Astronomy & Astrophysics manuscript no. paper
May 30, 2018

Finding flares in *Kepler* data using machine-learning tools

Krisztián Vida[1] and Rachael M. Roettenbacher[2]

[1] Konkoly Observatory, MTA CSFK, H-1121 Budapest, Konkoly Thege M. út 15-17, Hungary
e-mail: vidakris@konkoly.hu
[2] Department of Astronomy, Stockholm University, SE-106 91 Stockholm, Sweden

Received March 1, 2018; accepted March 1, 2018

ABSTRACT

Context. Archives of long photometric surveys, such as the *Kepler* database, are a great basis for studying flares. However, identifying the flares is a complex task; it is easily done in the case of single-target observations by visual inspection, but is nearly impossible for several year-long time series for several thousand targets. Although automated methods for this task exist, several problems are difficult (or impossible) to overcome with traditional fitting and analysis approaches.
Aims. We introduce a code for identifying and analyzing flares based on machine-learning methods, which are intrinsically adept at handling such data sets.
Methods. We used the RANSAC (RANdom SAmple Consensus) algorithm to model light curves, as it yields robust fits even in case of several outliers, such as flares. The light curves were divided into search windows, approximately on the order of the stellar rotation period. This search window was shifted over the data set, and a voting system was used to keep false positives to a minimum: only those flare candidate points were kept that were identified as a flare in several windows.
Results. The code was tested on short-cadence *K2* observations of TRAPPIST-1 and on long-cadence *Kepler* data of KIC 1722506. The detected flare events and flare energies are consistent with earlier results from manual inspections.

Key words. Methods: data analysis – Techniques: photometric – Stars: activity – Stars: flare – Stars: late-type – Stars: low-mass

1. Introduction

Flares are energetic eruptions that occur as a result of magnetic field line reconnection. These events can be found in almost all types of main-sequence stars, including hot and cool stars (Švanda & Karlický 2016; Shibayama et al. 2013); but flares are most numerous in low-mass, late-type M dwarfs (Walkowicz et al. 2011; Vida et al. 2016).

These energetic events have received increased interest since the advent of exoplanet research, as flares can have strong, deleterious effects on orbiting planets (Khodachenko et al. 2007; Yelle et al. 2008). Flares can also continuously transform exoplanetary atmospheres, which is disadvantageous for hosting life (see Vida et al. 2017; Roettenbacher & Kane 2017 and references therein).

Currently used definitions of the habitable zone are based only on the stellar irradiation and the distance of the planet from the host star. As flares can have strong effects on planetary environments, these definitions will likely need to be revised for a more accurate definition of habitability in order to include the effects of stellar activity. To do this, and to better understand stellar magnetism itself, it is essential to characterize flares: events need to be properly identified, and their strength and frequency need to be determined.

The data from the *Kepler* satellite proved to be a great resource for stellar activity research because they provide an almost continuous data set of unprecedented precision over four years from about 160 000 targets, which include thousands of active stars (e.g., Basri et al. 2010). Studies have been performed to understand stellar activity of individual stars (e.g., Roettenbacher et al. 2013) and of classes of stars (e.g., McQuillan et al. 2014; Davenport 2016).

Significant strides were made to detect the flares that are contained in the *Kepler* archive by Davenport (2016), who identified events in the light curves by detecting the shape of a flare. However, this method can misidentify other astrophysical phenomena as flares (e.g., KIC 1572802, an RR Lyrae star). Of course, there is no single perfect way to accurately detect and classify all flares: the diversity of observations (e.g., short- and long-cadence *Kepler* data or ground-based observations) and of the events themselves (the flare length and complexity due to multiple nearly simultaneous flaring events can result in several light-curve shapes) make the automated search for these eruptions a difficult task. A manual identification of flares is also impossible in practice for a large number of observations, as in the *Kepler* archive.

In this paper, we present an algorithm that is based on machine-learning, with which we identify flares in light curves[1], and we present our application to the flaring, planet-hosting star TRAPPIST-1, a popular target for habitability studies, land KIC 1722506, a rotationally variable star (e.g., Debosscher et al. 2011).

2. Flare-finding algorithm

2.1. Determining the stellar rotation period

The first step of our FLAre deTection With Ransac Method (FLATW'RM) algorithm is to determine the rotation period of the light curve, which FLATW'RM accomplishes by taking the photometric modulation of spotted active stars into account. The

[1] The code is available at https://github.com/vidakris/flatwrm/

Article published by EDP Sciences, to be cited as https://doi.org/10.1051/0004-6361/201833194

starspots, which are dark regions of suppressed convection of cool, active stars, rotate across the stellar surface in and out of view of the observer, causing periodic modulations to the light curve. Starspots are often longer-lived than sunspots, allowing for detection during multiple rotations. This provides a reliable approximation for the stellar rotation period.

Light-curve sections of approximately the length of the stellar rotation period (typically, on the order of days) are expected to be easily described by a relatively low-degree polynomial (as opposed to sections covering several rotations), and their lengths are longer than the timescale of a flare event (typically, on the order of hours). The light-curve sections are specifically defined such that a flare could be easily spotted in a data set by eye. The period search in FLATW'RM is made using `LombScargleFast`, the Lomb–Scargle periodogram implementation in `gatspy`[2]. For further analysis, these light-curve windows (with a length of $1.5 \times P$ by default, where P is the period found with the Lomb–Scargle method, which is generally assumed to be the rotation period) are used. Each light-curve window must also be standardized: it is a common requirement for many machine-learning estimators that individual features should look more or less like normally distributed data (Gaussian with zero mean and unit variance, see, e.g., Pedregosa et al. 2011; Müller & Guido 2017). For our purposes, it is enough to transform only the time axis by removing its mean and scaling the light curve by its standard deviation, since the scaling of the brightness variation is just a multiplicative factor in the coefficients of the polynomial used to fit the light-curve window.

2.2. Determining the order of the polynomial fit

In machine-learning, one major problem is determining the complexity of the model used to fit the data in order to avoid under- or overfitting (Müller & Guido 2017). In the case of underfitting, the model does not describe the data well, while in the case of overfitting, a too-complex model is used that tries to fit too many data points individually. While it might fit a training data set well, this will describe test data and future measurements poorly. This problem is generally solved by a cross-validation method. An example of a basic approach is k-fold cross-validation, where the training data are split into k smaller sets (so-called folds), and the model is trained on $k-1$ folds of the data. The remaining part of the data is used for validating the model. The sets are usually created by selecting random samples of the initial data set, but this is not very useful with time series. In these cases, sets of consecutive data points are therefore used (Pedregosa et al. 2011). Lengthy data sets, such as *Kepler* light curves, make it impractical to use all the available data for cross-validation, or to find the optimal polynomial order for each segment. Therefore, we selected a sample (five, by default) of light-curve windows from the observations, and performed a grid search of fit parameters on them to select the best model to describe the given data based on the median absolute error regression loss. In most cases, a polynomial up to ≈10th degree is sufficient to fit rotational modulation of the data in the window.

2.3. Outlier detection and selection of flare candidates

To model light curves, we used the RANdom SAmple Consensus (RANSAC) algorithm, as it is designed to give a robust fit to data with several outliers (Bolles & Fischler 1981). RANSAC is an iterative method that assumes that the data consist of inlier and outlier points (generally noise, but also flare events, in this case). The algorithm works as follows: first, a sample random subset is generated from the input data set, which is fit by the model. Then, the algorithm checks which elements of the original data set fit this model based on the residuals. The points that fit the model are considered inliers for the given iteration. These steps are iterated either a maximum number of given times or until one of the stop criteria is met (this can be a given number of inlier points or a stop score by a given metrics). The final model is based on all inlier samples (also called a consensus set) of the previously determined best model. An example of the outlier selection by the algorithm is shown in the second plot of Fig. 1.

We found that while RANSAC gives a good fit even for a light-curve section with several flare events, the marked outliers are not reliable enough for searching for flare candidates alone (see Fig. 1): it sometimes also marks the beginning of the light-curve windows. Thus we only used the RANSAC estimate of the inlier points for statistics and calculated the standard deviation of the light curve (with the rotational modulation and most of the flare points removed). We considered those points as first-order flare candidates that were above a given detection level (above 3σ by default). To achieve more robust results, we shifted the

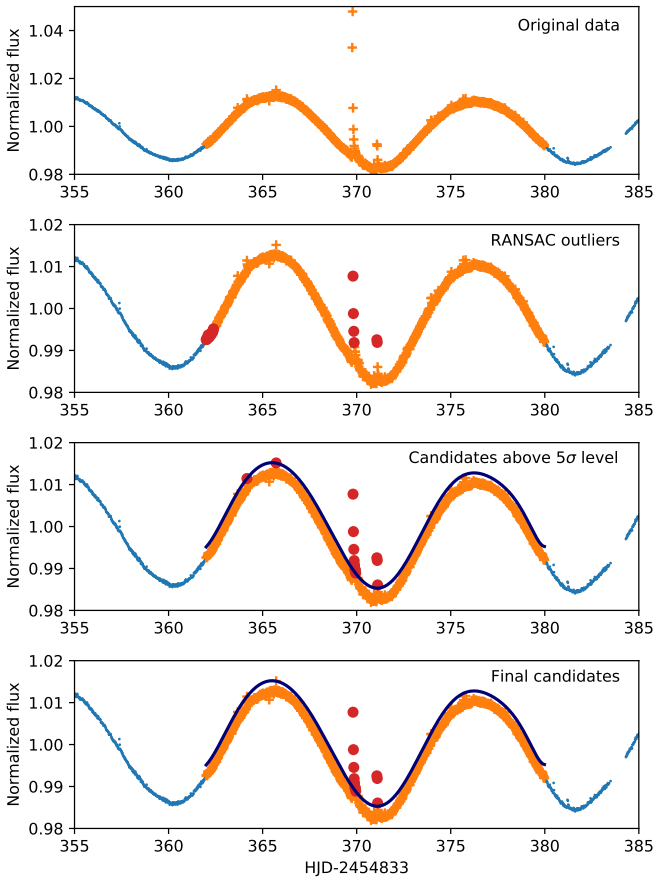

Fig. 1. Demonstration of the algorithm on a light-curve section of KIC 1722506. The top plot shows the original light-curve section. The second plot shows the outlier points found by RANSAC, marked with red dots. The third plot shows the 5σ detection level from the fit (continuous line) and the flare candidate points. In the bottom plot, the final flare candidates are shown, which have more than a given number (three, in this case) of consecutive data points. These points will get a vote for this light-curve section, indicating that the feature likely is a flare.

[2] `http://www.astroml.org/gatspy/`

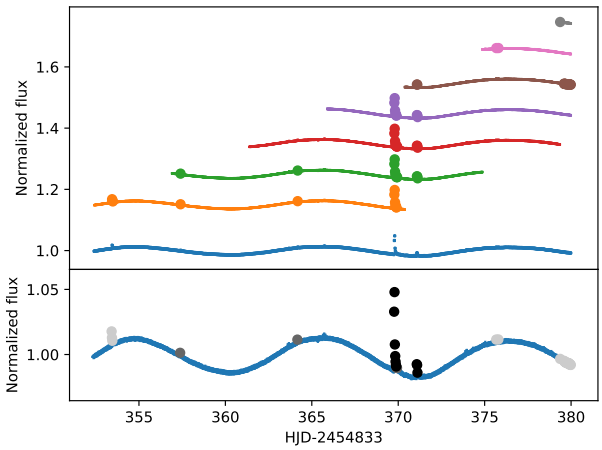

Fig. 2. Demonstration of the voting algorithm on a light-curve section of KIC 1722506. The top plot shows the original light curve (in blue, near a normalized flux of 1), and each light-curve segment in individual windows tested by FLATW'RM (each a different color), and the flare candidates are marked for the given run with circles. In the bottom plot, the candidates with one (light gray), two (medium gray), and at least three (black) votes are plotted. In this setup, the flares plotted in black are kept as final flare candidates.

search window through the light curve, by default in steps of one-fourth of the light-curve window (see Fig. 2, e.g.). In this way, every light-curve point was analyzed multiple times (with the exception of those that are at the beginning or at the end of the data set or large observational gaps), and if a light-curve point is considered as an outlier, that is, a flare point candidate, it received a "vote". A light-curve section was considered a flare when it received enough votes ($\geq$ 3, by default) in overlapping search windows and had at least a given number (2, by default) of consecutive points. We note that this last step was performed only after evaluating each light-curve segment. These criteria are basically the same as those defined by Equations 3 a–d of Chang et al. (2015) and are consistent with adaptations made by Davenport (2016). When running FLATW'RM from command line, the user can change the number of flare points needed for a flare, the detection level of the flares, and the full with at half maximum (FWHM) that is used for the analytic fit of the events. Optionally, the rotation period can be given as an input to run the code faster, or to fix it to a chosen value if the rotational modulation is too weak, and the polynomial degree can also be fixed. The number of votes needed for a flare candidate to be kept, the window size (compared to the rotation period), and the step length with which the search window is shifted cannot be changed from command line, but can be easily modified if the flare-finding function is imported to another code.

2.4. Summary of the selection process

For clarity, we explicitly state the procedure that FLATW'RM steps through in locating the flares in the light curve.

- A period search is first performed on the input light curve.
- The input light curve is divided into windows of $1.5 \times P$ that can be effectively fit with polynomials.
- Each light-curve window is modeled by the RANSAC algorithm to find the best-fitting polynomial.
- The data points are designated as inliers or outliers.
- The light-curve model is subtracted from the light-curve window.
- The standard deviation of the light curve is determined based only on the inlier data points.
- The data points that are above the given detection limit receive a vote as a flare candidate for the given window.
- After each window is analyzed, only those flare candidate data points are kept that have a given number of votes.
- Events that have at least a given number of selected candidate points are marked as flare events.

2.5. Fitting an analytic model

As output we considered two options: (1) the beginning and ending times of the flare and the maximum time and light-curve amplitude for each event, or (2) an analytic model can be fit to the data. In the latter case, the observed light curve is modeled again with the RANSAC algorithm (this time centered on the event) to remove the effect of rotational modulation, leaving only the light-curve changes caused by flares. This data set is then fit by the classical single-peak flare model defined by Davenport et al. (2014). The parameters for this function are the time of the flare peak, the FWHM (i.e., the timescale of the flare), and the amplitude of the flare. As an initial guess, we took the middle flare data point as the peak time, the amplitude of the selected light-curve data point, and one hour as the timescale of the flare, which fit most of the events (but this can be changed by the user). In the case of weak eruptions near stronger events, however, the fit might be distorted, and the fit could converge to the event with higher amplitude, effectively ignoring the weaker event. In the case of lower sampling, the fits could yield very high peaks (as a consequence of exponential decay) if only the declining phase is measured. To estimate flare energies, the integrated intensity (also known as equivalent duration) was also calculated for each event.

3. Caveats

While this method is an improvement over previous automated flare-detection efforts, we acknowledge that there are still a number of difficulties. Here, we list the significant caveats for using FLATW'RM.

- In the application of the code on a large number of different light curves (e.g., mixing short- and long-cadence *Kepler* data) without adjusting the searching parameters.
- Long and complex flare events (e.g., those observed by Kóspál et al. 2018, where eruptions were likely caused both by accretion and magnetic field reconnection) can cause failed outlier detection if there are not enough inlier points in the search window.
- Analytic fits to the events can yield unexpected results, especially in the case of long-cadence *Kepler* data, where only the exponential decay of the flare is observed (cf. Fig. 4). We emphasize that the analytic fits always need to be checked or the output from just the light curve should be used and analytic fitting be ignored in dubious cases.
- A weak rotation signal may not be properly identified with FLATW'RM's Lomb–Scargle period search and could a yield problematic light-curve search window length. Additionally, FLATW'RM could fail to find a rotation period. In these cases, the rotation period (or a reasonable size for the light-curve window) should be given as an input to the code.

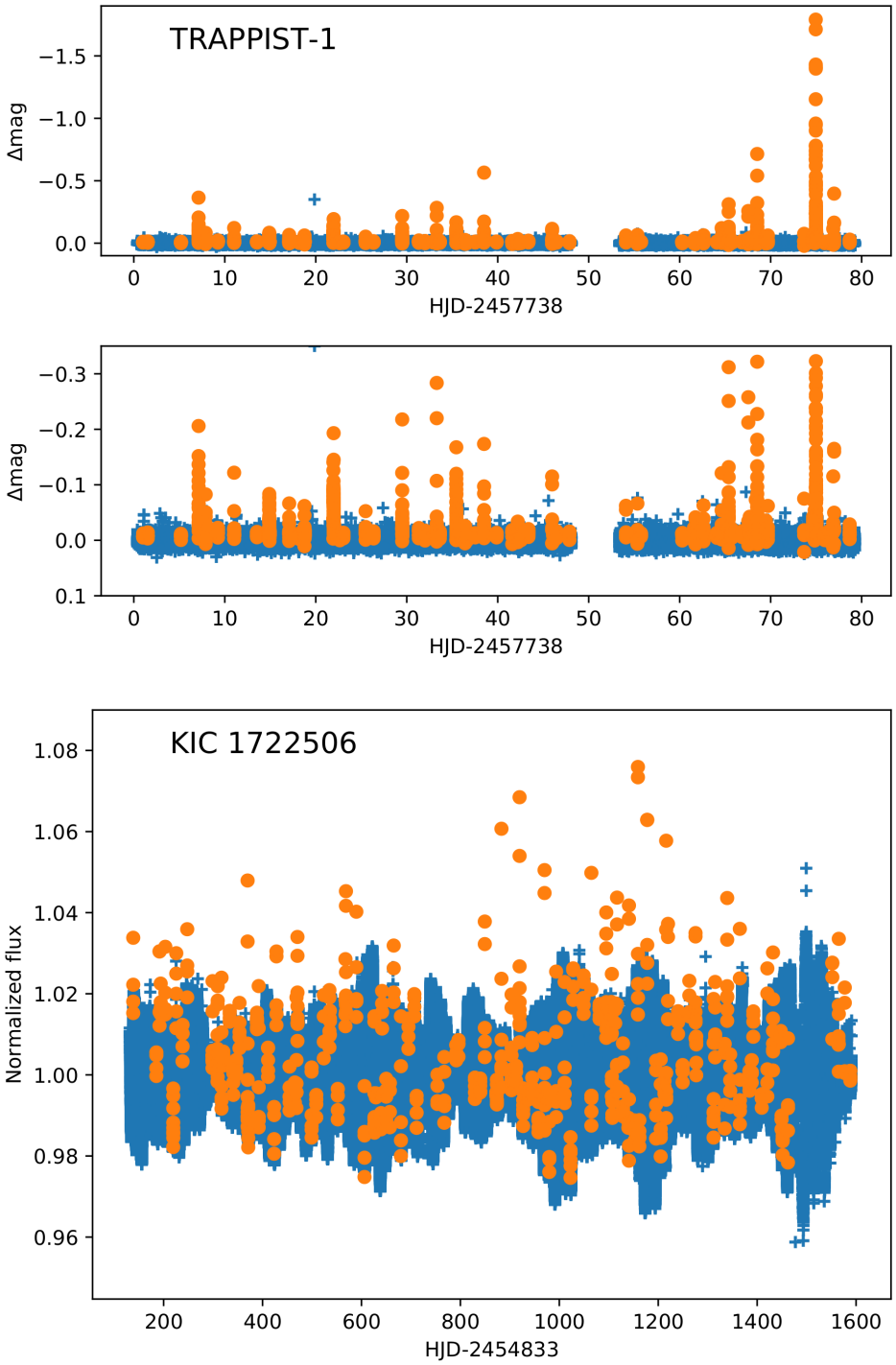

Fig. 3. Top: Selected flare candidates in the TRAPPIST-1 short-cadence *K2* data. The upper panel shows the total light curve, the middle plot is zoomed-in to show smaller events. Bottom: Similar analysis, but for long-cadence Kepler data of KIC 1722506.

4. Testing the FLATW'RM code

As a demonstration of the FLATW'RM code, we analyzed two data sets from the *Kepler* telescope: a short-cadence *K2* observation of TRAPPIST-1, and a long-cadence light curve of KIC 1722506 (the data were obtained in the 420–900 nm wavelength range with the maximum spectral response at 575 nm). In the case of TRAPPIST-1, we set the minimum number of data points needed for a flare (N3 in Chang et al. 2015) to 5, and the detection limit to 5σ. In the case of KIC 1722506 we used N3=3 with a detection limit of 3σ. The detected flare events are shown in Figure 3. Two samples from the recovered flares that demonstrate the analytic model fit by FLATW'RM are plotted in Figure 4.

With these parameters, FLATW'RM found 35 and 126 events in the case of TRAPPIST-1 and KIC 1722506, respectively. As a comparison, in the case of TRAPPIST-1, visual inspection by Vida et al. (2017) revealed 42 events.

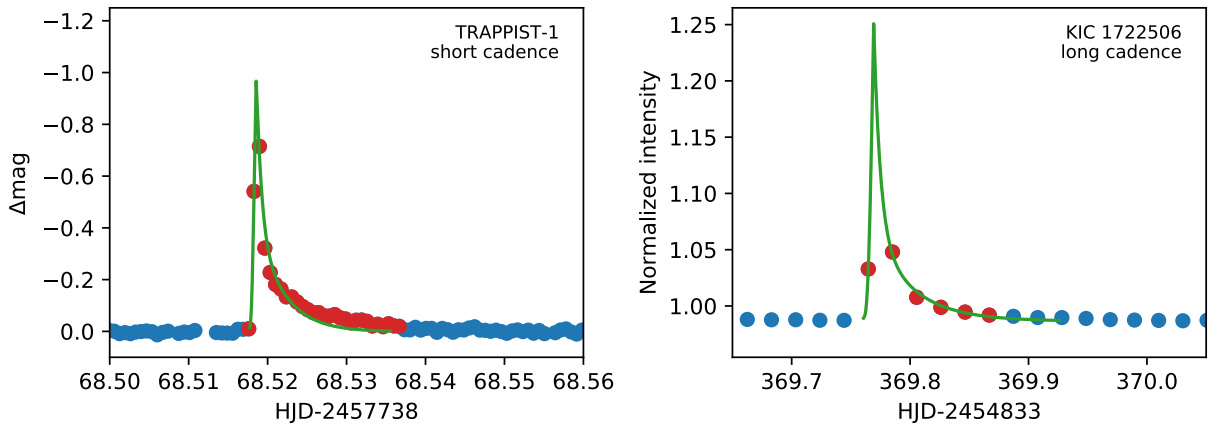

Fig. 4. Two samples from the recovered flares, from the short-cadence TRAPPIST-1 and the long-cadence KIC 1722506 data. Red points show the points selected for flares, the green line indicates the fitted analytical model from Davenport et al. (2014).

To estimate the flare energies, we followed the method of Kővári et al. (2007), which is based on integrating the flare intensity during the event:

$$\varepsilon_f = \int_{t_1}^{t_2} \left(\frac{I_{0+f}(t)}{I_0} - 1 \right) dt,$$

where t_1 and t_2 are the beginning and end times of an event, and I_{0+f} and I_0 are the intensities with and without a flare (i.e., their fraction is the normalized intensity). The integral above will yield the relative flare energy (or equivalent duration). During the analysis, FLATW'RM fits each flare event as described above and produces a spotless light curve that contains only the eruption itself, which has to be integrated over the duration of the event in order to obtain the relative flare energy. From this, the flare energy in the observed bandpass (E_f) can be calculated by multiplying by the quiescent flux ($F_\star$):

$$E_f = \varepsilon_f F_\star.$$

We estimated the quiescent flux by assuming blackbody radiation with an effective temperature of $T_{\rm eff} = 2550$K and stellar radius of $R = 0.117 R_\odot$ for TRAPPIST-1 (Gillon et al. 2016), and $T_{\rm eff} = 4270$K and $R = 0.845 R_\odot$ for KIC 1722506 (taken from the *Kepler* Input Catalogue[3]; Brown et al. 2011). The blackbody power function $\mathcal{F}(\lambda)$ was convolved with the *Kepler* response function $S_{\rm Kp}$, and integrated over wavelength to obtain the quiescent flux $F_\star$ in the *Kepler* passband:

$$F_\star = \int_{\lambda_1}^{\lambda_2} 4\pi R^2 \mathcal{F}(\lambda) S_{\rm Kp}(\lambda) d\lambda.$$

Following the analysis of Vida et al. (2017), we fit the cumulative flare frequency distribution (plotted in Fig. 5) with a linear fit that can be expressed as

$$\log \nu = a + \beta \log E,$$

where ν is the cumulative number of flares with energy higher than E. The slope of a linear fit yields $\beta = 1 - \alpha$, where α is often used to determine if flare energy dissipates by thermal, nonthermal, or magnetic processes (see Aschwanden et al. 2016). Alternatively, α can be determined using the maximum likelihood estimator (see Gizis et al. 2017):

$$(\alpha - 1) = n \left[\sum_{i=1}^{n} \ln \frac{E_i}{E_{\rm min}} \right]^{-1}.$$

Here, n is the number of detected events, while E_i and $E_{\rm min}$ are the individual and the lowest flare energies, respectively. According to Gizis et al. (2017), this result should be multiplied by $\frac{n-2}{n}$ for small samples to correct for the bias. The linear fit to the cumulative distribution yielded $\alpha = 1.53$ for TRAPPIST-1, while the maximum likelihood estimator gives $\alpha = 1.47$ (corrected for sample size), close to $\alpha = 1.59$, the value found by Vida et al. (2017). For KIC 1722506, the linear fit yielded $\alpha = 1.38$ for energies $\log E < 34.5$ (E given in ergs), where the distribution is close to linear, while the maximum likelihood estimator gave $\alpha = 1.50$.

5. Summary

We presented the FLATW'RM code, which uses machine-learning methods to identify flare events in light curves and calculates their relative energies. Characterizing these energetic events is crucial, since they can shape circumstellar environments, especially in the realm of planetary habitability. In the case of many targets and large data sets, as with the *Kepler* database, manual inspection is impossible, but machine-learning tools can help astronomers to effectively analyze such data. In the future, we plan to apply this method to a large set of *Kepler* stars in order to obtain a new view that is independent of currently available works.

Acknowledgements. The authors thank A. Moór for useful discussion and the anonymous referee for their careful review of this work. The authors acknowledge the Hungarian National Research, Development and Innovation Office grants OTKA K-109276, OTKA K-113117, and supports through the Lendület-2012 Program (LP2012-31) of the Hungarian Academy of Sciences, and the ESA PECS Contract No. 4000110889/14/NL/NDe. KV is supported by the Bolyai János Research Scholarship of the Hungarian Academy of Sciences. This work has used *K2* data from the proposal number GO12046. Funding for the *Kepler* and *K2* missions is provided by the NASA Science Mission directorate.

[3] `http://archive.stsci.edu/kepler/`, see also Kepler Mission Team 2009

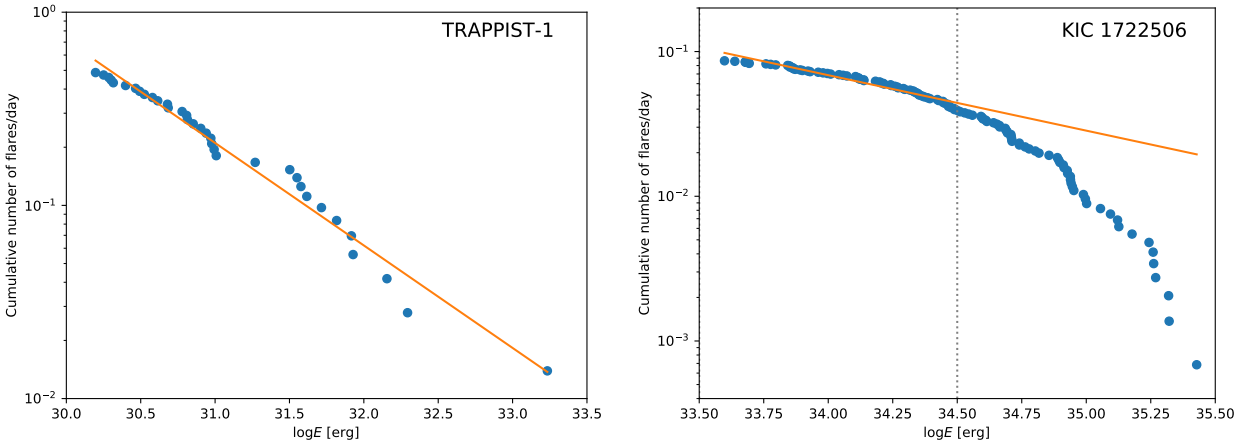

Fig. 5. Cumulative flare frequency distribution as a function flare energy, log E where E is given in ergs, of the TRAPPIST-1 *K2* data (left, cf. Vida et al. 2017) and for long-cadence Kepler data of KIC 1722506 (right). In the right plot, the dashed line marks the upper energy limit used for the fit.

.

References

Aschwanden, M. J., Holman, G., O'Flannagain, A., et al. 2016, ApJ, 832, 27
Basri, G., Walkowicz, L. M., Batalha, N., et al. 2010, ApJ, 713, L155
Bolles, R. C. & Fischler, M. A. 1981, in Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2, IJCAI'81 (San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.), 637–643
Brown, T. M., Latham, D. W., Everett, M. E., & Esquerdo, G. A. 2011, AJ, 142, 112
Chang, S.-W., Byun, Y.-I., & Hartman, J. D. 2015, ApJ, 814, 35
Davenport, J. R. A. 2016, ApJ, 829, 23
Davenport, J. R. A., Hawley, S. L., Hebb, L., et al. 2014, ApJ, 797, 122
Debosscher, J., Blomme, J., Aerts, C., & De Ridder, J. 2011, A&A, 529, A89
Gillon, M., Jehin, E., Lederer, S. M., et al. 2016, Nature, 533, 221
Gizis, J. E., Paudel, R. R., Mullan, D., et al. 2017, ApJ, 845, 33
Kepler Mission Team. 2009, VizieR Online Data Catalog, 5133
Khodachenko, M. L., Ribas, I., Lammer, H., et al. 2007, Astrobiology, 7, 167
Kővári, Zs., Vilardell, F., Ribas, I., et al. 2007, Astronomische Nachrichten, 328, 904
Kóspál, A., Ábrahám, P., Zsidi, G., et al. 2018, ApJ, submitted
McQuillan, A., Mazeh, T., & Aigrain, S. 2014, ApJS, 211, 24
Müller, A. C. & Guido, S. 2017, Introduction to Machine Learning with Python, 1st edn. (O'Reilly Media, Inc.)
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825, documentation available at `http://scikit-learn.org/stable/_downloads/scikit-learn-docs.pdf`
Roettenbacher, R. M. & Kane, S. R. 2017, ApJ, 851, 77
Roettenbacher, R. M., Monnier, J. D., Harmon, R. O., Barclay, T., & Still, M. 2013, ApJ, 767, 60
Shibayama, T., Maehara, H., Notsu, S., et al. 2013, ApJS, 209, 5
Švanda, M. & Karlický, M. 2016, ApJ, 831, 9
Vida, K., Kővári, Zs., Pál, A., Oláh, K., & Kriskovics, L. 2017, ApJ, 841, 124
Vida, K., Kriskovics, L., Oláh, K., et al. 2016, A&A, 590, A11
Walkowicz, L. M., Basri, G., Batalha, N., et al. 2011, AJ, 141, 50
Yelle, R., Lammer, H., & Ip, W.-H. 2008, Space Sci. Rev., 139, 437