Verification of space weather forecasting at the Regional Warning Center in Belgium

Andy Devos*, Cis Verbeeck, and Eva Robbrecht

Royal Observatory of Belgium (ROB) and Solar-Terrestrial Centre of Excellence (STCE), Ringlaan 3, 1180 Brussels, Belgium

*Corresponding author: andy.devos@oma.be

Received 21 March 2014 / Accepted 11 September 2014

ABSTRACT

The Solar Influences Data analysis Center (SIDC) in Brussels at the Royal Observatory of Belgium (ROB) has been providing daily space weather forecasts for more than a decade. A verification analysis was applied to evaluate the performance of the SIDC forecasts of fundamental space weather parameters such as the F10.7 radio flux, solar flare activity, and local geomagnetic index. Strengths and weaknesses are determined compared to common numerical models. Descriptive model statistics, common verification measures, error analysis and conditional plots related to forecasts and observations are presented. The verification analysis methods have been designed such that future improvements and additions can easily be included, for example with new forecasting models. The SIDC forecast (together with the persistence model) achieves the best performance for forecasting F10.7 on day 1, but has potential for improvement for a larger lead time mainly by applying estimates from the persistence and corrected recurrence models. The persistence model is superior for the forecast of flares, though corrected recurrence models are slightly better in foreseeing M- and X-class flares and the SIDC forecast estimates B- and C-class flares very well. The SIDC forecast scores better than all models in forecasting the local K -index. It best reproduces observations in the range of $K = 2-4$, but underestimates larger K values. The SIDC forecast provides a distribution that best matches the observations of the K -index. The analysis presented here demonstrates the influence of solar activity on the confidence level of the forecasts, as well as the hinted influence of the forecaster on duty due to the subjective nature of forecasting. The output aids to identify the strong and weak points of the SIDC forecast as well as those of the models considered. Though the presented analysis needs further extension, it already illustrates the opportunity to regularly reevaluate space weather forecasts and to stimulate ideas for improvement and increase the reliability of space weather forecasting.

Key words. space weather – forecast verification – geomagnetic K -index – solar flares – solar radio flux

1. Introduction

Since 2000, the Solar Influences Data analysis Center (SIDC) in Brussels has been a Regional Warning Center (RWC) and is a member of the International Space Environment Services (ISES). The SIDC provides daily forecast bulletins describing the current solar activity (with the GOES X-ray flux as a reference), the status of the solar wind, Solar Energetic Particles (SEPs), and geomagnetic activity. It also provides a 3-day forecast of the F10.7 radio flux (also denoted as F10.7; since 2002) and forecasts of solar flare activity and geomagnetic K -index (since 2004) for the next 48 h. Besides daily bulletins, several fast email alert messages are provided. Automated email messages are sent in case of strong flares detected by the GOES X-ray instrument. The halo CME detection software, CACTus (Robbrecht et al. 2009), sends out alerts for each CME detected with an estimated angular width of at least 150° using SOHO/C2 and C3 coronagraphic data. A threshold of 150° is set as a criterion to automatically identify potentially geo-effective CMEs. In addition, a Presto message is delivered through human intervention in case of any particular space weather event with potential consequences for Earth-based systems. A daily bulletin is sent each day at 12:30 UT, the other messages are event-based.

Forecasters are available via a weekly duty cycle during working hours, but bulletins are also delivered during the

weekend. Currently eight forecasters are in the SIDC forecast team. The service is freely available and any interested person or company can subscribe to any of the SIDC products via http://sidc.oma.be/registration/registration_main.php.

A very important part of forecasting any natural event is to assess the quality or skill of the forecasts through forecast verification (<http://www.cawcr.gov.au/projects/verification/>). Verification analysis has been already applied for many decades in terrestrial weather forecasting. The first paper on weather forecast verification even goes back to 1884 (Finley 1884). For space weather the application of forecast verification is more recent; see (Berghmans et al. 2005; Balch 2008; Crown 2012) for a few examples. A few RWCs have made an effort to apply forecast verification on their forecasts and make the results available. The Space Weather Prediction Center (SWPC) at the National Oceanic and Atmospheric Administration (NOAA) has verification measures for their forecasts from 1986 on http://www.swpc.noaa.gov/forecast_verification/index.html for most of their currently provided space weather forecasts. The National Institute of Information and Communications Technology (NICT) has an online platform http://seg-web.nict.go.jp/cgi-bin/forecast/eng_forecast_score.cgi for verification analysis of four different RWCs: Japan (Tokyo, NICT), Belgium (Brussels, SIDC), Australia (Sydney, Ionospheric Prediction Service (IPS)), and US (Boulder, SWPC). The platform compares the different forecasts using

Table 1. Levels of flare activity used in forecasting. The label and the wording occur in the daily bulletin. The flare class is the level defined to compare the forecasts to observations.

Level number	Label	Wording	Flare class
1	<50% probability of C-class flares	Quiet conditions	B or lower
2	C-class flares expected, >50% probability	Eruptive conditions	C
3	M-class flares expected, >50% probability	Active conditions	M
4	X-class flares expected, >50% probability	Major flares	X

percent correctness (PC) and underestimated rate (http://seg-web.nict.go.jp/forecast/eng_forecast_data_explanation.html) as intuitive measures for accuracy. However, an analysis focused on these two individual measures does not yield a complete view of the performance of the forecasts. It is not always straightforward to compare forecasts of different RWCs directly, since forecasts of different RWCs might be valid for different time scales (e.g. forecasts for days 1–3 separately or one forecast for the next period) or might be applicable for global or local measurements (e.g. global versus local index for geomagnetic activity).

Forecast verification yields information about the bias, measuring to what extent a forecast is generally leading to over- or underestimation. The accuracy indicates the size of the forecast errors. How well binary events are forecasted can be measured by the hit rate, while skill scores quantify the accuracy of a forecast with respect to a reference model. These and other concepts were developed over the past few decades in the context of terrestrial weather forecasting (Jolliffe & Stephenson 2012) and are applied in this paper for space weather forecasting.

The benefits of space weather forecast verification are threefold. Firstly, space weather forecast verification has the potential to reveal the strong and weak points of every forecast model and to pinpoint opportunities for improvements. Secondly, different forecast models can be compared to each other, which might provide insight from other (numerical) models which would be beneficial for the SIDC forecast. Thirdly, the evaluation of forecasts can be monitored over time to detect whether, for example, the forecast improved from previously gained insights.

Section 2 describes how space weather parameters are forecasted at SIDC. Section 3 sketches the background of forecasting verification for both continuous and binary parameters and introduces the numerical models which the human forecast is compared to in this paper. The verification results for each of the forecasted space weather parameters are presented in Section 4 and discussed in Section 5.

2. Space weather parameters forecasted at SIDC

This paper focuses on three space weather parameters that are forecasted at SIDC in a daily space weather bulletin: solar radio flux, solar flare activity and geomagnetic K -index. The forecasters provide a description of the key solar events of the past period, the effects for the Earth environment, and forecast the activity for the next period. Forecasting is based on human interpretation of the latest solar data, supported by software detection tools and modeling.

The F10.7 solar radio flux (Tapping 2013) is measured daily at Penticton (Canada). It measures the total emission from the full solar disk at a frequency of 2800 MHz (wavelength of 10.7 cm), made over a one hour period centered at

12:00 local time (20:00 UT). The F10.7 has been measured since 1947 and is regarded as an important indicator for solar activity, in addition to the International Solar Sunspot Number (ISSN) (Kruger 1979). The F10.7 could be influenced by large solar flares at the moment of measurement, though actions are implemented to correct for this influence. Forecasts of the radio flux are based on a combination of data in EUV imagery and radio flux measurements at local noon in stations that have been located more to the East (e.g. the Learmonth station measures the radio flux at 5:00 UT). The latest EUV imagery is used to visually relate the total amount of solar radiation sources that is rotating away from Earth and the total amount that is rotating toward the Earth viewpoint. Without explicitly calculating, forecasters estimate whether radiation toward Earth will increase or decrease. The radio flux measurement at the Learmonth station could provide another indication of potential changes in the solar radio flux measured at the Penticton station. The estimation is typically supported by the principles of both persistence, assuming the activity will remain at the same level, and reoccurrence, estimating activity may well return after a solar rotation, when the same active regions are again facing Earth – if they have survived.

Solar flares are identified via measurements of the GOES X-ray instrument. The rank of a flare is based on its X-ray energy output. Flares are classified by the order of magnitude of the peak burst intensity I measured at Earth in the 1–8 Ångström band (<http://spaceweather.com/glossary/flareclasses.html>). An X-ray flare of class B has an intensity between 10^{-5} W/m² and 10^{-6} W/m². The flare classes C and M are up to a factor of respectively 10 and 100 times stronger than a B-class flare. An X-ray flare of class X has an intensity of at least 10^{-4} W/m². Solar flare activity is forecasted for each individual active region based on the solar flares observed during the past few days and magnetic classification of each active region. The forecaster estimates the probability of flaring from which a range of probabilities for C-, M-, and X-class flares is determined for the full solar disk. The flare forecast included in the SIDC bulletin is applicable to the full solar disk and is set to one of the levels specified in the second column of Table 1.

The interplanetary magnetic field affected by CMEs and high speed solar wind streams can disturb the Earth's magnetic field when they reach the environment of the Earth. The degree of perturbation can be quantified by the geomagnetic index K (Bartels et al. 1939). The K -index ranges from 0 to 9, with a value from 5 to 9 corresponding to geomagnetic storm conditions. At several regional observatories around the world a local K -index is determined. The planetary Kp -index is calculated as a weighted average of K -indices from a defined network of regional geomagnetic observatories. At the RWC in Belgium the local K -index from Dourbes (50.10° N, 4.58° E) is forecasted. Unfortunately, from time to time the local K -index from Dourbes contains data gaps and the local K -index from Chambon-la-Forêt (48.06° N, 2.30° E) is more

reliable. Assuming there is only a slight difference due to the small geographical distance between the two stations, in this paper we will use the local K -index from Chambon-la-Forêt for the verification analysis. The K -index is calculated every 3 h and, since June 2004, SIDC provides a forecast for these time slots in a period of the next 3 days. The forecast of the geomagnetic K -index is based on the estimated arrival time and impact of observed (partial) halo CMEs with an expected Earth-directed component. For every observed CME the source region is determined based on EUV imagery. The CME speed is either retrieved from the CACTus message, estimated using stereoscopy if available or estimated by applying CME modeling on coronagraphic imagery (e.g. [Xue et al. 2005](#); [Thernisien et al. 2006](#)). Estimates of parameters such as the source region, CME width and speed are used to determine the potential likelihood of arrival, arrival time and geomagnetic impact. The arrival time can be estimated from CME propagation models such as ENLIL ([Odstrcil et al. 2004](#)) or the drag-based model (DBM) ([Vršnak et al. 2013](#)).

The forecast process is supported by near real-time detection tools developed at SIDC, such as CACTus ([Robbrecht et al. 2009](#)) for (partial) halo CMEs, Solar Demon (<http://solar demon.oma.be>) for flares, dimmings and EUV waves and flaremail (<http://sidc.be/products/flaremail/>) for flares using GOES X-ray data.

More details on the process of the forecast can be found in [Hochedez et al. \(2005\)](#) and [Berghmans et al. \(2005\)](#).

3. Forecast verification

3.1. Forecast verification measures

A forecast is evaluated by comparing the forecast value with the observed value. This can be done for continuous as well as binary data. Binary variables can only achieve two separate values such as yes/no or 1/0, while a continuous variable as real number can achieve any possible value. An example of a binary variable is the occurrence of a storm, while a measured proton flux value is typically a continuous variable. The calculations performed for continuous variables can also be applied to integer variables.

3.1.1. Continuous or integer variables

For continuous variables, an error analysis can be performed. Several performance measures can be calculated, such as the mean error, median error, root mean squared error (RMSE), skill score, and several quantiles of the errors ([Jolliffe & Stephenson 2012](#)). The RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum (o - f)^2}, \quad (1)$$

with o and f corresponding to observation and forecast respectively and n the number of observations used. The skill score is defined in terms of the mean-squared error (MSE) as:

$$\text{skill score} = 1 - \frac{\text{MSE}}{\text{MSE}_{\text{ref}}}, \quad (2)$$

range $[-\infty, 1]$, in which $\text{MSE} = \frac{1}{n} \sum (o - f)^2$. The value MSE_{ref} is the MSE of the reference model, which is the persistence model (see [Sect. 3.2](#)) in this paper. The closer the

Table 2. Contingency table for a binary event.

		Observation	
		Yes	No
Forecast	Yes	a	b
	No	c	d

skill score is to 1, the better the model performs, with a skill score of 1 corresponding to a perfect fit. A forecast with a skill score of 0 has the same MSE as the reference model, while a negative skill score reflects a performance worse than that of the reference model. The mean error (ME) is the error in a non-relative sense:

$$\text{ME} = \frac{1}{n} \sum (o - f), \quad (3)$$

while the mean absolute error (MAE) is the mean of the absolute value of the error calculated as:

$$\text{MAE} = \frac{1}{n} \sum |o - f|. \quad (4)$$

The quantile is a value that separates an ordered sample in two parts, one part with lower values and the other part with higher values. For instance, the 25% quantile separates the lowest 25% of the sample from the 75% that has higher values. The median error, the median value of the errors or 50% quantile of the errors is the value at which 50% of the errors in the sample is lower and 50% is higher. The interquartile range (IQR) of the errors is defined as the range between the 25% and 75% quantiles (aka first and third quartile) of the errors.

3.1.2. Binary variables

It is essential to predict a geomagnetic storm well, independent of the severity. In this respect a geomagnetic storm can be treated as a binary event; it occurs or does not occur. In case of a measurement with real or integer values, an event can be defined by comparing the value with respect to a specific threshold. For the analysis of binary events, several verification measures are introduced ([Jolliffe & Stephenson 2012](#)). A contingency table is a basic concept, since it is used to define several verification measures. It is defined by the crossing of the binary observational and forecast values (see [Table 2](#)).

The total number of observations corresponds to $n = a + b + c + d$. Using a contingency table the following measures can be defined, following [Jolliffe & Stephenson \(2012\)](#):

- The number of events corresponds to $(a + c)$, while the number of non-events (or lack of events) corresponds to $(b + d)$.
- A hit is a correctly forecasted event. The number of hits is denoted as a .
- A miss is defined as an event that was not forecasted. The number of misses corresponds to c .
- A false alarm is a forecast of an event, while no event was observed. The number of false alarms is b .
- A correct rejection is a forecast of a non-event, while indeed no event was observed. The number of correct rejections is counted as d .
- The base rate of a sample is the proportion of events occurring; defined as $s = (a + c)/n$.

- The Probability of Detection (POD) or hit rate is the ratio of the number of hits and the number of events; calculated as:

$$\text{POD} = \frac{a}{(a + c)}. \quad (5)$$

- The Proportion Correctness (PC) is the ratio of total number of correct forecasts and the total number of forecasts:

$$\text{PC} = \frac{(a + d)}{n}. \quad (6)$$

- The False Alarm Ratio (FAR) is a verification measure equal to the ratio of the number of false alarms by the total number of event forecasts:

$$\text{FAR} = \frac{b}{(a + b)}. \quad (7)$$

- The Probability of False Detection (POFD) is a verification measure equal to the ratio of the number of false alarms by the total number of observed non-events (<http://www.cawcr.gov.au/projects/verification/>):

$$\text{POFD} = \frac{b}{(b + d)}. \quad (8)$$

- Bias is the degree of correspondence between the mean forecast and the mean observation; as such it indicates whether observations are overestimated (bias higher than 1) or underestimated (bias lower than 1). For categorical forecasts, bias is defined as the ratio of the number of forecasts of occurrence to the number of actual occurrences:

$$B = \frac{(a + b)}{(a + c)}. \quad (9)$$

- Heidke Skill Score (HSS) is a skill score taking into account the number of correct random forecasts.

$$\text{HSS} = \frac{(\text{PC} - E)}{(1 - E)}, \quad (10)$$

with E being the proportion of correct random forecasts, assuming forecasts and observations are independent and assuming the same proportion of forecasts of occurrence to non-occurrence. HSS has a range from -1 to 1 , with 1 for a perfect forecast, 0 as good as a random forecast and -1 for the worst forecast.

- The True Skill Statistic (TSS) is a verification measure of categorical forecast similar to HSS. The TSS has desirable properties for rare event forecasts; for example random and constant forecasts get a TSS of 0 . The TSS ranges from -1 to 1 , with 1 a perfect forecast and 0 representing no skill. For extremely rare events, TSS tends to converge to POD. TSS is calculated as:

$$\text{TSS} = \frac{(ad - bc)}{(a + c)(b + d)} = \text{POD} - \text{POFD}. \quad (11)$$

The TSS is also called the Peirce Skill Score or the Hansen and Kuipers' Score. TSS and HSS have very favorable statistical characteristics for the verification of forecast models (Jolliffe & Stephenson 2012). A main difference between TSS and HSS is that TSS has the property being independent of the base rate, while HSS treats misses and false alarms equally.

Many other verification measures for binary variables are based on the above-mentioned numbers. Several of the above listed verification measures can be generalized to ordinal, multi-categorical parameters, such as the K -index. Such parameters have a number of discrete categories, with a specific order. However, in this paper we restrict ourselves to the common verification measures with favorable properties defined for continuous and binary parameters. The bias is the only specific verification measure for multi-categorical parameters that is applied in this paper. Assume a parameter with a sample probability distribution p_i and $\hat{p}_i$ of the observations and forecasts, respectively, with $i = 1, \dots, L$ and $L \geq 2$ the amount of possible values. The distribution p_i contains the probabilities of observing category i with $i = 1, \dots, L$, while $\hat{p}_i$ contains the probabilities for the forecast. Then for example the bias for multi-categorical parameters is defined as:

$$B_i = \frac{\hat{p}_i}{p_i} \quad \text{for } i = 1, \dots, L. \quad (12)$$

For further discussion on the properties of these and other measures, we refer to Jolliffe & Stephenson (2012) and references therein.

3.2. Forecast models

To evaluate the SIDC forecast of the space weather parameters, numerical models were applied to forecast the same parameters. The performance of the SIDC forecast was then compared to that obtained by the following simple numerical models, which reflect principles often implicitly applied when forecasting:

- The persistence model assumes solar activity will stay at the same level as yesterday. The space weather parameter of today is estimated to obtain exactly the same value as yesterday.
- The recurrence model focuses on the influence of the Carrington rotation (CR) of on average 27 days. The recurrence model assumes the active regions are as active as a full rotation ago and estimates the space weather parameter to obtain the same value as 27 days ago. In addition a recurrence model with a time shift of 14 days, which is half a rotation, was tested. The argument is that the activity pattern can be related to active longitudes (Berdyugina & Usoskin 2003).
- The corrected recurrence model seeks a compromise between the persistence and recurrence models. This model looks at the daily increment of 27 days ago and applies this to the value of yesterday. Similarly as for the recurrence model, the corrected recurrence model was applied with a time shift of 14 days and 27 days.
- Linear interpolation was also applied for the solar radio flux. A linear fit was estimated based on the values of the past 4 days and then applied to estimate the value of the next 3 days.

Fig. 1. Evolution of observations (a) and errors of the daily SIDC forecast (b) for days 1–3 of the F10.7 radio flux during year 2012.

3.3. Forecast verification at SIDC

The daily space weather forecast at the SIDC is carried out at 12:30 UT. The solar radio flux is forecasted for days 1–3 with day 1 being the UT day of the forecast. Hence, verification analysis is carried out for the F10.7 on days 1–3 separately.

The solar flare activity forecast is valid for the next 48 h from the moment of forecasting on (12:30 UT). In this analysis, the levels mentioned in the daily bulletin (see Table 1) are converted to the flare classes B, C, M, and X and as such are forecasted to be the highest flare level occurring in the next period. Verification analysis is applied by comparing this forecast to the maximum observed flare class across the same period.

Similar to the case of the flares, the forecast of the K -index is valid for the next 48 h. The K -index is forecasted for each time slot of 3 h. For the verification analysis the maximum K -index is calculated from these 3-hourly values over the next 48 h from 12:00 UT onwards. Unfortunately, the local K -index from Dourbes regularly has data gaps. The local K -index from Chambon-la-Forêt (48.06° N, 2.30° E) is more reliable. In order to use as many data as possible, the observed K -index from Chambon-la-Forêt is used for the verification analysis. This implies observations and numerical forecast models are based on the K -index from Chambon-la-Forêt, which can be justified by the small difference in the geographical coordinates of the two stations.

In the verification analysis described in this paper, data from 2002 till end of 2012 are included for the F10.7 radio flux and from June 2004 till end of 2012 for solar flares and the K -index. Throughout the present paper the term forecast is referring to a forecast in general, while SIDC forecast is referring to the human forecast provided by the SIDC.

The persistence, recurrence, and corrected recurrence models were applied for the estimation of solar radio flux, solar flares and geomagnetic index. Linear interpolation was only applied for the solar radio flux. The values obtained by linear

interpolation were rounded to the closest integer value. The estimated values were bound to the feasible range, for all models, except the persistence and recurrence models which achieve this already by definition. As a result the estimated radio flux obtains values of at least 64 sfu, the estimated solar flare activity integers 1–4, corresponding to B to X-class flares and the estimated K -index integers 0–9. The results presented in this paper are available on <http://www.sidc.be/forecastverification>, where also future results will be positioned.

4. Results

4.1. F10.7

For the F10.7 radio flux, basic analysis of observations versus forecast has been performed for data from the years 2002–2012. Figure 1 provides an example of the daily variability of solar activity by the solar radio flux as well as the parallel evolution of the errors of the SIDC forecast for days 1–3. Data from year 2012 are displayed to focus on the variability during one single year. The influence of the solar rotation of 27 days is prominent in the second half of the year, while in the first half of the year this reoccurrence is disturbed during the months February and April. The differences between the forecast for days 1–3 illustrate the tendency of the forecasters to implicitly apply the principle of persistence. The largest errors for day 3 typically follow a large peak in the observed F10.7, for example at mid July 2012.

Figure 2 displays the size (panel a) and the bias (panel b) of the errors pertaining to the different forecast models. The mean absolute error (colored bars) on day 1 is lowest for the SIDC forecast and persistence model. The skill score (crosses) is slightly better for the SIDC forecast on day 1, but is worse on days 2 and 3. The corrected recurrence models have a worse performance than the persistence model, but the skill score gets slightly better for days 2 and 3. The bottom panel shows the

Fig. 2. Mean size of errors for the forecast of the radio flux. The plots include results for the SIDC forecast and the numerical models persistence, recurrence and corrected recurrence of 14 and 27 days and linear fit. Errors are displayed separately for days 1–3. The left vertical axis displays the error (colored bars) and the right vertical axis the skill score (Eq. (2), crosses) for each of the forecast models with respect to the persistence model as reference model. Panel (a): mean absolute error (Eq. (4)). Panel (b): mean error (Eq. (3)).

mean error, which is positive for the SIDC forecast and recurrence models, while the mean error for the persistence, linear fit and corrected recurrence models, is close to zero (either positive or negative). The SIDC forecast is the best model (together with persistence) for day 1. For days 2 and 3 both the persistence and corrected 27 day recurrence models have the lowest mean errors and highest skill scores. These models also only have a limited bias, while the SIDC forecast has a positive bias between 0.2 (for day 1) and 0.7 (for day 3), which is small compared to the flux value. A positive bias indicates that the SIDC forecast has the tendency to more often overestimate than underestimate.

In order to investigate how the errors of the various forecast models vary and which one performs best as a function of the observed value, Figure 3 displays the errors, conditional on the observed flux values. The following models are included: persistence, corrected recurrence of 14 and 27 days and linear fit model. The interpretation of the plot is illustrated with the following example. Assume a flux value of 150 sfu is observed. The top three panels (for days 1, 2 and 3) show that the median error is about 0 sfu, which means that half of the times the error is lower or equal to 0 sfu, while the other half has a positive error. The fourth to sixth panels show the interquartile range (IQR) of the errors for days 1, 2 and 3. The plot for

Fig. 3. Errors conditional on the observed flux values. The plot includes results for the SIDC forecast and the numerical models persistence, corrected recurrence of 14 and 27 days and linear fit. The three top panels show the median error (Sect. 3.1.1) conditional on the observed flux for days 1–3 respectively. The three following panels display the interquartile range (Sect. 3.1.1, IQR) of the error conditional on the observed flux for days 1–3 respectively. The bottom panel shows the histogram of the observed values.

day 1 indicates an IQR of about 10 sfu, which means that half of the errors lie e.g. in the range of $150 \pm \frac{10}{2} = [145, 155]$ sfu. However this range may as well be asymmetric; e.g. [142, 152] sfu. The same principle for an observed flux of 200 sfu results in a range of e.g. $200 \pm \frac{20}{2}$ sfu.

On day 1, the SIDC forecast and persistence perform on average best with only limited error increases with larger

F10.7 values. For F10.7 values below 200 sfu, the IQR of their errors remains below 20 sfu. On days 2 and 3, persistence and corrected 27 day recurrence models perform best for all observed F10.7 values. For F10.7 values below 150 sfu (day 2) or 100 sfu (day 3), the IQR of their errors are below 20 sfu. In case of the highest observed F10.7 values, the corrected 27 day recurrence model scores the best on days 2 and 3.

Fig. 4. Matrix plot with the distributions and color grid for the different forecast models of the flare probabilities. *Top panel:* color grid for the different forecasts (SIDC, persistence, corrected recurrence of 14 and 27 days). The predicted flare level is on the horizontal axis (x-axis), while the observed one is on the vertical axis (y-axis). The more greenish or reddish the colored square, the more frequent the combination of forecasted and observed flare level occurs. Renormalized version where the sum of the probabilities across the whole grid is 1. The color code on a logarithmic scale is shown on the right side of the panel. *Second panel:* similar as top panel. Renormalized version where the sum of the probabilities along the x-axis is 1. *Third panel:* similar as top panel. Renormalized version where the sum of the probabilities along the y-axis is 1. *Bottom panel:* histograms of the different forecast models and observations on a logarithmic scale.

4.2 Solar flares

The flare forecasts were analyzed for the period June 2004–December 2012. Figure 4 displays color grids of selected models crossed with the observations, using different normalizations. The color bars at the right indicate to which value each color corresponds. The top panel shows the color grid using full normalization, meaning the sum of the values corresponding to the colors across the full grid is 1. The values themselves can be regarded as probabilities of occurrence of each observation-forecast pair.

The second panel displays the probabilities conditional on the observational value. This means it provides information on the most probable forecasted flare level for a given maximal observed flare level. For example on days where the maximal observed flare level was B-class (row corresponding to level 1), the probabilities that SIDC forecasts a B-, C-, M-, or X-class flare are displayed in the cells with levels 1–4 respectively of the row corresponding to level 1.

An analogous interpretation is valid for the third panel, except that the colors here indicate probabilities conditional on the forecast. In case M-class flares are forecasted, M-class flares are most often observed, but also observations with C- and X-class flares as maximal flare activity occurred, though less likely. The red diagonal line through each color grid can be used to detect the general tendency. The more the greenish and reddish colored cubes (i.e. high probabilities)

lie on the diagonal, the more observations and forecasts coincide. In addition, the more greenish and reddish colored cubes lie above (below) the diagonal, the more the forecast is underestimating (overestimating) the observed flare activity. Based on Figure 4, the following remarks can be made:

1. Though the histogram of the observations does not closely match any of the model histograms, the SIDC forecast and persistence model generate a smoother distribution around the diagonal in the color grids than the corrected recurrence models. This is in correspondence to the correlation values: 0.64 and 0.63 for SIDC forecast and persistence model respectively and 0.51 and 0.52 for corrected recurrence models of 14 and 27 days respectively.
2. Looking at the distributions in the lower panel of Figure 4, reveals that all forecasts result in an overestimation of B-class flares. The corrected recurrence and persistence models underestimate the occurrence of C-class flares, while the SIDC forecast returns an occurrence of C-class flares which resembles most the observations. The SIDC forecast and persistence model mainly underestimate the M- and X-class flares. Only the corrected recurrence models have a percentage of X-class flares in line with the observations.
3. In case of an observed C-class flare (using probabilities conditional on observations), the SIDC forecast has most often correctly forecasted a C-class flare. In case of an

M- or X-class flare more often a lower flare level was forecasted, while then the persistence and corrected recurrence models more often forecasted the correct flare level.

To evaluate the fit between each forecast and the observations the flare class level was also treated as a continuous parameter, enabling to compute the RMSE and skill score with respect to the persistence model. The SIDC forecast reached the highest skill score as the following scores were obtained: 0.05, 0, -0.43 , -0.40 for the SIDC forecast, persistence and the corrected 14 and 27 day recurrence models respectively. The SIDC forecast obtains a high skill score due to its high performance in predicting the more likely B- and C-class flares.

In a next step, the flare level was dichotomized by combining M- and X-class flares (further called event), while B- and C-class flares were regarded as a non-event. Figure 5 shows the skill scores across the years and for each individual forecaster. In 2004 many hits and only a limited proportion of misses and false alarms were obtained, but in the following years the proportion of misses was higher than the proportion of hits. Table 3 indicates that the SIDC forecast has a lower POD than the persistence and corrected recurrence models. The bias indicates that the corrected recurrence models tend to overestimate nor underestimate the amount of strong flares, while the SIDC forecast and persistence model in general tend to provide underestimations. This is in correspondence to the HSS and TSS skill scores. At a base rate $s < 1/2$ (which is the case here), the TSS treats underestimating models (i.e. with a low bias) more harshly than HSS and overestimating models more generously. This is reflected in the fact that the underestimating SIDC forecast and persistence model obtain a relatively lower TSS value, while the HSS values are better. The persistence model proves to be the best overall model for forecasting flare activity. Indeed, flare activity can be eruptive and hence its exact timing with the current knowledge can not be predicted, but the flare class may well fit with the persistence model. Most of the time, the Sun can be described as either eruptive yesterday and today, or quiet yesterday and today. In only a minority of cases does the flare level change from one day to the next. Figure 4 aids to verify this correspondence between the observations and the persistence model. The grid of the persistence model in the top panel displays the probabilities normalized across the full grid. The sum of the corresponding values on the diagonal is the total probability that the observations follow the persistence model. The sum of the values above the diagonal is the probability that the persistence model underestimates the flare level, while the sum of the values below the diagonal is the probability to overestimate. The persistence model underestimates the observations if the flare activity is increasing, while an overestimation occurs in case of decreasing activity. Indeed, in 66% of cases the observed flare level is forecasted by using persistence, while in 24% cases the flare level increased and in the remaining cases it decreased.

The SIDC forecast achieves a good performance on the FAR and HSS scores, in line with the tendency to underestimate. The TSS and POD scores are not much lower than for the persistence and corrected recurrence models.

The verification results per forecaster depend on the forecaster on duty, but also the level of activity has a substantial influence on the forecast performance. In periods of high activity it typically is more challenging to forecast the flare level,

since it is often unclear if and when complex active regions will become unstable and produce large flares. In case of limited amount of active regions, flare levels are assumed to remain quite low. Panel *b* of Figure 5 compares the scores of the SIDC forecast to that of numerical models for each individual forecaster. The verification results for forecasters with less than 50 days on duty (forecasters with number 8, 9, 11, and 13) are not reliable, as for example the event proportion of 100% for forecaster 13 is only based on 8 forecasting days. Hence, the results of these forecasters will be neglected in the summary below. In this interpretation the scores are compared relative to that of the numerical models. Since the persistence model achieves overall the highest skill scores (Table 3), the difference $TSS_{SIDC} - TSS_p$ between TSS_{SIDC} and TSS_p , respectively the TSS values of the SIDC forecast and the persistence model is adopted as a criterion to compare the performance of the forecasters.

1. Forecasters with number 0 and 14 obtain very high skill scores (PC, POD, HSS, TSS) with respect to numerical models. Forecaster 14 had lower scores than forecaster 0, but also the scores of the numerical models were much lower (with $TSS_{SIDC} - TSS_p = 0.22$), indicating the forecast job in general was more challenging. Forecasters 0 and 14 indeed are among the team members with a very strong background in solar physics.
2. Forecasters 1 and 6 also obtain scores which are very competitive with those of the numerical models, although slightly worse than for forecasters 0 and 14. Also forecasters 3 and 10 and, to a lesser extent, 2 and 12 obtain moderate skill scores close to the numerical models.
3. Forecasters 0 and 3 reach a very good bias, close to 1. To a lesser extent this holds for forecasters 6, 9, and 14. All other forecasters underestimate the flare activity more than the corrected 27 day recurrence and the persistence models do.

4.3. *K*-index

Basic analysis of observations versus forecast has been performed, using the *K*-index in the period June 2004–December 2012, though with some data gaps. Figure 6 displays the color grids to the observations for the same models as shown for the flares. No $K = 0$ has been covered in the histogram for the observations, the SIDC forecast and the corrected 27 day recurrence model. This is due to the fact that for the observations and these forecasts, $K = 0$ was never the maximum level across the next 48 h at the moment of forecasting. $K = 0$ does occur for the corrected 14 day recurrence and especially the persistence models.

Based on the color grids for the SIDC forecast (left column), the following remarks can be made:

1. Correlations between the forecasts and observations are most favorable for the SIDC forecast: 0.53, 0.18, 0.31 and 0.30 for the SIDC forecast, the persistence model, corrected recurrence models of 14 and 27 days, respectively.
2. For the rare cases with observed *K*-index $K_o = 1$ (6%) (second panel), a forecasted *K*-index $K_f = 2$ is more often given and sometimes $K_f = 3, 4$ are forecasted. Similarly a $K_f = 1$ (third panel) often corresponds to $K_o = 2$ or 3. This means it is difficult to predict the level of quiet conditions correctly.

forecaster	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
# days on duty	391	406	65	385	193	256	316	151	28	7	208	39	559	8	75
bias	0.9	0.3	0.5	1.0	0.4	0.2	0.7	0.3	0.4	0.7	0.5	NaN	0.5	0.8	0.8
$TSS_{SIDC} - TSS_p$	0.04	0.02	-0.12	-0.02	-0.18	-0.17	0.02	-0.32	-0.10	0.25	-0.05	NaN	-0.09	NaN	0.22

Fig. 5. Evaluation of the SIDC forecast and models to predict an X-ray flare of classes M or X, across the years (a) and for each individual forecaster (b). The following description holds for both plots. *Top panel:* POD (lines, left vertical axis, Eq. (5)) and HSS (bars, right vertical axis, Eq. (10)). *Middle panel:* PC (lines, left vertical axis, Eq. (6)) and the TSS (bars, right vertical axis, Eq. (11)). POD and PC are shown for all forecasts, while HSS and TSS are only for the SIDC forecast. *Bottom panel:* the proportion of days with an event, a hit, a miss and a false alarm (on a logarithmic scale). The proportion of hits and the proportion of misses sum up to the proportion of events. The number of days on duty, the bias and the difference of the TSS with the TSS of the persistence model are mentioned for each forecaster in the table below the figure. Note that some verification measures can not be calculated and are left undefined in case of lack of events or hits (e.g. year 2009 and forecaster 11).

Table 3. Skill scores for the SIDC forecast and numerical models for the forecast of flare events of classes M or X.

	SIDC forecast	Persistence	Recurrence 14 days	Recurrence 27 days	Corrected recurrence 14 days	Corrected recurrence 27 days
<i>n</i>	3119	3119	3119	3119	3119	3119
<i>a</i>	150	176	62	65	184	68
<i>b</i>	82	99	215	214	201	202
<i>c</i>	250	224	338	335	216	214
<i>d</i>	2637	2620	2504	2505	2518	2517
Hits (%)	4.81	5.64	1.99	2.08	5.90	5.96
False alarms (%)	2.63	3.17	6.89	6.86	6.44	6.48
Misses (%)	8.02	7.18	10.84	10.74	6.93	6.86
Correct rejections (%)	84.55	84.00	80.28	80.31	80.73	80.70
Events (%)	12.82	12.82	12.82	12.82	12.82	12.82
POD (%)	37.50	44.00	15.50	16.25	46.00	46.50
FAR (%)	35.34	36.00	77.62	76.70	52.21	52.06
PC (%)	89.63	89.64	82.27	82.40	86.63	86.66
Bias	0.58	0.69	0.69	0.70	0.96	0.97
HSS	0.42	0.47	0.09	0.10	0.39	0.40
TSS	0.34	0.40	0.08	0.08	0.39	0.39

- For a K -index of $K = 2, 3, 4$, the highest conditional probabilities lie on the diagonal, both when conditional on the observations (second panel) and when conditional on the forecast (third panel). Forecasted and observed K -index often coincide for these most frequent levels of geomagnetic activity. If for example the K -index $K_o = 4$, then the most probable forecasted index values by SIDC are $K_f = 3, 4$, while also $K_f = 1, 2, 5, 6$, and 7 are also forecasted but with lower (conditional) probability. All other values ($K_f = 0, 8, 9$) have hardly or not been forecasted in case of $K_o = 4$.
- Cases with $K_o = 5, 6, 7$ often have a lower K -index forecasted and so the observed K -index is underestimated by 1–3 index points. On the other hand, the third panel shows errors also occur the other way around as $K_f = 5–7$ can also correspond to a lower observed K -index.
- K -indices higher than 7 are rare such that they are even more difficult to forecast. Such values have a few times been correctly estimated but apparently more often underestimated.

The performance of the numerical models (columns two to four in Fig. 6) is worse than that of the SIDC forecast. The color grids of the numerical models reveal a large discrepancy between the observed K -index and the K -index forecasted by the numerical models. There is a larger underestimation of the K -index from $K \geq 4$ when based on the persistence model. The rare cases in which a high K -index was forecasted often lead to an overestimation, again most striking for the persistence model. The distribution of the SIDC forecasted K -index is most similar to that of the observations, while the persistence model is more skewed to a lower K -index and the corrected recurrence models have the tendency to more often predict a higher K -index. In more detail, based on the bias for multi-categorical parameters (Eq. (12)) and also visible in Figure 6, the SIDC forecast tends to slightly overestimate a K -index of 2–4 and underestimate K -indices of 5 and 6. The persistence model overestimates $K = 0–2$ and underestimates $K = 3–9$, while the corrected recurrence models underestimate $K = 1–3$ and overestimate $K = 4–9$. Overall the SIDC forecast color grid has the most smooth and symmetric distribution around the diagonal.

Similar to the analysis for flares, the forecast of the K -index was also evaluated using a skill score based on the RMSE. Again, the SIDC forecast outperforms the numerical models as following scores were obtained: 0.52, 0, 0.18, 0.13 for the SIDC forecast, persistence and the corrected 14 and 27 day recurrence models respectively. Figure 7 displays the mean errors and mean absolute errors of the applied forecasts. The SIDC forecast obtains the lowest mean absolute error and has almost no bias, while persistence and (uncorrected) recurrence models on average underestimate and the corrected recurrence models overestimate the K -index.

Next we dichotomized the K -index based on a threshold of $K = 5$: any $K \geq 5$ was coded as a geomagnetic storm (further also called event), while any other was regarded as a non-event. Figure 8 displays several skill scores for the forecast provided by SIDC and estimated by some numerical models.

The POD and PC are intuitive measures, but these scores clearly depend on the base rate. In case of very few events, for example in 2009 around solar minimum, the SIDC forecast has a very low POD, since it is very hard to predict these rare events. The PC was very high, since most non-events were predicted correctly. Hence, solely using the PC could provide a false impression. The lower panel of Figure 8 indicates that many events were missed in 2005 and several of the following years, while proportionally this happened less in 2004 which is reflected in the POD. Also in other years such as 2011 and 2012 many misses occur, but less false alarms were made, which returns a moderate PC.

Table 4 lists several verification measures for the different forecasts. The SIDC forecast has higher scores, including on HSS and TSS, than the persistence and corrected recurrence models. The bias indicates that the SIDC forecast and persistence model are more often underestimating, while the corrected recurrence models are overestimating. This is in line with remarks made above on Figure 6. The difference between the models is less prominent for TSS than for HSS. Due to the properties of HSS and TSS, the overestimating corrected recurrence models obtain a relatively high TSS value, while the HSS values are worse (see Table 4). Though the SIDC forecast tends to frequently underestimate the high values of the K -index, it is outperforming the numerical models in

Fig. 6. Matrix plot with the distributions and color grid for the different forecast models of the K -index. *Top panel:* color grid for the different forecasts (SIDC, persistence, corrected recurrence of 14 and 27 days). The predicted K -level is on the horizontal axis (x -axis), while the observed one is on the vertical axis (y -axis). The interpretation of all subplots is analogous to that of Figure 2. The more greenish or reddish the colored square, the more frequent the combination of forecasted and observed flare level occurs. Renormalized version where the sum of the probabilities across the whole grid is 1. The color code on a logarithmic scale is shown on the right side of the panel. *Second panel:* similar as top panel. Renormalized version where the sum of the probabilities along the x -axis is 1. *Third panel:* similar as top panel. Renormalized version where the sum of the probabilities along the y -axis is 1. As for $K_f = 0$ of the corrected 14 days recurrence model only $K_o = 2$ is observed, this conditional probability is 1 (red square). *Bottom panel:* histograms of the different forecast models and observations on a logarithmic scale.

Fig. 7. Mean size of errors for the forecast of the local K -index. The plots include results for the SIDC forecast and the numerical models persistence, recurrence and corrected recurrence of 14 and 27 days. The left vertical axis displays the error (colored bars) and the right vertical axis displays the skill score (crosses) for each of the forecast models with respect to the persistence model as reference model. The mean errors (red bars, Eq. (3)) as well as the mean absolute errors (green bars, Eq. (4)) are displayed.

Fig. 8. Evaluation of the SIDC forecast and numerical models to predict a geomagnetic storm with K of at least 5, across the years. *Top panel:* POD (lines, left vertical axis, Eq. (5)) and HSS (bars, right vertical axis, Eq. (10)). *Middle panel:* PC (lines, left vertical axis, Eq. (6)) and the TSS (bars, right vertical axis, Eq. (11)). POD and PC are shown for all forecasts, while HSS and TSS only for the SIDC forecast. *Bottom panel:* the proportion of days with an event, a hit, a miss and a false alarm (on a logarithmic scale). The proportion of hits and the proportion of misses sum up to the proportion of events. Note that some or all verification measures cannot be calculated and are left undefined in case of lack of events or hits (e.g. year 2009).

Table 4. Skill scores for the SIDC forecast and numerical models for the forecast of a geomagnetic storm ($K \geq 5$).

	SIDC forecast	Persistence	Recurrence 14 days	Recurrence 27 days	Corrected recurrence 14 days	Corrected recurrence 27 days
n	2985	2985	2985	2985	2985	2985
a	115	37	23	30	89	101
b	130	149	163	156	352	373
c	207	285	299	292	233	221
d	2533	2514	2500	2507	2311	2290
Hits (%)	3.85	1.24	0.77	1.01	2.98	3.38
False alarms (%)	4.36	4.99	5.46	5.23	11.79	12.50
Misses (%)	6.93	9.55	10.02	9.78	7.81	7.40
Correct rejections (%)	84.86	84.22	83.75	83.99	77.42	76.72
Events (%)	10.79	10.79	10.79	10.79	10.79	10.79
POD (%)	35.71	11.49	7.14	9.32	27.64	31.37
FAR (%)	53.06	80.11	87.63	83.87	79.82	78.69
PC (%)	88.71	85.46	84.52	84.99	80.40	80.10
Bias	0.76	0.58	0.58	0.58	1.37	1.47
HSS	0.34	0.07	0.01	0.04	0.12	0.14
TSS	0.31	0.06	0.01	0.03	0.14	0.17

estimating the occurrence of a geomagnetic storm. In addition, it even has the lowest false alarm ratio.

5. Discussion

For the F10.7 radio flux, the SIDC forecast has overall the highest skill score for day 1, while it is worse for days 2 and 3. The errors get larger for larger lead times for each of the models, except for the (uncorrected) recurrence models. Forecasters base their flux estimate on the properties and

evolution of active regions, such as the number of spots, area and position on the solar disk.

The solar flux observed from the front side of the solar disk is expected to increase due to active regions rotating around the east limb and decrease due to active regions rotating over the west limb. However it is difficult to foresee the evolution of active regions, probably leading to the largest errors in human forecasting. Indeed, flux values might change a lot over the course of a few days due to the development, growth, and decline of specific active regions. Flux emergence

and cancellation observed in magnetograms typically lead to a change in radio flux, which was potentially not expected a few days before while forecasting.

The persistence, corrected recurrence, and the linear fit models have a relatively low error for days 2 and 3. Especially for the persistence and corrected recurrence model of 27 days, the errors are larger for days 2 and 3 than day 1, but the difference is less prominent than for the SIDC forecast. A potential explanation is that the corrected recurrence model combines observations of yesterday as well as of 27 days ago with respect to the target day. Inherently, the corrected recurrence model uses the changes of one rotation ago to estimate the changes for the next few days. For the observed flux values above 150 sfu the corrected recurrence model for 27 days roughly predicts best the value for day 3. The corrected recurrence models perform much better than the uncorrected recurrence models. Exploiting the principle of the persistence and corrected recurrence models for days 2 and 3 could be beneficial for the SIDC forecast. To illustrate, blind usage of the persistence model would result in an average reduction of the errors in absolute value with 1 to 2 sfu for days 2 and 3. However the challenge would be to specify these results in more detail and to translate them to explicit and practical forecast guidelines. The uncorrected recurrence model has the highest errors, but is not dependent on the lead time. This might be due to low changes in flux values from one day to the next, which is logical since we are using a recurrence term of 27 days (or 14 days) with respect to the target date.

The persistence model achieves in general the best scores for the forecast of large flares. The SIDC forecast for flares provides a low false alarm ratio and a high percentage correctness and HSS, but does worse on the detection of events (lower POD) compared to the persistence and corrected recurrence models. From these first results, there is an indication that an increase up to 9% in the detection rate of the strong flares could be achieved by blindly following the corrected recurrence models. However, this would probably also lead to substantially more false alarms. Many strong flares were missed by the SIDC forecast and only in 2004 more hits than misses were obtained. A striking aspect is that this trend to a lower probability of detection also occurs for the numerical models, especially during the years of the solar minimum. Further analysis will be needed to understand reasons for this.

The SIDC forecast of the geomagnetic K -index has across the years higher skill scores than all numerical models. As the applied numerical models are only depending on the past values of the K -index, they do not take into account the properties of the solar wind and any CMEs that erupted in the past days. These aspects are typically monitored by forecasters when estimating K -index for the next forecast period. The (corrected) recurrence models (with a recurrence time of 27 days) have the potential to deal with the effects of recurrent coronal holes by determining the K -index using the value of a full rotation ago. This probably aids the (corrected) recurrence models to forecast K -index values relatively well in the range of 3–6, which is observed as a consequence of the arrival of a high speed stream related to recurrent coronal holes. However, further detailed investigation of past events is required to confirm the key factors leading to the most common errors in predicting geomagnetic storms.

6. Conclusions and next steps

For each of the forecasted space weather parameters, the strengths and weaknesses were identified in comparison to simple numerical models and several opportunities for improvement were revealed. However this is preliminary, since additional and deeper analysis is required to fully understand the current results.

The SIDC F10.7 forecasts can gain accuracy by using the persistence model for days 2 and 3, except when observed flux values exceed 150 sfu on days 2 and 3. In the latter case, the corrected 27 day recurrence model is advised (see Fig. 3). Strong flares are well forecasted by the SIDC forecast, but the persistence model outperforms any of the discussed forecast models, indicated by almost all skill scores. Also the corrected 27 day recurrence model has a high ability to foresee flares, reflected by the high POD. Especially applying the persistence model more often could aid the forecast of strong flares, though it returns errors when the maximum flare class changes from one day to the next. The SIDC forecast outperforms the numerical models for the forecast of geomagnetic storms, though there is a tendency to underestimate the K -index.

In order to gain improvement from the above results, we could implement (conditional) error bars on our forecast. Once there is an agreement on the most essential analyses, these will be rerun on a regular basis in order to continuously reevaluate the SIDC space weather forecasting and track its performance. A comparison of similar situations with correct versus erroneous forecast might reveal underlying factors that explain the reasons for some discrepancies. For example certain geomagnetic storms are predicted very well, while for others a large underestimation is obtained. To pinpoint the common causes, such as potential deflection or interaction of CMEs, would be very useful. A separate analysis on the forecast of geomagnetic storms due to coronal holes and CMEs seems worthwhile.

For the multi-categorical parameters such as the flare activity and K -index, a generalization of the skill scores could be applied. More specifically a generalization of HSS, TSS, and other skill scores can provide an objective comparison of the forecast methods without losing information by regarding flare and geomagnetic activity only as binary events or continuous parameters. An appropriate weighted skill score (other than one based on the RMSE) might be advised that also reflects the ability to forecast more extreme events.

For X-ray solar flares, probabilities of the individual flare classes are computed during the daily forecast. These probabilities are estimated based on the forecaster's judgment of flaring chances of each currently visible sunspot group. The probabilities serve as a forecasting guideline to define the level of forecasted flare activity. The evaluation of these flare probabilities is in the pipeline and would possibly aid to provide more granularity in the flare forecast. In the analysis described in this paper, only simple numerical models were used. It would be interesting to see how the SIDC forecast performs compared to more sophisticated numerical models or forecasts of other RWCs. In order to make such analysis, forecasts need to be treated using common conditions, such as period of forecast, possible levels of activity, using maximum or mean forecast.

A selection of forecasters obtains a higher performance than others in forecasting strong solar flares. These forecasters typically obtain a bias closer to 1, indicating they are better in notifying early signatures of upcoming strong flare activity. An extension of this analysis to the forecast of the geomagnetic activity and radio flux is planned. Detailed analyses should reveal the factors for the forecaster dependency of the skill scores. Such exercise should provide learnings to better and more objectively streamline the forecast procedure at SIDC. Especially, the forecast of high flare activity and strong geomagnetic storms remains a challenge. Elaborating and extending further the discussed analyses should provide more insight in the value of specific simple and more advanced approaches for space weather forecasting.

Acknowledgements. This work has received funding from the European Commission FP7 Projects AFFECTS (263506) and COMES-SEP (263252). The authors thank the SIDC forecasters for discussions and their colleague Daniel Ryan for proofreading the manuscript. The editor thanks Larisa Trichtchenko and Rodney Viereck for their assistance in evaluating this paper.

References

- Balch, C.C., Updated verification of the Space Weather Prediction Center's solar energetic particle prediction model, *Space Weather*, **6** (1), DOI: [10.1029/2007SW000337](https://doi.org/10.1029/2007SW000337), 2008.
- Bartels, J., N.H. Heck, and H.F. Johnston, The three-hour range index measuring geomagnetic activity, *Geophys. Res.*, **44**, 411–454, DOI: [10.1029/TE044i004p00411](https://doi.org/10.1029/TE044i004p00411), 1939.
- Berdyugina, S.V., and I.G. Usoskin, Active longitudes in sunspot activity: Century scale persistence, *A&A*, **405**, 1121–1128, DOI: [10.1051/0004-6361:20030748](https://doi.org/10.1051/0004-6361:20030748), 2003.
- Berghmans, D., R.A.M. Van der Linden, P. Vanlommel, R. Warnant, and A.N. Zhukov, et al., Solar activity: nowcasting and forecasting at the SIDC, First European Space Weather Week (ESTEC, Noordwijk November 2004), *Ann. Geophys.*, **23** (9), 3115–3128, DOI: [10.5194/angeo-23-3115-2005](https://doi.org/10.5194/angeo-23-3115-2005), 2005.
- Crown, M.D., Validation of the NOAA Space Weather Prediction Center's solar flare forecasting look-up table and forecaster-issued probabilities, *Space Weather*, **10**, 6006, DOI: [10.1029/2011SW000760](https://doi.org/10.1029/2011SW000760), 2012.
- Finley, J.P., Tornado predictions, *Am. Meteorol. J.*, **1**, 85–88, 1884.
- Hochedez, J.-F., A. Zhukov, E. Robbrecht, R. Van der Linden, D. Berghmans, P. Vanlommel, A. Theissen, and F. Clette, Solar weather monitoring, *Ann. Geophys.*, **23**, 3149–3161, DOI: [10.5194/angeo-23-3149-2005](https://doi.org/10.5194/angeo-23-3149-2005), 2005.
- Jolliffe, I.T., and D.B. Stephenson, Forecast verification. A practitioners Guide in Atmospheric Science, 2nd edn., Wiley-Blackwell, 2012.
- Kruger, A., Introduction to solar radio astronomy and radio physics, D. Reidel Publ. Co., Dordrecht, Holland, 1979.
- Odstrcil, D., V. J. Pizzo, J. A. Linker, P. Riley, R. Lionello, Z. Mikic, and J. G. Luhmann, Initial coupling of coronal and heliospheric numerical magnetohydrodynamic codes, *J. Atmos. Sol. Terr. Phys.*, **66**, 1311–1326, Doi: [10.1016/j.jastp.2004.04.007](https://doi.org/10.1016/j.jastp.2004.04.007), 2004.
- Robbrecht, E., D. Berghmans, and R.A.M. Van der Linden, Automated LASCO CME Catalog for Solar Cycle 23: Are CMEs Scale Invariant? *ApJ*, **691**, 12–22, DOI: [10.1088/0004-637X/691/2/1222](https://doi.org/10.1088/0004-637X/691/2/1222), 2009.
- Tapping, K.F., The 10.7 cm solar radio flux (F10.7), *Space Weather*, **11**, 394–406, DOI: [10.1002/swc.20064](https://doi.org/10.1002/swc.20064), 2013.
- Thernisien, A.F.R., R.A. Howard, and A. Vourlidas, Modeling of flux rope coronal mass ejections, *ApJ*, **652** (1), 763–773, DOI: [10.1086/508254](https://doi.org/10.1086/508254), 2006.
- Vršnak, B., T. Žic, D. Vrbanec, M. Temmer, T. Rollett, C. Möstl, A. Veronig, J. Čalogović, M. Dumbović, S. Lulić, Y.-J. Moon, and A. Shanmugaraju, Propagation of Interplanetary Coronal Mass Ejections: The Drag-Based Model, *Sol. Phys.*, **285** (1–2), 295–315, DOI: [10.1007/s11207-012-0035-4](https://doi.org/10.1007/s11207-012-0035-4), 2013.
- Xue, X.H., C.B. Wang, and X.K. Dou, An ice-cream cone model for coronal mass ejections, *J. Geophys. Res. Space Phys.*, **110** (8), DOI: [10.1029/2004JA010698](https://doi.org/10.1029/2004JA010698), 2005.

Cite this article as: Devos A, Verbeeck C & Robbrecht E: Verification of space weather forecasting at the Regional Warning Center in Belgium. *J. Space Weather Space Clim.*, 2014, **4**, A29.