Solar Phys (2018) 293:94
https://doi.org/10.1007/s11207-018-1312-7

An Event-Based Verification Scheme for the Real-Time Flare Detection System at Kanzelhöhe Observatory

W. Pötzi[1] · **A.M. Veronig**[1,2] · **M. Temmer**[2]

Received: 16 March 2018 / Accepted: 30 May 2018 / Published online: 20 June 2018

Abstract In the framework of the Space Situational Awareness program of the European Space Agency (ESA/SSA), an automatic flare detection system was developed at Kanzelhöhe Observatory (KSO). The system has been in operation since mid-2013. The event detection algorithm was upgraded in September 2017. All data back to 2014 was reprocessed using the new algorithm. In order to evaluate both algorithms, we apply verification measures that are commonly used for forecast validation. In order to overcome the problem of rare events, which biases the verification measures, we introduce a new event-based method. We divide the timeline of the Hα observations into positive events (flaring period) and negative events (quiet period), independent of the length of each event. In total, 329 positive and negative events were detected between 2014 and 2016. The hit rate for the new algorithm reached 96% (just five events were missed) and a false-alarm ratio of 17%. This is a significant improvement of the algorithm, as the original system had a hit rate of 85% and a false-alarm ratio of 33%. The true skill score and the Heidke skill score both reach values of 0.8 for the new algorithm; originally, they were at 0.5. The mean flare positions are accurate within ± 1 heliographic degree for both algorithms, and the peak times improve from a mean difference of 1.7 ± 2.9 minutes to 1.3 ± 2.3 minutes. The flare start times that had been systematically late by about 3 minutes as determined by the original algorithm, now match the visual inspection within -0.47 ± 4.10 minutes.

Keywords Flares, forecasting · Instrumentation and data management

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11207-018-1312-7) contains supplementary material, which is available to authorized users.

✉ W. Pötzi
werner.poetzi@uni-graz.at

A.M. Veronig
astrid.veronig@uni-graz.at

M. Temmer
manuela.temmer@uni-graz.at

[1] Kanzelhöhe Observatory for Solar and Environmental Research, University of Graz, Graz, Austria

[2] Institute of Physics/IGAM, University of Graz, Graz, Austria

1. Introduction

Solar flares are sudden enhancements of radiation in a wide range of wavelengths within regions of strong magnetic fields on the Sun, the so-called active regions, which have a complex magnetic configuration (*e.g.* Sammis, Tang, and Zirin, 2000). The flare energy is converted into the acceleration of high-energy particles, mass motions, and heating of the solar plasma (*e.g.* reviews by Priest and Forbes, 2002; Benz, 2017). They are well observable optically from ground-based observatories (*e.g.* review by Veronig and Pötzi, 2016). In addition to regular visual detection, reporting, and classification of solar Hα flares by a network of observing stations distributed over the globe and archive at NOAA's[1] National Geophysical Data Center (NGDC), recent efforts have also been made to develop automatic flare detection routines. The detection methods range from comparatively simple image recognition methods based on intensity variations derived from running-difference images (Piazzesi *et al.*, 2012), region-growing and edge-based techniques (Veronig *et al.*, 2000), to more complex algorithms using machine learning (Fernandez Borda *et al.*, 2002; Ahmed *et al.*, 2013), or support vector machine classifiers (Qu *et al.*, 2003). These methods have been applied to space-borne image sequences in the extreme ultraviolet (EUV) and soft X-ray range (*e.g.* Qahwaji, Ahmed, and Colak, 2010; Bonte *et al.*, 2013), but also to ground-based Hα filtergrams (*e.g.* Veronig *et al.*, 2000; Kirk *et al.*, 2013; Pötzi *et al.*, 2015).

In order to test the quality of an automatic system, to improve the algorithms, and to compare different systems, a verification scheme has to be applied (*e.g.* Devos, Verbeeck, and Robbrecht, 2014). First verification analyses have been performed more than 100 years ago for weather forecasts (Finley, 1884). Almost all forecast systems, including space weather tools, today undergo regular verification (Balch, 2008; Kubo, Den, and Ishii, 2017), applying the same methods (Henley *et al.*, 2015). For this, the Space Weather Prediction Center (SWPC) of NOAA provides verification measures for most of their products back to 1986.[2] A comprehensive comparison of the performance of different flare forecasting tools was recently performed by Barnes *et al.* (2016) (see also the review by Green *et al.*, 2018). The National Institute of Information and Communications Technology (NICT) of Japan[3] provides an online tool to compare the flare and geomagnetic forecast performance of six different regional space weather warning centres that started in 1999.

In this article we present an enhanced version of the automatic flare detection system described in Pötzi *et al.* (2015), and we validate the detection algorithm, taking into account the fact that flares are rare events and that thus a normal verification scheme would be biased by this strong imbalance between positive and negative events. To overcome this problem, we introduce a method for handling the different lengths of flaring and non-flaring periods and retrieve a comparable number of events in the two classes.

2. Observations and Methods

2.1. Observations

Kanzelhöhe Observatory for Solar and Environmental Research (KSO) is operated throughout the year at a mountain ridge in southern Austria near Villach. Patrol full-disc observations of the Sun are possible for about 300 days a year, typically 1000 – 1200 hours of

[1] National Oceanic and Atmospheric Administration.

[2] NOAA, Space Weather Prediction Center – Forecast Verification, http://www.swpc.noaa.gov/content/forecast-verification.

[3] NICT, http://seg-web.nict.go.jp/cgi-bin/forecast/eng_forecast_score.cgi.

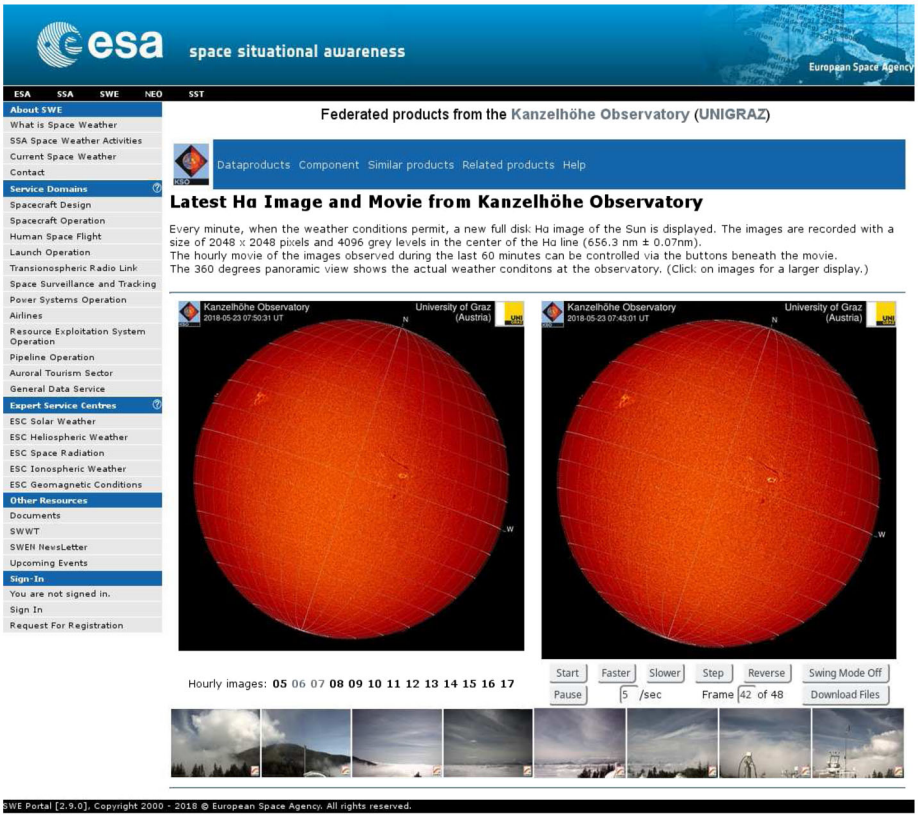

Figure 1 Kanzelhöhe Hα services on the ESA SSA space weather portal http://swe.ssa.esa.int/web/guest/kso-federated. The submenu contains the links to white-light solar images, flare detection data, and the flare alerts.

observations with high quality. The Sun is observed in the Hα spectral line (Otruba and Pötzi, 2003; Pötzi *et al.*, 2015), the Ca II K spectral line (Hirtenfellner-Polanec *et al.*, 2011), and in white light (Otruba, Freislich, and Hanslmeier, 2008).

The KSO Hα telescope is a refractor with an aperture ratio number of $d/f = 100/2000$ and a Lyot band-pass filter centred at the Hα spectral line ($\lambda = 656.3$ nm) with a full-width at half-maximum (FWHM) of 0.07 nm. The CCD camera has 2048 × 2048 pixels, a dynamic range of 12 bit, and a gigabit ethernet interface. A frame rate of nearly seven images *per* second permits the application of frame selection to benefit from moments of good seeing. Every six seconds, the best of the last ten images is recorded. Automatic exposure-time calculation is in place to avoid overexposure and saturation during flares (typically in the range from 1.5 to 3.0 ms). The spatial sampling of the full-disc observations is ~ 1 arcsec, corresponding to about 725 km on the Sun.

2.2. Flare Detection Algorithm

Each Hα image is immediately checked for quality after acquisition and classified into three categories: *bad* – images are sorted out and moved to a temporary archive, *fair* – images are stored in the archive and used for visual inspections, and *good* – these images are suitable for further automatic processing (Veronig and Pötzi, 2016). The image processing consists

```
20150409 102226 1024.00 1024.00  943.03      49 1136.19  850.33    144  2.0861  0.1493  1.7991  2.4469 1155.00  854.00
20150409 102226 1024.00 1024.00  943.03      58  872.52 1242.10     31  1.9942  0.0806  1.7794  2.1072  871.00 1243.00
20150409 102238
20150409 102238 1024.00 1024.00  943.38      49 1136.22  851.78    144  2.0959  0.1405  1.7875  2.4159 1119.00  846.00
20150409 102244
20150409 102244 1024.00 1024.00  942.86      49 1134.48  850.23    133  2.1177  0.1571  1.8521  2.5157 1118.00  847.00
20150409 102251
20150409 102251 1024.00 1024.00  943.10      49 1135.70  851.24    122  2.0863  0.1201  1.8549  2.3966 1156.00  856.00
20150409 102302
20150409 102302 1024.00 1024.00  943.30      49 1135.36  852.11    134  2.0886  0.1479  1.8036  2.4755 1119.00  851.00
20150409 102308
20150409 102308 1024.00 1024.00  942.88      49 1135.56  851.14    132  2.0478  0.1275  1.7177  2.3384 1154.00  855.00
20150409 102333
20150409 102333 1024.00 1024.00  943.27      49 1136.72  851.07    124  2.1198  0.1574  1.7715  2.4749 1120.00  850.00
20150409 102333 1024.00 1024.00  943.27      62 1145.11 1251.52     27  2.1500  0.1511  1.9216  2.4709 1145.00 1249.00
20150409 102346
20150409 102346 1024.00 1024.00  943.26      49 1137.80  851.25    151  2.0794  0.1447  1.7688  2.4805 1156.00  855.00
20150409 102346 1024.00 1024.00  943.26      62 1144.32 1250.83     47  2.1165  0.1217  1.9072  2.4014 1145.00 1252.00
20150409 102416
20150409 102416 1024.00 1024.00  942.58      49 1137.00  851.50    116  2.0593  0.1191  1.7940  2.3783 1121.00  850.00
20150409 102416 1024.00 1024.00  942.58      62 1145.68 1250.53     79  2.3206  0.3322  1.9046  3.0337 1147.00 1250.00
20150409 102429
20150409 102429 1024.00 1024.00  943.73      49 1134.82  851.03    121  2.0466  0.1167  1.8159  2.3061 1157.00  853.00
20150409 102429 1024.00 1024.00  943.73      62 1145.59 1251.88     66  2.3130  0.2834  1.7922  2.8950 1147.00 1254.00
20150409 102435
20150409 102435 1024.00 1024.00  943.36      49 1137.20  850.48    135  2.0795  0.1387  1.7636  2.4279 1121.00  851.00
20150409 102435 1024.00 1024.00  943.36      62 1145.80 1252.05     64  2.3041  0.2656  1.8546  2.9380 1147.00 1253.00
20150409 102518
```

Figure 2 Part of a log file produced by the flare recognition algorithm. The file includes (from *left* to *right*) date, time, disc centre coordinates, solar radius, ID of the region, position of the region (centre of gravity), size of the region, four brightness values (mean, root mean square, minimum, and maximum), and position of the brightest pixel. If no bright region (*i.e.* flare candidate) is detected, only a time stamp is set.

of two main steps. In the first step, the images are preprocessed and prepared for the near real-time provision on the ESA/SSA space weather portal,[4] where they are available as jpeg and fits files within seconds after recording (see Figure 1). In the second step, an image recognition algorithm segments the solar disc into bright regions (*i.e.* potential flare regions), background, filaments, and sunspots. This segmentation is done in four steps:

i) Preprocessing: The image is normalised and features are enhanced by application of bandpass filters, which remove large-scale inhomogeneities and noise.
ii) Feature extraction: Each pixel is assigned a class probability for the classes sunspot, filament, background, and flare.
iii) Multilabel segmentation: The noisy distribution of pixel classes is regularised.
iv) Postprocessing: Each flare and filament is identified and tracked. Characteristics such as area, brightness, and position are written into a log file (*e.g.* Figure 2). This file is updated with each new image that enters the processing pipeline.

A detailed description of the algorithm, the implementation in the KSO observing pipeline, and its performance in real time can be found in Riegler *et al.* (2013), Riegler (2013), and Pötzi *et al.* (2015).

The data of the log files (see Figure 2) are used to extract flare events and their characteristics, such as heliographic position, flare classification, and flare start, peak, and end times. As each flaring region is assigned a unique ID that is propagated from image to image, its evolution can be tracked. The flare area defines the importance class, flares with areas smaller than 100 microhemispheres (µhem) are called subflares, importance 1 class flares extend to 250 µhem, importance 2 class to 600 µhem, importance 3 class flares to 1200 µhem, and larger flares are of importance class 4. The flare brightness is originally defined by the brightness enhancement in the Hα line core (Svestka, 1976), but this enhancement depends strongly on the FWHM and the characteristics of the filter. In our case, it is defined by the maximum brightness within the flaring region compared to the background intensity: the faint (F) level is defined between three and six times the normalised background intensity, then up to nine times the normal (N) level, and everything higher than

[4] http://swe.ssa.esa.int/web/guest/kso-federated.

Table 1 Differences between the original algorithm that was in use at KSO before September 2017 (Pötzi *et al.*, 2015) and the new improved flare detection algorithm (this study).

	Original algorithm	New algorithm
Preflare brightness	Three times higher than the faint level	Reset counter, if brightness is lower than the faint level
Brightness threshold	(Maximum − mean) brightness	Maximum brightness
Centre-to-limb variation correction	$clv = 1$	$clv = 0.55$
Bandpass correction	Stepwise	Continuous
Foreshortening and 3D	$\frac{1}{(\cos\rho)^{0.6}}$	$\frac{1}{\sqrt[4]{1-\mu^2}} \leq 2$
Lower area threshold	50 µhem	25 (10) µhem (brightness N or B)
Flare position	Brightest pixel	Centre of gravity
Long-lasting bright flares	Keep flare level	End flare if brightness < half of maximum
Data gaps	Flare ends if data gap > 5 min	Flare ends if brightness falls below threshold or data gap > 20 min

this is a bright (B) flare. Small flares of type S and 1 are mostly of type F, importance 2 flares typically reach N, and the larger flares are very often of B type, but in some cases, subflares can also reach brightness B. In the automatic flare detection system, we focus on flares of Hα importance classes ≥1, *i.e.* we ignore subflares.

Based on flare observations in 2012, one year before the original algorithm was introduced, the thresholds for the flare brightness classes were obtained from the brightness values in the log files. This was mostly performed with flares of importance classes 1 and higher.

The main changes from the original to the new flare detection algorithm are the following (see also Table 1). The brightness handling is corrected in a threefold manner, *i.e.* in order to determine the start of a flare, a threshold level had to be reached at three time steps in the original algorithm. Very often, the brightness jitters near the threshold level, however, it can exceed or fall below the threshold sometimes for a few minutes. In the new algorithm, the counter is therefore reset if the brightness repeatedly falls below the threshold. The brightness threshold is now determined via the maximum flare brightness, which is more stable than the original brightness difference method.

The centre-to-limb correction is based on the formula from Scheffler and Elsasser (1990), $\frac{1+clv}{1+clv\sqrt{1-\mu^2}}$, where clv is the centre-to-limb variation factor that covers the decreasing limb brightness and the distance to the centre of the solar disc, μ, as a fraction of the radius. This clv factor is derived experimentally and depends on the wavelength and also on the stray light in the atmosphere and telescope. The area correction for the bandpass filtering, which tends to increase the detected area especially for thin elongated structures (see Figure 3), is continuous for the new algorithm. In addition, the correction for foreshortening of flare structures towards the limb has been changed. The foreshortening should theoretically behave like $\frac{1}{\cos\rho}$, where ρ is the angular distance to the disc centre. This is not the case for bright features (Godoli and Monsignori Fossi, 1967), however, and 3D effects also play a role near the limb. As many regions seemed to increase their flare sizes near the limb, the factor for the foreshortening is reduced. The calculation basis is changed from ρ to μ ($\cos\rho = (1 - \mu^2)$). The foreshortening correction now has an upper limit of 2 in order to prevent overcorrecting flare areas near the limb.

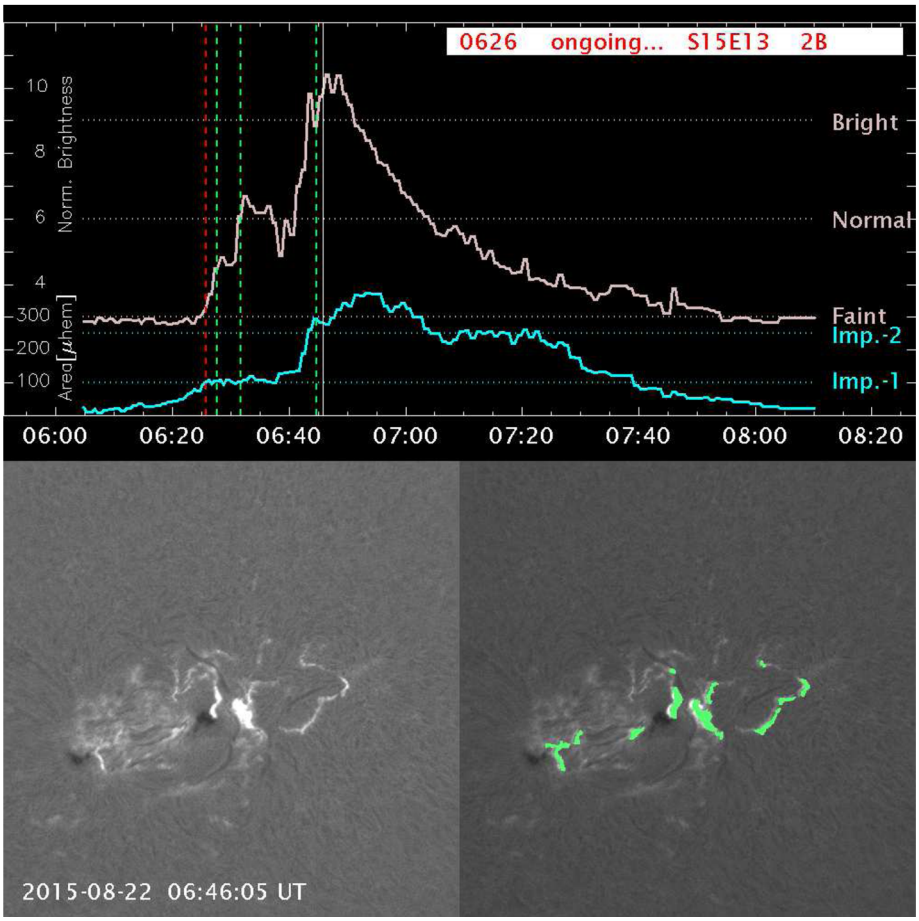

Figure 3 Schematic presentation of the flare detection algorithm for a 2B flare on 22 August 2015. The *upper panel* shows the evolution of the brightness (*grey*) and area (*blue*) extracted from the log file (*e.g.* Figure 2) together with the brightness and area threshold levels. The *vertical dashed red line* indicates the flare start time and the *vertical dashed green lines* show the alert times for flare start and when it reaches importance class 1 and 2. The *vertical solid line* is the time stamp for the lower panel in which the original image of the flare (*left*) and an overlay of the detected flare region (*right*) is shown. A movie of the whole event is available in the online version of this article (20150822_2B.avi).

The new algorithm also includes the detection of small subflares when their brightness reaches the N level (defined as twice the faint level). Very long-lasting (over several hours) bright flares are now split up when the brightness falls below half of the peak brightness and rises again afterwards, *i.e.* every new significant brightness peak is defined as a new individual flare. As we process ground-based data, data gaps due to clouds or varying seeing conditions are common. A flare does not stop immediately with each data gap; the brightness level after the data gap defines the flare status. A data gap of more than 20 min defines the end of a flare.

Figure 3 shows an example of the flare extraction based on a log file such as is shown in Figure 2. The flare start is defined via an area threshold level (≥ 50 μhem) and a brightness threshold ($\geq$ faint). Both have to be reached in order to define the detection and start time

of a flare. Therefore the flare did not start before 6:26 UT because the area was below the defined threshold and the brightness threshold level of faint was not reached. The first alert was issued 2 min after the detected flare start time (first green dashed vertical line), and the following alerts were also issued within 1 or 2 min; these are updates to the initial alert, should the flare reach the next importance class. The red text in the white bar in the top right corner shows the message that is given on the space weather portal. It corresponds to the time stamp indicated as a vertical solid line at 6:46 UT. The green areas in the lower panel show the detected flare region. These areas tend to be somewhat larger than the real flare ribbons, which is due to the bandpass filtering that is performed in the first step of the image recognition process. A movie of the whole event is available in the online version of this article (20150822_2B.avi).

An evaluation of the results of the original flare detection algorithm for the period July 2013 to December 2015 (Pötzi *et al.*, 2015; Veronig and Pötzi, 2016) gave the following results for events that were of importance class 1 and higher: for 95% of the 174 events taken in consideration, the peak times were within ± 5 min of the peak times listed in the visual flare reports of KSO. The determined heliographic positions were mostly better than $\pm 1°$, and 95% of all flares were detected. Based on the observations and results from these early years, the algorithm has been improved. Table 1 summarises the differences between the original algorithm and the new algorithm that is in use since September 2017 and that is used in this study.

2.3. Verification Scores for Classification Accuracy

A very detailed description of the forecast verification can be found in Jolliffe and Stephenson (2012). In this study we only concentrate on a few terms and the simple form of events, the dichotomous events. Dichotomous events are of the category "yes – there is an event" or "no – there is no event". For such events a very simple contingency table or confusion matrix (Table 2) can be established. The table contains the following entries:

TP *true positive* = *hit* – an event was predicted or detected and it occurred.
FP *false positive* = *false alarm* – an event was predicted or detected, but none occurred; this is also known as a false-positive event.
FN *false negative* = *miss* – no event was predicted or detected, but one occurred, *i.e.* a false-negative event.
TN *true negative* – no event was predicted or detected and none occurred, *i.e.* a correct negative event.

From this contingency table, various numbers of verification measures can be obtained: The *accuracy* gives the fraction of the correct detections:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}. \tag{1}$$

The values range from 0 to 1, where 1 means a perfect detection (FP = 0 and FN = 0).

The *bias score* measures the ratio of predicted events to the observed events:

$$\mathrm{BS} = \frac{\mathrm{TP} + \mathrm{FP}}{\mathrm{TP} + \mathrm{FN}}. \tag{2}$$

It does not measure how well the detection works, but it indicates whether a system underdetects (< 1) or overdetects (> 1) events. The values range from 0 to ∞, with a perfect score of 1 (same number of FP and FN).

Table 2 Contingency table for dichotomous events (also called confusion matrix) for the classes hit (TP = true positive), false alarm (FP = false positive), miss (FN = false negative), or true negative (TN), depending on the forecast and observation of the event. The rows give the sum of positive or negative forecasts, the columns give the sums for positive or negative observations.

		Observation		
		yes	no	
Detection	yes	TP	FP	TP + FP = $\sum$ (detected)
	no	FN	TN	FN + TN = $\sum$ (not detected)
		P = TP + FN = $\sum$ (observed)	N = FP + TN = $\sum$ (not observed)	

The *hit rate* (TPR, *true positive rate*) gives the fraction of the observed events that were detected:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{3}$$

The values range from 0 to 1 with a perfect score of 1 (FN = 0). It ignores all negative events (FP and TN) and is therefore very sensitive to hits.

The opposite of the hit rate is the *false-positive rate* (FPR):

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}}. \tag{4}$$

It ranges from 0 to 1 with a perfect score of 0 (FP = 0), but here all positive events are ignored.

The hit rate should always be checked against the *false-discovery rate* (FDR), which gives the fraction of detected events that where not observed:

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}}. \tag{5}$$

The values range from 0 to 1 with a perfect score of 0. The FDR is sensitive to false alarms and ignores misses. The FPR and the FDR should always be presented in a good validation scheme, as the first includes the TN values and the second the TP values, so that it depends strongly on the ratio of these values.

The *threat score* (*TS*) or *critical success index* (CSI) indicates how well the positive detections correspond to the observed events:

$$\mathrm{TS} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{6}$$

The values range from 0 to 1 with a perfect score of 1 (FP + FN = 0). It can be interpreted as the accuracy score of the system neglecting negative events.

In order to quantify how well positive events are separated from negative events, the *Hanssen and Kuipers discriminant* or *true skill statistic* (TSS) is used:

$$\mathrm{TSS} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} - \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}} = \mathrm{FDR} - \mathrm{FPR}. \tag{7}$$

It ranges from -1 to 1 with a perfect score of 1. It can be interpreted as the difference between the hit rate and the false-positive rate. A large number of TN increases this measure. As a more generalised skill score that measures the fraction of correct detections after eliminating random detections, the *Heidke skill score* (HSS) is used:

$$\mathrm{HSS} = \frac{2 \cdot (\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN})}{(\mathrm{TP} + \mathrm{FN}) \cdot (\mathrm{TN} + \mathrm{FN}) + (\mathrm{TP} + \mathrm{FP}) \cdot (\mathrm{TN} + \mathrm{FP})}. \tag{8}$$

It ranges from -1 to 1 with a perfect score of 1 ($\mathrm{FN} = 0$ and $\mathrm{FP} = 0$). HSS becomes 0 in the case $\mathrm{TP} = \mathrm{FN} = \mathrm{FP} = \mathrm{TN}$.

If the number of hits and true negatives is not balanced, some of the above measures can give misleading numbers, as they do not include all of the values of the contingency table. A famous example for such an imbalanced system was the tornado forecast in 1884 of Sergeant John P. Finley (1884). As there were many regions with low tornado probabilities, the accuracy of his forecast was very high, although he had a low hit rate. This became well known as the Finley Affair (Murphy, 1996). This behaviour of a system is known as the rare-events problem (Murphy, 1991).

3. Results

In order to evaluate the new automatic flare detection system at KSO and to compare the outcomes of the new and the original algorithms (*i.e.* the test group), independent measurements have to be used. The verification group is made up of NOAA and KSOv (visual KSO flare reports) datasets. The first is based on data from the NOAA Space Weather Prediction Center (SWPC), which provides official lists of solar events available online at https://www.swpc.noaa.gov/products/solar-and-geophysical-event-reports. The information on the flare events is collected from different observing stations from all over the world. The KSOv dataset comes with the visual KSO flare reports (KSOv) that are available online at http://cesar.kso.ac.at/flare_data/kh_flares_query.php. These flare reports are also sent to NOAA on a monthly basis. They are compiled every few days from the visual inspection of data that also includes images of fair quality. We expect that the results of the automatic detections are on average closer to the visual KSO flare reports than the NOAA reports, as they are based on the data from the same observatory. Furthermore, the NOAA event reports are not always complete; sometimes, even larger events are missing (see *e.g.* Figure 5). However, it is also important to compare the outcome against the NOAA reports, as they provide an independent set of visual flare reports.

In Section 3.1 we present the results obtained for the verification of the improved automatic flare detection system. In Section 3.2 we evaluate the accuracy of the automatically determined flare parameters (importance class, heliographic position, and start and peak times).

3.1. Algorithm Performance in Flare Recognition

For the analysis of the flare recognition, all flares between January 2014 and December 2016 within 60° central meridian distance (CMD) that issued an alert via the ESA space weather portal were taken into account. The criteria for such flares are at least importance class 1 and a position within 60° CMD. During the evaluation period, we observed 142 flare events fulfilling these criteria. Smaller flares and flares outside 60° CMD are not considered, as they are most probably not relevant for space weather.

Table 3 Summary of yearly sunshine duration, yearly observation times, total flare times, and the international sunspot number (ISN, Clette and Lefèvre, 2016) at the Kanzelhöhe Observatory from 2014 to 2016.

	2014	2015	2016
Sunshine hours (yearly)	1984.0 hrs	2441.3 hrs	2240.9 hrs
Observation times (good + fair quality)	952.7 hrs	1266.1 hrs	1246.2 hrs
Flare times: KSO visual	7.6%	3.9%	0.32%
Flare times: original algorithm	3.9%	1.8%	0.07%
Flare times: new algorithm	6.1%	3.5%	0.11%
Yearly averaged ISN	113.3	69.8	39.8

To account for area uncertainties of the simple threshold method used in the visual inspection, we include all flares of > 80 µhem in order not to miss any flares of importance class ≥ 1. A database query resulted in 227 flares obtained by visual inspections in this period. Of these, 121 flares had importance class 1, 12 had importance class 2, and only 2 had class 3. These numbers also cover flares outside 60° CMD, which are excluded in our investigations. This results in a total number of 142 flares for the evaluation.

For solar flares, the situation is equivalent to the tornado forecast system, especially outside the maximum of a solar cycle, as from 2014 to 2016. During these years, a very low percentage of the total observing time is covered by flare events (see Table 3). The flare times decrease from a few percent in 2014 to tenths of percent in 2016, *i.e.* in 2016, only for 239 of 74 772 observing minutes a flare was detected. We note the differences in the time coverage between the three datasets in Table 3. For visual detections, images of fair quality are also taken into account, whereas the automatic system is fed only with images of good quality. The original system only detects flares exceeding areas of at least 50 µhem, the new system and the visual detection do not have this limitation. In this table all flares that have been detected, *i.e.* also flares that were too small for issuing an alert, are taken into account.

In order to obtain reliable and meaningful verification measures, the value of TP + FN + FP should be comparable to TN. In this case, verification measures that include TN, like the accuracy, are not biased. To achieve this, the evaluation method is made independent of time, *i.e.* the length of flaring periods or the length of quiet times should not count. The solution is to divide the time line into positive and negative events, if there is, *e.g.* one flare during one week, we only count three events: a negative event, followed by a positive event, and at last a negative event. By this method the number of positive events and negative events is in the same range, unless there is a longer time span when one flare directly follows the next, which can occur if there is more than one flaring region at the same time.

Figure 4 shows with an example how the event-based evaluation is performed. The background of the plot shows whether there were no observations (grey, *e.g.* night, clouds or very poor seeing), or the camera was switched on (white = good quality, green = fair quality, and orange = bad quality). The test group is marked in red (original algorithm) and blue (new algorithm) and the verification group in magenta (KSO visual reports, KSOv) and green (NOAA visual reports). The times at which a flare was detected or observed are marked with a coloured bar. The vertical line above each bar marks the peak time of the flare, and it is annotated with the associated flare type. The following rules have been used to define events:

- At least two sources must have observed or detected a flare, at least one of them must be from the verification group (KSOv or NOAA).

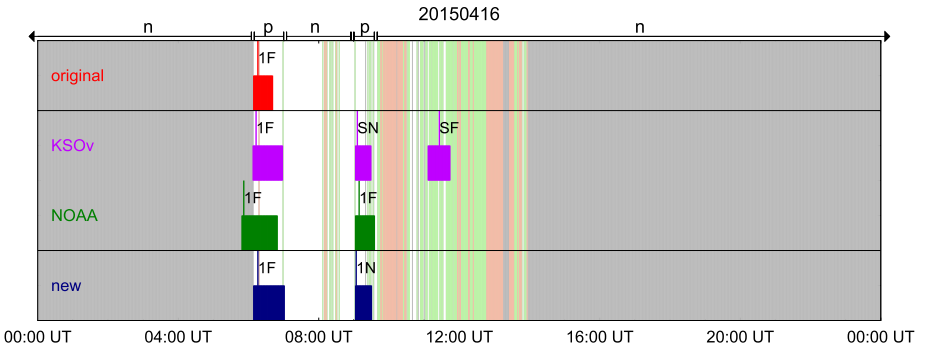

Figure 4 16 April 2015 divided into positive events (marked p above the plot) and negative events (n). Four sources are compared: the original flare detection algorithm (*red*), KSO visual (KSOv) (*magenta*), NOAA (*green*), and the new algorithm (*blue*). The time at which the flare is active is marked with a *coloured bar*. The *vertical line* above the *bar* denotes the peak time of the flare, and it is annotated with the flare type.

- Good image quality during flare peak time (white background).
- Flare peak times of the test group and the verification group lie within 5 min.
- If the observation day ends with an event and the next day begins with an event, the time in between will not be counted as a negative event.
- The time between two positive events counts as one negative event even though the time spans several days.
- If a flare peak time differs by more than 5 min from the verification group, it is defined as a false-positive event.

To give an example about how this scheme works, the observation day of 16 April 2015 is shown in Figure 4. According to the rules above, we obtain five elements:

Negative event: From midnight until 6:30 UT, this negative event reaches back until the last positive event because there are no observations and no event was detected.

Positive event: From 6:30 until 6:50 UT, an event was detected by all observation sources. NOAA was earlier in detecting the flare start due to observations when the KSO dome was still closed.

Negative event: Between 6:50 and 8:50 UT, no flare was detected.

Positive event: Between 8:50 and 9:30 UT, a flare was reported by both verification sources (NOAA and KSOv) and was detected by the new detection algorithm, thus it counts as a hit (TP). Instead, this event was not detected by the original algorithm, therefore it counts as a missed event (FN).

Negative event: The remaining day until the next positive event will count as a negative event. The event detected visually by the KSO is not considered, as the observation conditions were not good (green background), and this event was not detected by any other source.

For the day shown in Figure 4, the new algorithm would give TP = 2, FN = 0, FP = 0, and TN = 3, and the original algorithm would give TP = 1, FN = 1, FP = 0, and TN = 3.

An example of how false alerts or multiple alerts are produced is shown in Figure 5, on a day with some small subflares and one large 2B flare. The curves in the lower panel show the intensities for the original algorithm (red) and the new algorithm (blue). The 2B flare that peaks at 13:55 UT is interrupted by data gaps (clouds), the original algorithm splits this flare at 14:07 UT and produces a false alarm. The intensities for the new algorithm

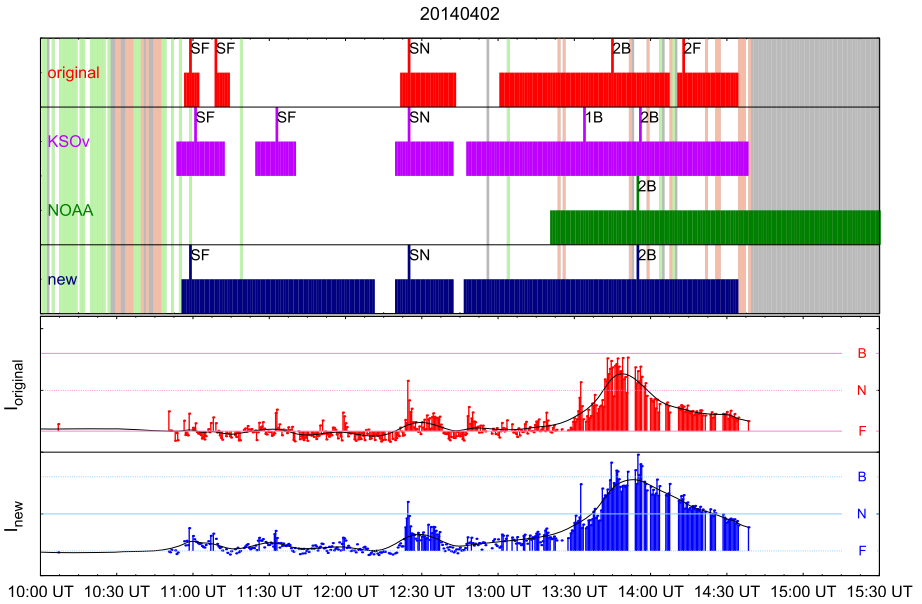

Figure 5 Observations from 2 April 2014 as an example for when the original algorithm splits great flares as a result of data gaps. The *lower plot* shows the intensity levels for both algorithms, the smoothed curve is overlaid in *black*. Note the missing events in the NOAA data.

are slightly higher due to the change in the intensity calculation. The changing observation conditions smooth the intensity maximum in unsharp images. Thus the time of the flare peak also differs between the original algorithm and the other datasets. This splitting of bright large flares was a problem of the original algorithm, as these flares often last for hours and therefore the probability of data gaps increases. This example also shows the problem of missing events using NOAA data we describe above: the subflares before 13 UT are not listed, although one of them was of brightness N.

The complete evaluation for 2014, 2015, and 2016 gives a total of 329 positive events and negative events, with 142 flares of importance class $\geq$1 within 60° CMD that are detected by at least one automatic routine and observed by at least one visual method. Table 4 shows the results of the comparison between the original and the new algorithm with the KSOv data. The sum of TP + FN + FP + TN = 329 gives the total number of positive events and negative events. The number of positive events is the sum of TP + FN = 142, and the number of negative events is the sum of FP + TN = 187. The latter number is larger due to periods without flares that are interrupted by false detections. If there is a false detection during a negative event period, this produces an additional positive event. The sum of the false events FN + FP clearly shows that the new algorithm has a much better flare detection performance than the original one, the value decreased from 82 (24.9%) to 34 (10.3%). Nonetheless, there are 29 false alarms (FP) in the new algorithm, but this number is more than twice as high (61) in the original algorithm.

The accuracy increased from 75% to nearly 90%, or in other words, the original algorithm issued one false alarm out of four, whereas in the new algorithm it is one out of 10. The hit rate, which ignores false-positive alarms, increased to 96.5%, as only five flares are missed. The false alarms are reflected in the bias score, which improved from 1.28 to 1.16, *i.e.* the "overdetection" is reduced. The false-alarm ratio is reduced from 33.5% to 17.5%, and

Table 4 Comparison of the original and the new flare detection algorithm with the visual KSO detections (KSOv) using the event-based scheme for flares of importance class 1 and higher between 2014 and 2016.

Rates/scores	Original algorithm	New algorithm
TP	121	137
FN	21	5
FP	61	29
TN	126	158
Accuracy	75.1%	89.67%
Bias score	1.3	1.2
Hit rate (TPR)	85.2%	96.5%
False-alarm ratio (FPR)	33.5%	17.5%
False-discovery rate	32.6%	15.5%
Threat score	0.60	0.80
TSS	0.53	0.81
HSS	0.51	0.79

the false-discovery rate from 32.6% to 15.5%; both show nearly the same reduction as the number of TP and TN are similar.

The threat score emphasising the hits rises from 0.60 for the old algorithm to 0.88 for the new algorithm, *i.e.* from six correct hits out of ten detected (plus missed) to eight out of ten. The TSS and the HSS both improved from about 0.5 to 0.8, emphasising the very good performance of the overall system for the new flare detection algorithm.

3.2. Algorithm Performance on Flare Location and Timing

The dataset for the flare location and timing analysis covers the same period as the dataset for the above analysis. In order to improve the statistics, we also analyse subflares. For the original algorithm we included subflares with areas ≥ 50 µhem and for the new algorithm, we also include smaller subflares, ≥ 25 µhem and ≥ 10 µhem, if the brightness reaches level N. This led to 413 events for the original flare detection algorithm and 451 for the new algorithm.

Figure 6 shows the distribution of the differences of the flare peak times and flare start times of the original and new detection algorithm with respect to KSOv. Most of the flare peak times, defined as the time of the highest flare brightness, are within 5 min, only 8.2% deviate by more than 5 min for the original algorithm and 4.8% for the new algorithm. The mean of the absolute differences is 1.72 ± 2.95 min for the old algorithm and 1.29 ± 2.30 min, both numbers including the outliers with more than 5 min difference. The flare start times improves from a difference of 2.57 ± 3.90 min for the original detection algorithm to -0.47 ± 4.10 min for the new algorithm. This improvement mostly arises because the area limitation of 50 µhem does not apply to the new algorithm.

Figure 7 shows the difference between the heliographic latitude and longitude of the KSO visual data and the two test group datasets for 2014 to 2016 for all flares exceeding 50 µhem. In this case, we recall that the flare position in the original data and the KSOv data is defined via the brightest pixel and in the new algorithm via the centre of gravity. Nonetheless, the agreement between the detections and the inspections only changed from $0.28 \pm 0.68°$ in latitude to $0.66 \pm 1.02°$ and from $0.03 \pm 0.17°$ to $0.10 \pm 0.30°$ for the longitude values.

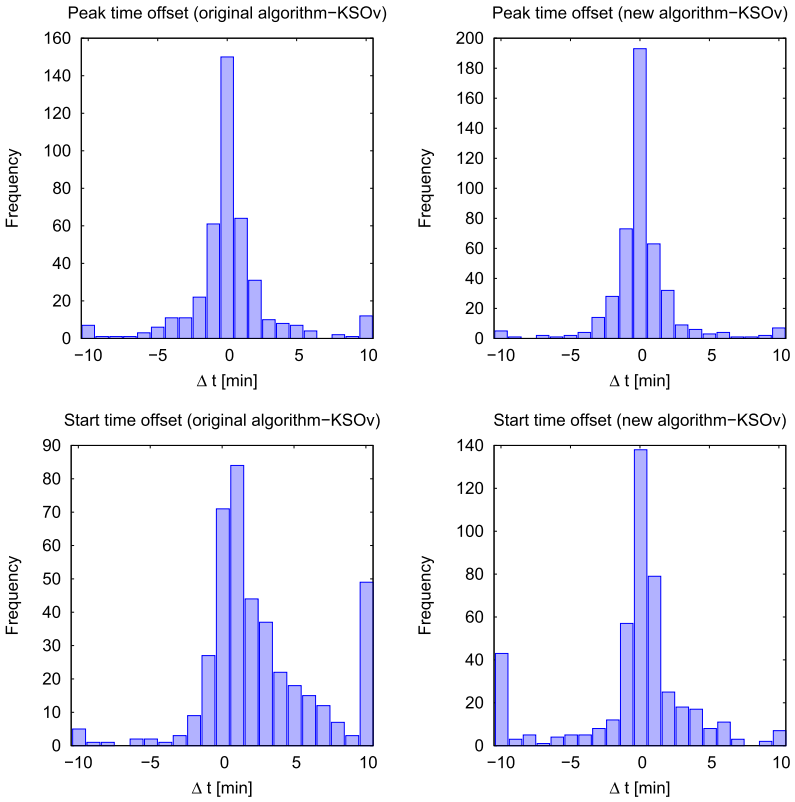

Figure 6 Distribution of the absolute differences of the flare peak times (*upper row*) and flare start times (*lower row*) between KSOv and the original algorithm (*left*) and KSOv and the new algorithm (*right*) for all flares exceeding 50 µhem from 2014 until 2016, *i.e.* also the flares below the alert threshold of importance 1 class.

4. Discussion

In order to overcome the problem of rare events, which could result in misleading verification scores, we introduce an event-based approach. If the verification scores are calculated with a time-based method (*i.e.* setting fixed time steps in which the status of the system is checked), the number of TN would be considerably higher than the other numbers (see Table 3). In such a case, the accuracy and the false-alarm ratio, which are strongly biased by the high TN values, would give rather good results, but they would not reveal the state of the system. With the event-based approach, however, the number of "events" and "no events" is well balanced, and the verification scores are mostly affected by false alarms and missed events. These TN and FN are the numbers of interest when a system has to be validated.

The new automatic flare detection algorithm that was implemented at the KSO Hα observing system in September 2017 shows a clear improvement over the original algorithm. The most relevant advantage of the new algorithm is the smaller number of false detections and the dramatically reduced number of missed events. The original algorithm issued 182 alerts of which only 121 were correct, one-third were false alarms, or as Figure 5 shows, a double or triple alarm for the same event. The new algorithm would have issued 156 alerts

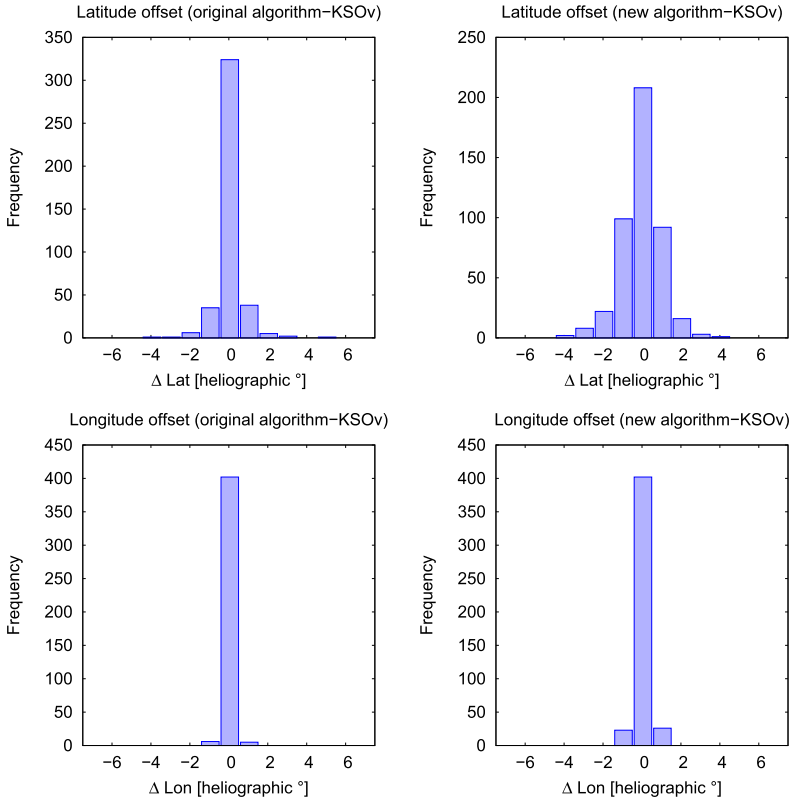

Figure 7 Distribution of the absolute differences of the flare heliographic latitude and longitude between KSOv and the original algorithm (*left*) and KSOv and the new algorithm (*right*) for all flares exceeding 50 µhem from 2014 until 2016, *i.e.* also the flares below the alert threshold of importance 1 class.

of which only 29 would have been incorrect, *i.e.* only one-fifth. Again, most of them are multiple alerts for one event. The number of missed events is mainly reduced because of the brightness correction and the changed threshold detection. The new method gives peak times and start times that coincide substantially better with the visual observations (Figure 6).

There is space for further improvements for the treatment of data gaps (due to seeing or actual missing data) to account for the cases when a single flare is misclassified as several separate flares. However, this modification would at present need human intervention and it is not within the scope of the actual approach. This scope is a uniform and unbiased flare detection algorithm that can be automatised and run with no subjective criteria.

5. Conclusion

With the event-based approach for the evaluation of a system that produces rare events, results cannot be improved by selecting verification scores that overvalue true negatives, *e.g.* high number of true negatives can result in high accuracy values independent of the system performance. Additionally, the scores can be grouped together and it is not so important which of the scores is selected:

- The success of the system can be either displayed via hit rate or accuracy.
- The failures can be seen in the false-alarm ratio (FPR), the false-discovery rate (FDR), or the bias. Only a very large number of false alarms would produce different results here.
- The overall quality is represented by the threat score, the true skill statistics, or the Heidke skill score.

A forecast system is based on a fixed time interval, whereas an observing system just waits for a detection of events, therefore this event approach is a robust method for testing and comparing systems that detect events and do not forecast them. This is especially the case for data modulated by the solar activity cycle, which shows extended periods in the rare events regime.

The evaluation in 2014 – 2016, covering 142 flares with importance higher than class 1 within 60° CMD, shows a significant increase in almost all measurements for the new developed algorithm. The flare location performances of the old and new algorithm are consistent, considering the different definition of flare centre between the KSOv verification data and the old algorithm, with that of the new one. The flare peak times are now closer to the visual observations: they improved from 1.7 ± 3.0 min to 1.3 ± 2.3 min. The flare start times improved significantly from a systematic delay of 2.6 ± 4.0 min, and are now accurate to within -0.5 ± 4.1 min. The hit rate increases from 85% to 96% for the new algorithm, and the false-alarm ratio decreases from 33% to 17%. Only 4% of the events are missed in the new algorithm compared to 15% for the original method. The main reason for the large number of false alerts in the original algorithm was the splitting of flares into multiple events. This splitting is reduced in the new algorithm by adapting the brightness handling. The TSS and HSS skill scores rise from 0.5 to 0.8.

Acknowledgements Open access funding provided by University of Graz. The automatic flare detection system was developed within the framework of the Space Weather Segment of the ESA Space Situational Awareness Programme.

Disclosure of Potential Conflicts of Interests The authors declare that they have no conflicts of interest.

References

Ahmed, O.W., Qahwaji, R., Colak, T., Higgins, P.A., Gallagher, P.T., Bloomfield, D.S.: 2013, Solar flare prediction using advanced feature extraction, machine learning, and feature selection. *Solar Phys.* **283**, 157. DOI. ADS.

Balch, C.C.: 2008, Updated verification of the Space Weather Prediction Center's solar energetic particle prediction model. *Space Weather* **6**, S01001. DOI. ADS.

Barnes, G., Leka, K.D., Schrijver, C.J., Colak, T., Qahwaji, R., Ashamari, O.W., Yuan, Y., Zhang, J., McAteer, R.T.J., Bloomfield, D.S., Higgins, P.A., Gallagher, P.T., Falconer, D.A., Georgoulis, M.K., Wheatland, M.S., Balch, C., Dunn, T., Wagner, E.L.: 2016, A comparison of flare forecasting methods. I. Results from the "All-Clear" Workshop. *Astrophys. J.* **829**, 89. DOI. ADS.

Benz, A.O.: 2017, Flare observations. *Living Rev. Solar Phys.* **14**, 2. DOI. ADS.

Bonte, K., Berghmans, D., De Groof, A., Steed, K., Poedts, S.: 2013, SoFAST: automated flare detection with the PROBA2/SWAP EUV Imager. *Solar Phys.* **286**, 185. DOI. ADS.

Clette, F., Lefèvre, L.: 2016, The new sunspot number: assembling all corrections. *Solar Phys.* **291**, 2629. DOI. ADS.

Devos, A., Verbeeck, C., Robbrecht, E.: 2014, Verification of space weather forecasting at the Regional Warning Center in Belgium. *J. Space Weather Space Clim.* **4**(27), A29. DOI. ADS.

Fernandez Borda, R.A., Mininni, P.D., Mandrini, C.H., Gómez, D.O., Bauer, O.H., Rovira, M.G.: 2002, Automatic solar flare detection using neural network techniques. *Solar Phys.* **206**, 347. DOI. ADS.

Finley, J.P.: 1884, Tornado predictions. *Am. Meteorol. J.* **1**, 85.

Godoli, G., Monsignori Fossi, B.C.: 1967, On the correction for foreshortening for Ca plages. *Solar Phys.* **1**, 148. DOI. ADS.

Green, L.M., Török, T., Vršnak, B., Manchester, W., Veronig, A.: 2018, The origin, early evolution and predictability of solar eruptions. *Space Sci. Rev.* **214**, 46. DOI. ADS.

Henley, E., Murray, S., Pope, E., Stephenson, D., Sharpe, M., Bingham, S., Jackson, D.: 2015, Verification of space weather forecasts using terrestrial weather approaches. *AGU Fall Meeting Abs.*, SH21B-2412. ADS.

Hirtenfellner-Polanec, W., Temmer, M., Pötzi, W., Freislich, H., Veronig, A.M., Hanslmeier, A.: 2011, Implementation of a calcium telescope at Kanzelhöhe Observatory (KSO). *Cent. Eur. Astrophys. Bull.* **35**, 205. ADS.

Jolliffe, I.T., Stephenson, D.B. (eds.): 2012, *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, Wiley, New York. DOI.

Kirk, M.S., Balasubramaniam, K.S., Jackiewicz, J., McNamara, B.J., McAteer, R.T.J.: 2013, An automated algorithm to distinguish and characterize solar flares and associated sequential chromospheric brightenings. *Solar Phys.* **283**, 97. DOI. ADS.

Kubo, Y., Den, M., Ishii, M.: 2017, Verification of operational solar flare forecast: case of Regional Warning Center Japan. *J. Space Weather Space Clim.* **7**(27), A20. DOI. ADS.

Murphy, A.H.: 1991, Probabilities, odds, and forecasts of rare events. *Weather Forecast.* **6**(2), 302. DOI.

Murphy, A.H.: 1996, The Finley affair: a signal event in the history of forecast verification. *Weather Forecast.* **11**(1), 3. DOI.

Otruba, W., Freislich, H., Hanslmeier, A.: 2008, Kanzelhöhe Photosphere Telescope (KPT). *Cent. Eur. Astrophys. Bull.* **32**, 1. ADS.

Otruba, W., Pötzi, W.: 2003, The new high-speed Hα imaging system at Kanzelhöhe Solar Observatory. *Hvar Obs. Bull.* **27**, 189. ADS.

Piazzesi, R., Berrilli, F., Del Moro, D., Egidi, A.: 2012, Algorithm for real time flare detection. *Mem. Soc. Astron. Ital. Suppl.* **19**, 109. ADS.

Pötzi, W., Veronig, A.M., Riegler, G., Amerstorfer, U., Pock, T., Temmer, M., Polanec, W., Baumgartner, D.J.: 2015, Real-time flare detection in ground-based Hα imaging at Kanzelhöhe Observatory. *Solar Phys.* **290**, 951. DOI. ADS.

Priest, E.R., Forbes, T.G.: 2002, The magnetic nature of solar flares. *Astron. Astrophys. Rev.* **10**, 313. DOI. ADS.

Qahwaji, R., Ahmed, O., Colak, T.: 2010, Automated feature detection and solar flare prediction using SDO data. *COSPAR Sci. Assem.* **38**, 2877. ADS.

Qu, M., Shih, F.Y., Jing, J., Wang, H.: 2003, Automatic solar flare detection using MLP, RBF, and SVM. *Solar Phys.* **217**, 157. DOI. ADS.

Riegler, G.E.: 2013, Flare and filament detection in Hα solar image sequences. Master thesis, Graz Univ. of Technology. https://online.tugraz.at/tug_online/wbAbs.showThesis?pThesisNr=52944&pOrgNr=2376#.

Riegler, G., Pock, T., Pötzi, W., Veronig, A.: 2013, Filament and flare detection in Hα image sequences. *ArXiv e-prints*. ADS.

Sammis, I., Tang, F., Zirin, H.: 2000, The dependence of large flare occurrence on the magnetic structure of sunspots. *Astrophys. J.* **540**, 583. DOI. ADS.

Scheffler, H., Elsasser, H.: 1990, *Physik der Sterne und der Sonne*, BI-Wissenschaftsverlag, Mannheim. ADS.

Svestka, Z.: 1976, *Solar Flares*, Springer, Berlin/Heidelberg, 415. ADS.

Veronig, A.M., Pötzi, W.: 2016, Ground-based observations of the solar sources of space weather. In: Dorotovic, I., Fischer, C.E., Temmer, M. (eds.) *Coimbra Solar Physics Meeting: Ground-based Solar Observations in the Space Instrumentation Era*, *Astron. Soc. Pacific Conf. Ser.* **504**, 247. ADS.

Veronig, A., Steinegger, M., Otruba, W., Hanslmeier, A., Messerotti, M., Temmer, M., Brunner, G., Gonzi, S.: 2000, Automatic image segmentation and feature detection in solar full-disk images. In: Wilson, A. (ed.) *The Solar Cycle and Terrestrial Climate, Solar and Space Weather*, *ESA* **SP-463**, ESA, Noordwijk, 455. ADS.