitable to reflect the use-case requirements.
- Select metrics/scores that are most suitable to reflect the use-case requirements.
- Justify the selection.

7.2.5 Selection of test time period(s)

- The ESC should assess the applicability of common test time periods.
- The ESC should provide recommendations for suitable test time periods.

7.3 Execution of validation campaign and presentation of validation results

This is the most comprehensive part of the validation report. Usually, a very large amount of validation results is generated during one campaign. It is necessary to condense the information and generate an appropriate overview.

7.4 Summary and conclusions

- Summarize the results of the validation campaign, ideally presenting some key numbers, which represent the quality of the product(s) under investigation.
- Indicate any limitations on the adopted validation plan (e.g. small number of test time intervals)
- Discussion
    - of the results with respect to the use case.
    - if the validation work allows to evaluate if user requirements are met
    - if the results reveal anything about the products that could feed in future developments of the products

[7] At this stage, it is also recommended to consider existing community accepted validation approaches, wherever applicable (e.g. radiation exposure to cosmic radiation in aviation) in conjunction to the guidelines provided in the present document.

- Conclusions: Provide some recommendations for users concerning the use and applicability of the product(s). Ideally, the conclusions are written in a way that it can be copied to the federated product website content (e.g. in the help or quality sections).

8 RECOMMENDATIONS FOR CONTINUOUS VALIDATION

8.1 Purpose of the guidelines

These recommendations shall guide the test manager to generate appropriate and easy interpretable validation products provided continuously as accompanying information attached to the main product on the SSA SWE Portal. This helps tracking the quality and performance of the products. However, continuous validation is not always applicable; some validation processes require visual inspections, samples collected over a long period to acquire sufficient statistics or sophisticated comparisons (e.g. warnings for forthcoming disturbances).

Ideally, continuous validation is provided in near real-time. But, there are also many cases where ground-truth or reference data is not available in near real-time. Thus, long time delays of continuous validation are acceptable, too.

Next to continuous validation, continuous quality control is necessary to inform the user about the reliability of a product. The quality control provides valuable information about the current product concerning availability and quality of input data and errors propagated by the uncertainty of input data. Also for the quality control, it is necessary to provide a description how to interpret the quality information.

8.2 Applicable validation methods

Generally, the same methods as for the validation campaigns can be applied. Usually, validation campaigns capture the mean performance of the product, while the continuous validation captures the latest performance of the product. Because continuous validation is done in real-time, it often has to deal with caveats. E.g. corrupted data needs to be filtered automatically. In many cases, this is not as reliable as manual inspections, which can be applied for validation campaigns. For correct interpretation of the presented validation results, it is necessary to describe known potential caveats with the reference data in the part where the continuous validation is presented.

8.3 Validation time period and update rate

While the validation campaigns consider a fixed period of time, the continuous validation is applied on a moving time period, which is usually close to real-time. The definition of the length of this time period is usually a compromise between a sufficient sample size to generate a meaningful result and computational efforts. Thus, the length of the time period can be product dependent (e.g. a function of the product time resolution). The update rate of the continuous validation result can be used as a parameter to balance out sample size and computational efforts. An increase of the update rate usually decreases the sum of computational efforts.

8.4 Presentation of the validation results

The results of the validation are expected to be ingested directly by users on the product website. Therefore, following recommendations should be taken into account:

- Apply common and easy interpretable methods (c.f. section recommendations for validation methods e.g. scatter plot, Pearson correlation coefficient), in combination with good explanation what is shown.

- Generate plots which are well annotated with titles, axis labels, etc.
- Provision of skill scores is recommended

Since very often, the sample size for continuous validation is rather small, it is very essential to provide uncertainties along with the validation measures, because validation metrics of small samples are not as reliable as validation metrics computed from large samples. Validation results are only comparable if uncertainties in the metrics are given.

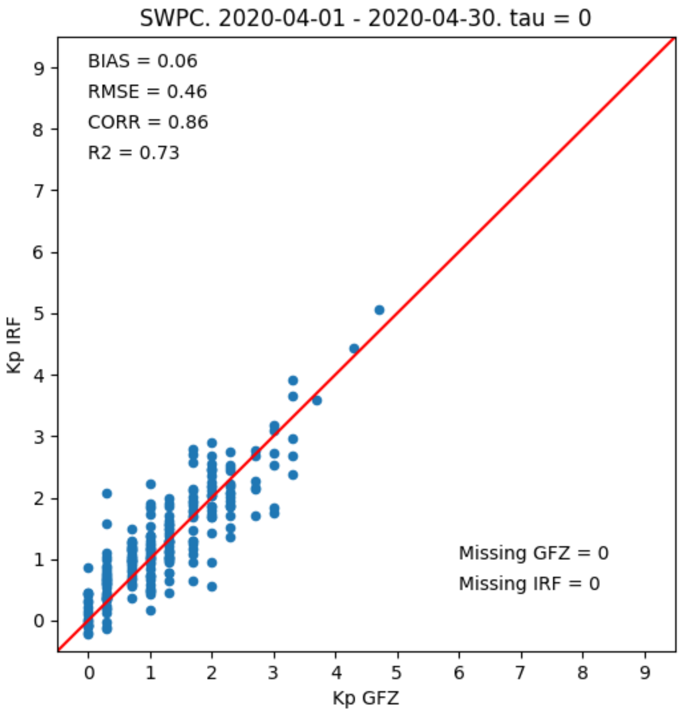


Figure 10: Example of continuous validation applied for Kp forecast provided by IRF, which is validated against GFZ Kp on a monthly basis. The plot is updated with every new Kp forecast. Skill scores are provided in the plot (c.f. http://swe.ssa.esa.int/web/guest/irf-federated).

8.5 User guidance

The aim of continuous validation is to help the user evaluating the reliability of the provided product. Since many users may not be well experienced with statistical methods, it is suggested to consider following recommendations for the provision of continuous validation results on federated websites:

- Provide descriptions how to interpret the validation results (E.g. thresholds for scores. Describe the range of values/ meaning of the metric. Reference to a validation campaign report/ result for more information on interpretation and typically expected results).
- Allow quick and easy assessment of the information
- Best way for users would be simple table reference vs. predicted. One reference is easiest to interpret for users. If there is more than one data set available, only the best reference should be chosen. If this cannot be identified, uncertainties of the reference should be indicated.

- If applicable provide thresholds when the quality of the product is considered to be good or bad
- Allow to browse within historic validation results
- Level of the product's reliability in terms of presenting possible impacts interesting for users (noting that this may be difficult to achieve due to limited information available from the end user community in terms of actual impacts experienced)

9 RECOMMENDATIONS FOR VALIDATION RESULTS DISSEMINATION

Validation tests may be executed through special campaigns or in a continuous basis. Both approaches are encouraged within the SSA SWE Network. The guidelines provided in the present document are providing recommendations for campaign-based validation and support developments for continuous validation solutions.

For the dissemination of the validation results it is essential to take into account that these results are addressed to the end users of the SSA SWE Network, the ESCs and the wider scientific community. In this respect:

- For end users' needs: **A combination of the provision of results obtained by both continuous validation and campaign results** is recommended as they would be most beneficial for users:
  - i) The continuous validation for products provides the latest information on the quality of the product. It is suggested to deliver this information as additional quality product(s) along with the product’s outcome.
  - ii) Results obtained through campaign based long-term assessment of the products performance deliver value added information in form of evaluation and interpretation of the validation results.

  It is suggested to include the validation campaign reports (or executive summaries) and publications in the **user manuals on the federated/product websites** and if appropriate in the **service usage guidance on the SSA SWE portal**.

- For the ESCs and the wider scientific community: The results obtained through campaign-based validation it is recommended to be made available through **peer-reviewed publications**. Published validation schemes will enable teams internal/external to network to compare and assess developing capabilities against equal references.