



The K2 M67 Study: Establishing the Limits of Stellar Rotation Period Measurements in M67 with K2 Campaign 5 Data

Rebecca Esselstein¹, Suzanne Aigrain¹ , Andrew Vanderburg^{2,3,8} , Jeffrey C. Smith^{4,5}, Soren Meibom³, Jennifer Van Saders⁶, and Robert Mathieu⁷

¹ Department of Astrophysics, University of Oxford Denys, Wilkinson Building, Keble Rd. Oxford, OX1 3RH, UK; rebecca.esselstein@physics.ox.ac.uk

² Department of Astronomy, University of Texas at Austin 2515 Speedway, Stop C1400 Austin, TX 78712-1205, USA

³ Harvard-Smithsonian Center for Astrophysics, 60 Garden St., Cambridge, MA 02138, USA

⁴ SETI Institute, 189 Bernardo Ave., Suite 100, Mountain View, CA 94043, USA

⁵ NASA Ames Research Center, Moffett Field, CA 94035, USA

⁶ Institute for Astronomy—Mānoa, University of Hawai‘i, 2680 Woodlawn Dr., Honolulu, HI 96822, USA

⁷ Department of Astronomy, University of Wisconsin-Madison, 475 N., Charter St., Madison, WI 53706, USA

Received 2018 April 4; revised 2018 April 30; accepted 2018 April 30; published 2018 June 5

Abstract

The open cluster M67 offers a unique opportunity to measure rotation periods for solar-age stars across a range of masses, potentially filling a critical gap in the understanding of angular momentum loss in older main sequence stars. The observation of M67 by NASA K2 Campaign 5 provided light curves with high enough precision to make this task possible, albeit challenging, as the pointing instability, 75 day observation window, crowded field, and typically low-amplitude signals mean that determining accurate rotation periods on the order of 25–30 days is inherently difficult. Lingering, non-astrophysical signals with power at ≥ 25 days found in a set of Campaign 5 A and F stars compounds the problem. To achieve a quantitative understanding of the best-case scenario limits for reliable period detection imposed by these inconveniences, we embarked on a comprehensive set of injection tests, injecting 120,000 sinusoidal signals with periods ranging from 5 to 35 days and amplitudes from 0.05% to 3.0% into real Campaign 5 M67 light curves processed using two different pipelines. We attempted to recover the signals using a normalized version of the Lomb–Scargle periodogram and setting a detection threshold. We find that, while the reliability of detected periods is high, the completeness (sensitivity) drops rapidly with increasing period and decreasing amplitude, maxing at a 15% recovery rate for the solar case (i.e., 25 day period, 0.1% amplitude). This study highlights the need for caution in determining M67 rotation periods from Campaign 5 data, but this can be extended to other clusters observed by K2 (and soon, *TESS*).

Key words: methods: data analysis – open clusters and associations: individual (M67) – stars: rotation – stars: solar-type – techniques: photometric

1. Introduction

The measurement of rotation periods in open clusters can be a powerful and convenient tool in understanding stellar angular momentum evolution; it offers the opportunity to connect a star’s age to its rotation period and color or, alternatively, mass. Skumanich (1972) first noticed the relationship between period and age, though the color–period relation has long been observed in the Hyades, where there is a notable dependency of stellar rotation period on spectral type for stars cooler than about mid-F (Radick et al. 1987; Delorme et al. 2011). Kawaler (1988) concluded that, by the age of the Hyades, a star’s initial angular momentum no longer plays a significant role in determining its rotation period, so another mechanism, possibly stellar winds coupled with magnetic fields, “brakes” the rotation of a star as it ages. Kawaler (1989) then acknowledged the use of rotation, a direct observable independent of distance, as a potential age estimator. Barnes (2003) developed the first semi-empirical model for deriving ages from the colors and rotation periods of FGK dwarfs, introducing the term “gyrochronology.” Following the improvements of Barnes (2007), there was a need for precise rotation period measurements in clusters to test gyrochronology and probe its limitations. The availability of large field-of-view (FOV) CCD cameras on smaller (1 m class) telescopes enabled the

measurement of thousands of rotation periods for stars in young open clusters via the rotational modulation of stellar flux (Irwin et al. 2007, 2009; Collier Cameron et al. 2009; Hartman et al. 2009, 2010; Meibom et al. 2009, 2011; James et al. 2010), using co-eval populations to enable the creation of period–age–mass (P–t–M) surfaces based on different gyrochronology models.

NASA’s *Kepler* mission, dedicated to the detection of transiting exoplanets of nearby bright stars, offered unprecedented precision for the study of stellar variability, particularly stellar rotation. Key *Kepler* large-scale rotation period studies include those of McQuillan et al. (2013a), Reinhold et al. (2013), McQuillan et al. (2013b) (letter), McQuillan et al. (2014), Nielsen et al. (2013), García et al. (2014), and Reinhold & Gizon (2015). These studies measured the rotation periods of tens of thousands of *Kepler* field stars, including planetary hosts. Along with the ground-based cluster surveys, they helped bring to light a number of interesting trends: for example, the puzzling bimodal period distribution shown by late K and M-dwarfs, as well as the existence of a correlation between stellar temperature, rotation period, and the spread thereof, which may be related to differential rotation, active region evolution, or a combination of the two.

The *Kepler* Cluster Study extended the P–t–M surface from ~600 Myr (Hyades) to 2.5 Gyr by using *Kepler* data to measure rotation periods of FGK dwarfs in NGC 6811

⁸ NASA Sagan Fellow.

(1 Gyr; Meibom et al. 2011) and NGC 6819 (2.5 Gyr; Meibom et al. 2015). These measurements agreed with the predictions of the most recent gyrochronology relations by Barnes (2010). Other approaches using *Kepler* data, however, have focused on asteroseismic age determination (García et al. 2014; Angus et al. 2015; van Saders et al. 2016), and these studies have called into question the behavior of stars at older ages. For instance, van Saders et al. (2016) attempts to explain the anomalously rapid rotation of old field stars by proposing a “weakened magnetic braking” theory. Their theory states that, around solar age, stars with masses similar to and above that of the Sun should begin to diverge from the “standard” gyrochronology relations, which their cooler counterparts are still expected to follow. There is a clear need for high-precision observations of solar-age clusters to test this new theory, alongside the established gyrochronology relations.

The best candidate for an in-depth rotation study on a solar-age cluster is M67. M67 is estimated to be ~ 4.3 Gyr old, and roughly 800–900 pc from the Sun (Carraro et al. 1994; Balaguer-Núñez et al. 2007; Geller et al. 2015). The cluster’s age is not its only similarity to the Sun; calculated metallicity values from the literature show that its [Fe/H] ranges from about -0.1 to 0.1 , with average values lying around -0.01 (Fan et al. 1996; Balaguer-Núñez et al. 2007; Taylor 2007; Jacobson et al. 2011, the most recent of which reports -0.01 ± 0.05). The members of M67 have been cataloged several times, based on proper motion (Sanders 1977; Loktin 2005; Yadav et al. 2008) and radial velocities (Pasquini et al. 2008; Yadav et al. 2008; Geller et al. 2015). Geller et al. (2015; hereafter G15) report 1278 candidate members, making M67 the richest solar-age open cluster that is relatively nearby. There is a general tendency for middle-aged, main-sequence stars to have longer activity cycles and thus lower variability amplitudes than their younger counterparts (Radick et al. 1995, 1998; Hempelmann et al. 1996; Baliunas et al. 1998; Mamajek & Hillenbrand 2008), making their signals more difficult to detect. Thus, high photometric precision is necessary for a study of rotation in M67. *K2*, the second phase of the *Kepler* mission following the loss of two of the spacecraft’s reaction wheels by 2013, presents such an opportunity. However, the mission’s reduced pointing accuracy, relative to *Kepler*, leads to lower photometric precision due to an increase in systematics associated with inter- and intra-pixel sensitivity variations, as well as aperture losses (Howell et al. 2014). Early estimates of *K2*’s photometric precision indicated, for stars with a magnitude of $V = 12$, a precision of approximately 400 parts per million (ppm) for the long-cadence (or 30 minute) observations, and 80 ppm over the course of 6 hr (Howell et al. 2014). Previously, *Kepler* reached a precision level of 10 ppm at a magnitude of $K_p = 10$ for 6 hr observations (Christiansen et al. 2012; Vanderburg & Johnson 2014).

Members of the community have since developed methods to improve the photometric precision of the *K2* light curves, particularly to correct the pointing-related variations. These include the Vanderburg & Johnson (hereafter VJ) pipeline (Vanderburg & Johnson 2014; Vanderburg et al. 2016), the *K2* Systematics Correction (K2SC) pipeline (Aigrain et al. 2015a, 2016), EVEREST (Luger et al. 2016), K2VARCAT (Armstrong et al. 2014, 2016), and the PSF-fitting used by Libralato et al. (2016) for *K2* observations of M35 and NGC 2158. Several of these pipelines match the photometric precision of *Kepler* for stars with $K_p = 12.5$ and perform within a factor of two of the original mission for fainter stars. This, combined with the much

more diverse sample of stars surveyed by *K2* compared to the *Kepler* prime mission, including numerous open clusters, makes *K2* data a treasure trove for stellar angular momentum evolution studies. Indeed, the *K2* mission has produced spectacular results for nearby, well-studied, young open clusters, including the Pleiades (Rebull et al. 2016a, 2016b; Stauffer et al. 2016) and Hyades (Douglas et al. 2016). More critical to this study, however, *K2* observed M67 during Campaign 5.⁹ Barnes et al. (2016) and Gonzalez (2016a, 2016b) have already used those data to measure rotation periods for M67 members and investigate the implications of their results in terms of gyrochronology. However, both studies worked only with stars observed using individual postage stamps, avoiding the crowded central regions of the cluster. Moreover, while Barnes et al. (2016) obtained a good match to pre-existing gyrochronology relations, this was based on a very small sample of 20 stars, with the majority of the reported periods not convincingly confirmed when examined by eye, based on the provided light curves. By contrast, Gonzalez (2016a) showed no dependence of rotation period on mass (color) and significant scatter for all masses. Furthermore, there are discrepancies in the measured periods for several stars in common between the two studies, and they inferred mutually inconsistent gyrochronological ages for M67 (4.2 ± 0.2 and 5.0 ± 0.2 Gyr, respectively).

In addition to exploiting only part of the data available, the published *K2* M67 rotation studies made specific choices in terms of light curve extraction, detrending, and period search methods, without investigating possible alternatives in any detail, and both methods involved a large component of human decision-making, which is difficult to reproduce. Furthermore, it is useful to consider just how challenging it is to determine accurate rotation periods in M67 data. Based on existing gyrochronology relations and what we know about the Sun itself, these stars are expected to have rotation periods on the order of 20–30 days, and amplitudes of a few tenths of a percent, at most. Even in the full *Kepler* field star sample of McQuillan et al. (2014), based on four years of continuous observations, there are relatively few detections in this area of parameter space, due either to sensitivity limits or the underlying period–amplitude distribution. The limited duration of a *K2* campaign (~ 75 days; about 2–3 rotation cycles for a typical M67 star) will only make the process of detecting accurate rotational modulation harder. In addition, the reduced photometric precision of *K2* light curves (relative to the *Kepler* prime mission) is compounded by the large *Kepler* pixel size, as well as the significant degree of crowding of cluster stars, which creates additional systematics.

In addition to the difficulties mentioned above, a sample of 89 hot A and F stars from *K2* Campaign 5 has highlighted the presence of 25 day+, non-astrophysical signals in the data, as can be seen in Figure 1. This figure shows the rotation periods of the data set, acquired using the autocorrelation function (ACF), the standard Lomb–Scargle periodogram, and a normalized version of the Lomb–Scargle periodogram (which we explain and implement in this study in Section 3). The hot star sample selection is detailed in Section 3.1. We should not expect periods much greater than ~ 5 days for these stars. The significant number of measurements from either the ACF or the standard Lomb–Scargle periodogram above 25 days shows that there is a substantial risk of making erroneous period

⁹ <https://keplerscience.arc.nasa.gov/k2-fields.html>

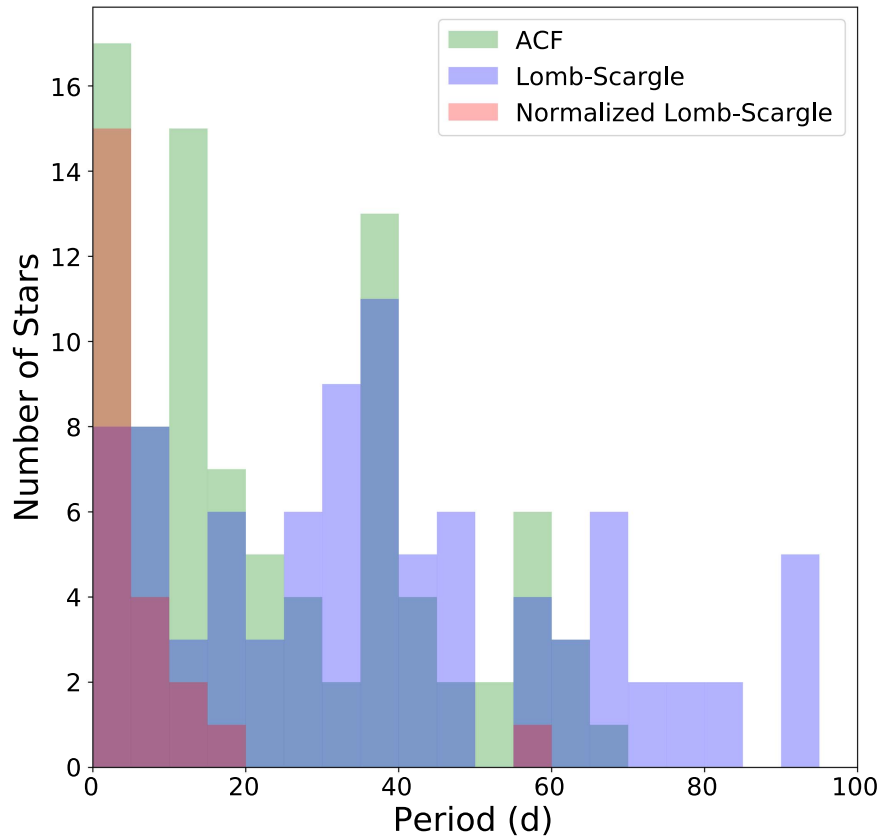


Figure 1. Periods acquired from a set of 89 *K2* Campaign 5 A and F stars, using the autocorrelation function (ACF) (green) and the Lomb–Scargle periodogram (blue). We also include a normalized version of the Lomb–Scargle periodogram (red), which we implement later in this paper, as a demonstration of how we can set detection thresholds and remove potentially anomalous measurements.

measurements on the order of 25–30 days in M67. All of this should lead to a healthy degree of a priori skepticism regarding any period detections in M67, for G and K stars at least, and any such detections need to be backed up by detailed completeness (sensitivity) and reliability tests.

Using all the cluster members and the aforementioned set of A and F stars observed during *K2* Campaign 5, we carefully compare the effects of different light curve extraction and detrending methods. We perform systematic tests with simulated, sinusoidal signals injected into the Campaign 5 M67 *K2* light curves, in order to properly evaluate the biases and best-case sensitivity limits of each detrending pipeline. This paper reports on the first set of injection tests carried out as part of our team’s *K2* M67 rotation study, which ultimately seeks to produce a list of periods from this critical cluster in which we can quantify our level of confidence. Later, a round of smaller injection tests will use more realistic signals, generated by the star spot models from Aigrain et al. (2015b), to help us reach that final goal.

The structure of the rest of this paper is as follows. Section 2 discusses the *K2* observations, light curve extraction, and light curve detrending. Section 3 describes the injection tests, including the generation of the injected signals, the period search method, and the definition of a systematic detection criterion. In that section, we also discuss how we handle intrinsically variable stars (i.e., stars with an astrophysical signal before injection) and what we consider a “valid” detection. The results are presented in Section 4, where we compute the completeness and reliability of the period search

as a function of period and amplitude, before they are discussed in Section 5. We summarize our conclusions in Section 6.

2. Observations and Light Curve Preparation

2.1. *K2* Observations

M67 was observed during *K2* Campaign 5, which commenced on 2015 April 27 and ended on July 10. The pointing was centered on R.A. = 130^h09^m27^s.53 and decl. = 16°49′46″61.¹⁰ Individual *K2* targets for Campaign 5 were proposed for observation by members of the community. Our proposal (PI. R. Matthieu) included 12,935 candidate members, with 12,521 drawn from the EPIC database, 266 from the 2MASS calibration field (see, e.g., Sarajedini et al. 2009), and the last 148 targets filling in those missing from the EPIC database around $K_p \sim 19$ (R. L. Gilliland 2018, private communication). Membership for stars brighter than $K_p = 15$ is based on radial velocity and proper motion, as explained in G15, taking those with membership probabilities of 20% or greater to make a more complete catalog. For the remainder, we accepted those stars whose photometric membership made them likely members, unless contradicted by known radial velocities or proper motions. These stars formed our master target list. M67 was located on channels 1 and 2 of CCD module 6, near the edge of the FOV. Figure 2 shows the entire Campaign 5 FOV, with M67 targets represented by the teal circles on CCD module 6, along with a scattering of orange diamonds that

¹⁰ See <https://keplerscience.arc.nasa.gov/k2-data-release-notes.html#k2-campaign-5>.

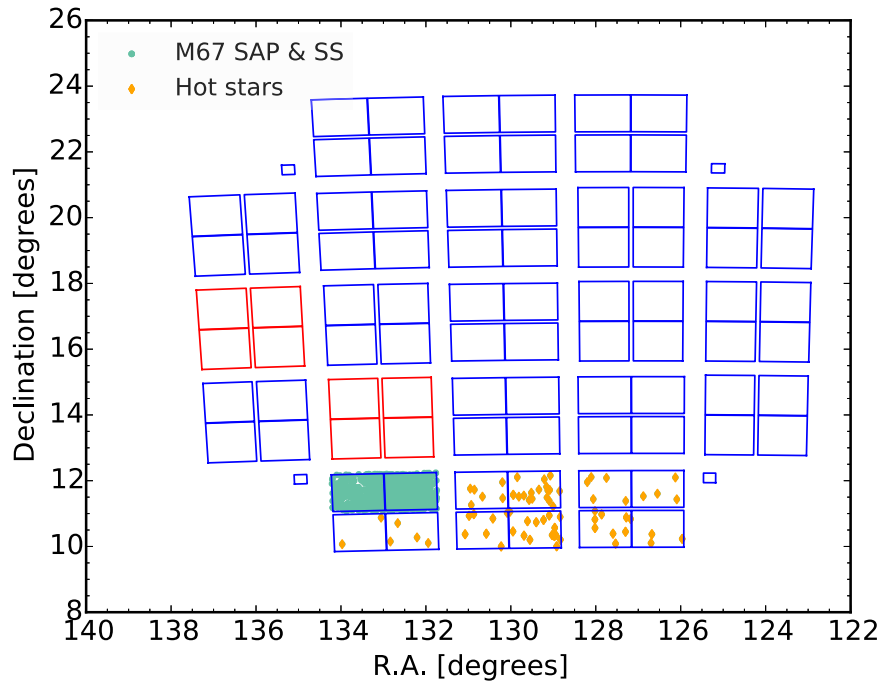


Figure 2. Location of the 1877 M67 SAP and 976 superstamp stars (teal circles) used in this study, along with an additional 89 hot stars (orange diamonds), on the *Kepler* Campaign 5 field of view.

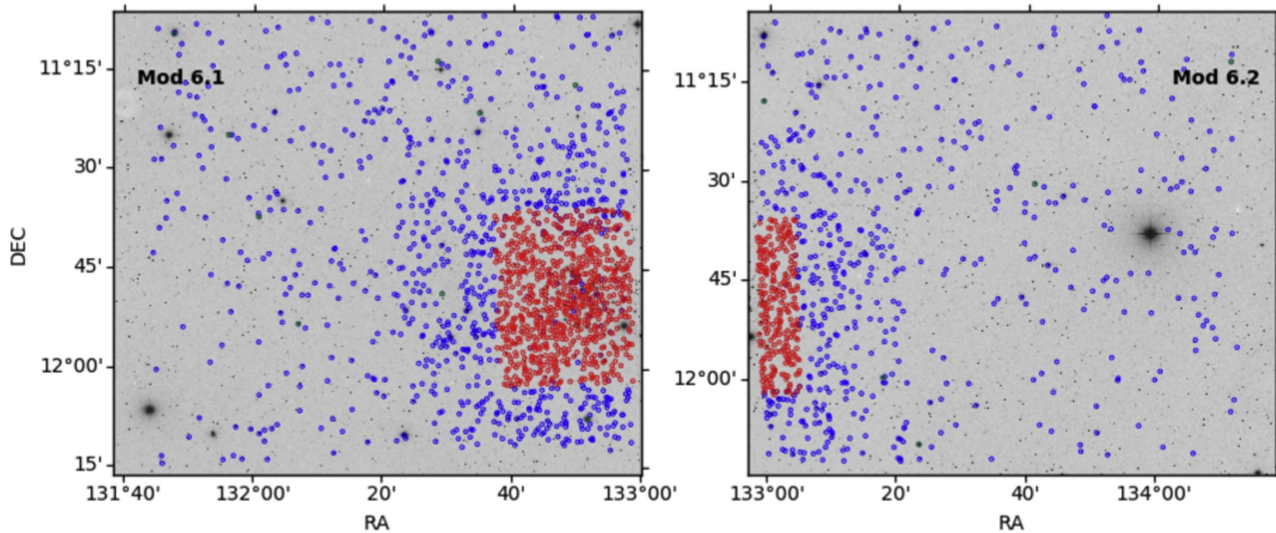


Figure 3. Zoomed-in look at the M67 SAP and superstamp stars, using two images of M67 retrieved from the ESO Online Digitized Sky Survey. The SAP stars are surrounded by blue circles, while the SS stars are surrounded by red circles.

represent the sample of hot stars used in this analysis, which will be discussed later.

The stars located in the less-crowded, outer parts of the cluster were observed in the standard manner, by downloading a small window of pixels centered on each star of interest. A total of $\sim 25,000$ individual star postage stamps were collected during Campaign 5; of these, 1877 fell on output channels 6.1 and 6.2. For the crowded inner regions of the cluster, a “superstamp” was used: a large, contiguous region created by juxtaposing many postage stamps. Approximately 2210 of the proposed M67 targets were located within the superstamp (the number of stars in the superstamp for which light curves are actually extracted depends on the method used). Figure 3

shows the locations of the M67 SAP sample (in blue) and the superstamp stars (in red) where they fall on channels 6.1 and 6.2 using two images of M67 taken from the ESO Online Digitized Sky Survey, centered on the channels’ respective pointing.

Most postage stamps were read out in “long-cadence” mode, or one observation every 30 minutes, with a total of 3663 cadences during the campaign. After basic reduction (correction of pixel-level instrumental effects), each series of postage stamps was combined into a Target Pixel File (TPF), and the TPFs were made publicly available at the Mikulski Archive for Space Telescopes (MAST) a few months after the end of the campaign.

In the rest of this section, we describe the two parallel light curve preparation procedures, or “pipelines,” from two different institutions (Oxford and Harvard-CfA in collaboration with NASA Ames), which we tested against each other in this study. The *K2* M67 study team includes core members of the teams behind the K2SC and VJ pipelines, which we implement here. While other *K2* pipelines have been successful in various contexts, we did not include them in this study for a variety of reasons. The EVEREST pipeline, for example, does not perform optimally with a crowded field (Luger et al. 2016, 2017). Furthermore, the *K2* light curves from K2VARCAT (Armstrong et al. 2014, 2016) only encompass Campaigns 0–4. Alternatively, Cody et al. (2018) have recently released Campaign 5 M67 light curves, using an approach to superstamp data similar to that of Aigrain et al. (2015a) (which we employ here). Nardiello et al. (2016) have also produced Campaign 5 M67 light curves using the PSF-fitting approach of Libralato et al. (2016) in addition to aperture and optimal-mask photometry. However, the light curves from Cody et al. (2018) were unavailable when we started this study, and we did not realize those of Nardiello et al. (2016) were publicly available until after we completed it. A future comparison with both sets of light curves would be interesting, especially since the root mean square error of the latter using aperture photometry at the bright end is similar to that of the pipelines we use (see Section 2.4).

Therefore, for each star, we applied our two pipelines to extract the light curves, correct for pointing-related systematics, and correct for residual common-mode systematics. In an ideal world, one would try every possible combination of each of these steps, resulting in eight versions of each light curve from the three separate data sets, each of which would have been injected with 42 simulated signals. Unfortunately, this would have led to an unmanageably large set of tests (almost 1,000,000 in total), and would have taken far too long to complete, especially when applying Pre-search Data Conditioning (PDC)–maximum a posteriori (MAP) for common-mode correction (see Section 2.3.3). While mixing the pipelines may have produced slightly more optimal results, we treat each pipeline as a unit, and produced only two versions of each light curve.

2.2. The “Oxford” Pipeline

The “Oxford” pipeline is comprised of the following.

1. For light curve extraction: the Simple Aperture Photometry (SAP) component of the standard *Kepler* pipeline, or the procedure described in Aigrain et al. (2015a) for cases where the SAP light curves are not available.
2. For pointing systematics correction: the K2SC pipeline Aigrain et al. (2016).
3. For the correction of residual common-mode systematics: a Principal Component Analysis (PCA) step.

2.2.1. Light Curve Extraction

Stars located outside the superstamp were processed on the ground by the SAP component of the *Kepler* pipeline. SAP light curves are extracted using a fixed, pixelated aperture, the boundaries of which are defined so as to maximize the signal-to-noise ratio of the resulting light curve (Van Cleve et al. 2016). They are corrected for known, pixel-level

instrumental effects only, and contain significant instrumental systematics, primarily due to the pointing variations of the telescope. These light curves are publicly available at MAST. The MAST light curve files also include a second version of each light curve, produced by the PDC component of the *Kepler* pipeline. The PDC step was designed to remove common-mode systematics and prepare the light curves for planetary transit searches. It was also designed to preserve all astrophysical signals, but it can reduce the amplitude of these intrinsic signals at timescales greater than 15 days (Gilliland et al. 2015). Additionally, in the case of *K2* data, it was never designed to remove the sawtooth systematics associated with the spacecraft roll error (Smith et al. 2012; Stumpe et al. 2012, 2014). Therefore, the use of PDC light curves can be problematic for variability studies with *K2*, and so we work with the SAP rather than the PDC versions.

As previously mentioned, MAST light curves are not available for the superstamp. We therefore use the method of Aigrain et al. (2015a) to extract the superstamp light curves ourselves. We give a brief summary of it here. We processed channels 6.1 and 6.2 in turn. First, a Full Frame Image (FFI) is downloaded and a preliminary astrometric solution obtained using `astrometry.net` (Lang et al. 2010). For each epoch, we then stitch together all the superstamp postage stamps from a given channel to create a reconstructed image, which is blank where no data was downloaded and is initialized with the preliminary astrometric solution from the FFI. We also construct a binary mask that indicates where data was collected. We then use the publicly available `CASUTOOLS` routines `nebuliser`, `imcore`, and `wcsfit` to model and subtract the background, identify stars within each image, and obtain a refined astrometric solution by cross-matching the catalog generated from each image with 2MASS. The variable background subtraction is new (compared to Aigrain et al. 2015a) and is important because of the high density of unresolved stars in the central regions of the cluster.

Next, we construct a master image by stacking the first 20 valid images using the astrometric solution for each frame. We repeat the source detection, photometry, and astrometric solution steps on this master image to obtain a master catalog. We then extract light curves for every star on the master catalog by placing a circular aperture at the sky position of each star on each image (as recorded in the master catalog). We compute the median of each star’s light curve and a zero-point correction for each frame by taking the weighted average of the median-subtracted fluxes, using the inverse square scatter to weight them, on that frame. This removes some systematics (particularly those associated with aperture losses), but significant residual systematics are nonetheless present in the extracted light curves.

The light curve extraction is done using several aperture radii (1, 2, $2\sqrt{2}$, 4, $4\sqrt{2}$, and 8 pixels) for every star in the superstamp. In general, larger apertures give the best results for bright stars, while fainter stars require smaller apertures to minimize the background noise. However, the best aperture for any given star also depends on the relative positions and brightness of neighboring stars. In practice, we select the aperture that minimizes the star’s calculated scatter (see Section 2.4). Eventually, however, this aperture needs to be checked for contamination from other stars, especially if it is larger than about two pixels.

We extracted light curves for 1342 stars located in the superstamp. We then cross-matched our master catalog to that of G15, finding a total of 976 matches. Most of the stars in our catalog without G15 matches are faint stars with poor quality K2 photometry, so we focus the remainder of our analysis of superstamp stars on those with G15 matches, hereafter referred to as the “SS” sample.

2.2.2. Star-by-star Systematics Correction

The dominant systematics in K2 light curves are due to the spacecraft’s pointing variations. We correct these using the K2SC pipeline. We briefly summarize K2SC here, but the more interested reader is referred to Aigrain et al. (2016). K2SC models each light curve as the sum of two unknown functions, one of which depends on the star’s pixel position and represents the pointing systematics, and another that depends on time and represents the star’s intrinsic variations, as well as any residual systematics not captured by the position-dependent component. The model is formulated as an additive Gaussian process, which allows us to require only that each function have a certain degree of smoothness, without having to specify its exact form. It also enables us to separate the time- and position-dependent components. Removing the former from the original light curve is helpful when looking for planetary transits, while subtracting the latter, which is dominated by the ~ 6 hr drift of the spacecraft, only leaves us with a light curve corrected for position-dependent systematics, with minimal impact on astrophysical variability. K2SC also flags significant outliers in the data, which we subsequently remove because the rotational signals in which we are interested are relatively smooth, for this study.

2.2.3. Common-mode Systematics Correction

Visual inspection of the K2SC-processed light curves reveals significant, residual, long-term trends common to many light curves, which appear to be present whatever the extraction and detrending method used. Barnes et al. (2016), who used the MAST PDC versions of the light curves, already noticed them; they used a PCA approach to remove them. These particular trends are likely due to stellar drift from differential velocity aberration (DVA), which causes photometric aperture losses or variations from sensitivity differences within a pixel.

We also use a PCA approach to remove residual systematics. We normalize each light curve by dividing by its median. We then construct a matrix where each row is the logarithm of the normalized light curve. We also experimented with the more standard approach of re-scaling each light curve to have zero mean and unit variance, but found that this gave poorer results. We then use singular value decomposition to compute: the eigenvectors of the matrix, which are the principal components (PCs); the eigenvalues, which tell us what fraction of the total variance of the data set is explained by each trend; and a matrix of coefficients relating each light curve to each PC. The first few PCs represent the dominant trends in the data, and these can be removed by subtracting each of them, multiplied by the corresponding coefficient from each light curve.

Because M67 falls on two different CCD channels, the two subsets might be expected to show different systematics. However, we found no significant difference between the sets of PCs extracted from the entire set or from each output

channel in turn. We therefore processed all the SAP stars on the two channels, both cluster members and non-members, together. On the other hand, there were much more significant differences between the PCs extracted from the SAP and SS sets, as illustrated in Figure 4, so these were processed separately. The SS PCs appear noisier than the SAP PCs, as a result of the relative dominance of the common-mode systematics in each set. The first, most dominant PC of the SAP set accounts for roughly 84% of the total variance of the light curves, while the first PC of the SS set only accounts for about 43%. The amplitude of the SS PCs are therefore much smaller, relative to the noise, than the SAP PCs.

If PCA is applied blindly, large-amplitude features in individual light curves can dominate the PCs, skewing the correction and introducing these features in the corrected light curves of other stars. This can be diagnosed easily, however, as the PCs are linear combinations of the light curves, and the linear combination becomes dominated by a single star. Two problematic stars in the SAP set were identified in this way, and excluded from the PC estimation: EPIC 211391083 (an eclipsing binary) and EPIC 211327533 (whose light curve contains a large discontinuity). We also excluded the first 85 cadences (~ 1.8 days) from the light curves before evaluating the PCs, as the systematics were particularly pronounced in that early phase of the campaign, and the PCA correction was of significantly lower quality when they were included (perhaps because the assumption of linearity, which is intrinsic in PCA, was violated more radically in that period).

Usually, a threshold in the fraction of explained variance (i.e., the eigenvalue) associated with each PC is used to decide how many PCs to include in the correction. We experimented with different values for this threshold, and settled on thresholds of 2.5% and 7.5% for the SAP and SS sets, respectively, which correspond to two PCs in both cases; see Figure 4. We tested a range of thresholds for each data set and found that overly high thresholds only produced one PC, which was not sufficient to explain all of the dominant, common modes seen in the light curves when examined by eye. Going too low, on the other hand, produced three or more PCs, in which case the extras were superfluous or introduced unwanted features in the data. A range of thresholds can produce the same two PCs, but the ones we chose were the on the lower end of those ranges for both data sets.

2.3. The “CfA” Pipeline

In analogy with the Oxford pipeline, the “CfA” pipeline is comprised of the following three segments:

1. Light curve extraction from calibrated K2 TPFs, as described by Vanderburg & Johnson (2014) and Vanderburg et al. (2016).
2. Correction for the 6 hr roll systematics, again as described by Vanderburg & Johnson (2014) and Vanderburg et al. (2016).
3. Correction of non-roll systematics using the *Kepler* team’s PDC technique (Smith et al. 2012; Stumpe et al. 2014).

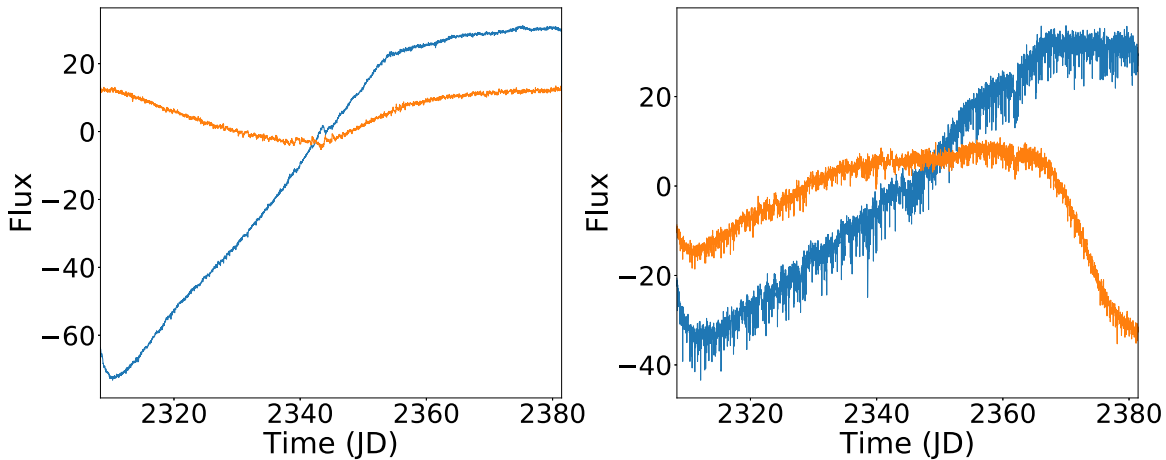


Figure 4. Principal components used to correct long-term systematics in the SAP (left) and superstamp (right) data sets, as part of the Oxford pipeline. The primary PCs are in blue, while the secondary PCs are in orange.

2.3.1. Light Curve Extraction

Light curve extraction for the CfA pipeline proceeds following the prescription of Vanderburg et al. (2016), with a few modifications for working on cluster light curves and in crowded regions. Unlike the Oxford pipeline, the CfA pipeline extracts light curves from the calibrated *K2* TPFs for both stars in the superstamp and in individual target postage stamps. We begin by using the World Coordinate System astrometric solution produced by the *K2* pipeline to identify the star of interest in either the target postage stamp or the superstamp sub-aperture.¹¹ We then laid down 20 different stationary, pixelated photometric apertures over the star in question: 10 of these apertures are circular, with radii logarithmically spaced between 1.5 and 13 pixels; the other 10 are shaped like the *Kepler* pixel response function (PRF) with different sizes. The PRF-shaped apertures tend to be smaller than the circular apertures. After defining the 20 different apertures, we summed the calibrated, background-subtracted pixels within those apertures at each time stamp to create raw light curves.

2.3.2. Star-by-star Systematics Correction

After producing the raw light curves, we removed the dominant systematic errors on 6 hr timescales introduced by *K2*'s unstable pointing, using the method described by Vanderburg & Johnson (2014) and Vanderburg et al. (2016). This correction works by decorrelating variability due to the roll of the spacecraft (which is correlated with the spacecraft's roll angle) from astrophysical variability (which is not correlated with the roll angle). Over the course of a *K2* campaign, the motion of the spacecraft causes stars to trace out a path back and forth on the detector, which slowly wanders as DVA causes the apparent positions of stars to move over the course of a campaign. We break up the campaign into a number of shorter time segments, when the motion of the spacecraft can be approximated as purely one-dimensional, and perform our systematics corrections in those divisions individually. We decorrelate the systematics from the astrophysical variability by

iteratively dividing away low-frequency variations, determining the relationship between the measured flux and *Kepler*'s roll angle, fitting a piecewise linear function to that relationship (with outlier exclusion to prevent transits, flares, or other short-timescale variability from being mistaken as a spacecraft systematic and removed). Once we have a piecewise linear function that models the roll-angle dependence in the raw light curve, we divide it away, remove any leftover low-frequency variations from the resulting light curve, and repeat the process. After a few iterations, the process converges, and we divide out the best-fit roll-angle variability from the raw light curve to yield a systematics-corrected light curve, with stellar variability and transits preserved.

After removing systematics from each of the 20 raw light curves produced from the various photometric apertures, we selected a "best" aperture to use in our analysis by determining which aperture produced the light curve with the best photometric precision. In the crowded M67 region, we only allowed the PRF-shaped apertures and circular apertures smaller than three pixels in radius to be chosen, in order to prevent additional stars from falling within the photometric aperture.

2.3.3. Common-mode Systematics Correction

The CfA star-by-star systematic correction above performs very well at removing roll-induced errors in each light curve. However, just as with the Oxford light curves, there still exist significant common-mode systematics as a result of focus changes, along with other long-term drifts due to DVA and other stochastic errors (such as attitude tweaks, safe modes, etc.). PDC-MAP, on the other hand, was designed to address the focus-induced systematics that were dominant in the original *Kepler* mission. Therefore, an ideal *K2* pipeline can be created by applying the CfA correction first, followed by the PDC correction. PDC uses a method somewhat similar to PCA, but it follows a Bayesian MAP approach, where a subset of highly correlated and quiet stars is used to generate a co-trending basis vector set, which is in turn used to establish a range of robust, "reasonable" fit parameters. These parameters are then used to generate Bayesian prior and posterior probability distribution functions (PDFs) that, when maximized, find the best fit that simultaneously removes systematic effects while reducing the signal distortion and noise injection

¹¹ Unlike the Oxford pipeline, we did not splice together the superstamp sub-apertures for our analysis, and instead worked with the 50×50 pixel sub-apertures created by the *Kepler* pipeline. When stars fell near the edge of one of the sub-apertures, we spliced together neighboring sub-apertures to produce the light curves.

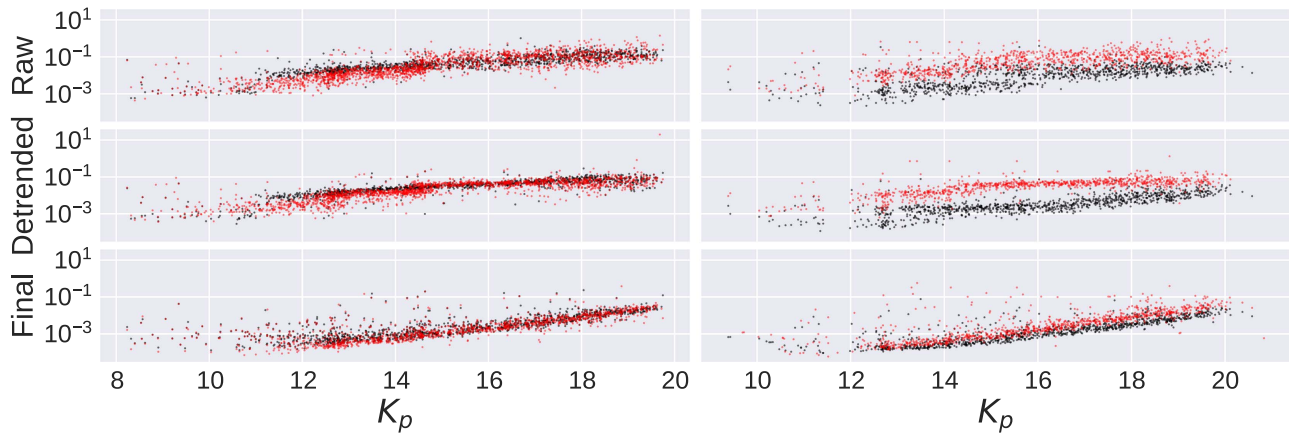


Figure 5. Scatter vs. K_p magnitude for the SAP (left column) and SS (right column) samples. The results of the Oxford and CfA pipelines are shown in black and red, respectively. From top to bottom, we plot the scatter in the raw light curves, after star-by-star systematics correction, and after common-mode systematics correction.

that commonly afflict simple least-squares fitting. A numerical and empirical approach is taken where the Bayesian prior PDFs are generated from fits to the light-curve distributions themselves.

PDC has two modes of operation: “Single-Scale” (ssMAP) and “Multi-Scale” (msMAP). The former, ssMAP (Smith et al. 2012), performs the MAP correction in a single band-pass. This has the advantage of instilling a strong regularization on the basis vector fit coefficients, thereby minimizing the removal of stellar signals. However, this is at the expense of more bias and less systematic error removal. msMAP (Stumpe et al. 2014), on the other hand, utilizes an over-complete, discrete wavelet transform, dividing each light curve into multiple channels, or bands. The light curves in each band are then corrected separately, allowing for a better separation of characteristic signals and improved removal of the systematic errors, but at the expense of more stellar signal removal, especially at longer periods. Both methods have their advantages, but generally speaking, msMAP performs better at preserving and cleaning light curves at transit timescales, and ssMAP performs better at preserving long-period signals. It is therefore expected that ssMAP will perform better for the studies carried out in this paper.

2.4. Scatter Comparison

We here compute the scatter of the M67 SAP and SS stars as a function of *Kepler* magnitude at each step of the Oxford and CfA pipelines, in order to conduct a preliminary assessment of pipeline performance prior to the injection tests. We use a variation of the median of absolute deviations (MAD) (Hoaglin et al. 1983) scatter, estimated in the following manner:

$$\text{MAD} = 1.48 \text{median}(|f_i - m_f|), \quad (1)$$

where f_i is the flux at each cadence i , and m_f is the median flux. We use the median as opposed to the mean because it is less influenced by outliers. Technically, the median term itself is the MAD scatter; we include a factor of 1.48 as a scaling factor that, under the assumption of a Gaussian distribution of f , makes the scatter equivalent to the standard deviation.

The scatter is given in Figure 5. The Oxford pipeline results are shown in black and the CfA, using single-scale PDC-MAP, in red. From top to bottom, the figure shows the scatter for the raw light curves, after star-by-star systematics correction, and

after common-mode systematics removal, with the SAP sample on the left and the SS sample on the right. In the SAP sample, the raw scatters produced by the two pipelines are fairly similar, though the CfA pipeline performs slightly better. For both pipelines, the star-by-star systematic correction reduces the scatter somewhat over the whole magnitude range and forms tighter relationships, but the final common-mode systematic correction has the most drastic effect, leading to scatter values of a few tens of ppm at the bright end. The final scatter obtained by the CfA pipeline is slightly lower than that obtained by the Oxford pipeline over the entire magnitude range. The final median scatters at $K_p = 12$ for the Oxford and CfA SAP samples are 544 ppm and 321 ppm, respectively.

By contrast, the raw outputs of the two pipelines are very different for the SS sample. The extraction step used by the Oxford pipeline, which involves moving apertures, already lessens the pointing-related systematics in the raw light curves of fainter stars. The effect of the star-by-star systematic correction on the scatter is then relatively minor for both pipelines, but the common-mode systematic correction is particularly effective over the whole magnitude range for the CfA pipeline, and for bright stars in the Oxford pipeline. The final results of the two pipelines are fairly similar, though the CfA pipeline performs slightly better for brighter stars (albeit not to the extent of the SAP sample). Among the fainter stars, there are several instances where the CfA pipeline leads to final scatters on the order of 1%, whereas the corresponding scatter for the Oxford pipeline is much lower. In these cases, the Oxford pipeline generally tends to correct the systematics better. The final median SS scatter at $K_p = 12$ is 203 ppm for the Oxford pipeline and 275 ppm for the CfA pipeline.

The final scatter distributions are very similar for the SAP and SS samples, as well as for the two pipelines. It is worth noting, however, that the scatter between the CfA SAP and SS data sets are closer to each other than their counterparts in the Oxford pipeline. This is most likely due to the fact that the CfA light curves are all extracted using the same technique, while the two Oxford data sets are extracted via different methods, so they have different starting points with respect to noise levels. Regardless, these numbers should give us increased confidence that the light curves we are using, from both pipelines, are relatively robust, and there is no obvious effect in either sample that is not handled reasonably well by either pipeline.

3. Injection Tests

We seek to quantify the limits of our ability to measure stellar rotation periods in K2 M67 light curves. To do this, we need to simulate realistic light curves that share the same time sampling and noise properties as those we wish to analyze, but that also contain signals of known period and amplitude (as well as any other parameter that might affect the detectability of the signal). In the absence of a detailed generative model for the various noise sources and systematics in K2 data, we are not able to simulate realistic light curves from scratch. Instead, we inject known rotation signals into the raw versions of the observed SAP and SS light curves (i.e., just after extraction). This introduces a few problems—the most obvious being that some light curves may already contain strong astrophysical variability—to which we later return. We then separately apply the detrending steps (pointing-related and common-mode systematics correction) of the Oxford and CfA pipelines, before attempting to recover the periodic signals. This enables us to assess the extent to which the noise present in the data, as well as the pre-processing applied to reduce this noise, affects the detectability of the periods.

As described in Section 1, this paper deals only with sinusoidal injected signals, but the tests are designed to be comprehensive, in that every signal is injected into every light curve. A smaller set of tests, involving more realistic signals and comparing different period search methods, will form the subject of a later paper.

3.1. The Stellar Samples

For the purposes of the injection tests, we define the following three stellar samples.

1. The SAP sample consists of the 1877 stars observed with individual postage stamps in channels 6.1 and 6.2, 1319 of which match to our master catalog and 236 of which are confirmed members from G15.
2. The SS sample consists of the 976 stars included in the superstamp with matches in the master catalog, 359 of which are confirmed members from G15. Seventy-five of the SS overlap with the SAP.
3. In addition, we define a hot star sample consisting of 89 main sequence stars with spectral types A through F located on CCD modules 6, 11, and 16, with effective temperatures ranging from 6300 to 10715 K (the majority falling between 6300 and 7300 K), $B - V$ values of -1.0 to 0.45 , and $\log g$ values greater than 4.0 . This information was all acquired from MAST. These stars were chosen because they lack an outer convection zone, or have only a very thin one. This means they are not able to support a large-scale magnetic field, and so do not spin down quickly (Schatzman 1962). Thus, if they do display periodic variability, it is expected to occur on timescales of ~ 2 days or less, whether it is due to pulsations (many of these stars lie in the classical instability strip) or rotation (as these hot stars are rapid rotators) (Nielsen et al. 2013). Therefore, their light curves are expected to contain only noise and systematics, at least on timescales longer than ~ 2 – 7 days, which makes them ideal targets for injection tests. Twenty-two of the hot stars overlap with the SAP set.

3.2. The Injected Signals

We inject sinusoidal signals by directly multiplying the following into the raw light curves of each test sample:

$$f_{\text{inj}} = \frac{a_{\text{inj}}}{2} \sin\left(\frac{2\pi}{P_{\text{inj}}}t + \phi_{\text{inj}}\right) + 1, \quad (2)$$

where f_{inj} is the injected signal, a_{inj} is the injected amplitude, P_{inj} is the injected period, ϕ_{inj} is the injected phase, and t is time in days. We vary P_{inj} from five to 35 days, in intervals of five days, leading to seven period values. We also vary a_{inj} , as follows: 0.05%, 0.10%, 0.30%, 0.50%, 1.00%, and 3.00%. Each light curve, then, is injected with 42 different combinations of periods and amplitudes, leading to over 120,000 injections in total for each pipeline to process. The phase, ϕ_{inj} , is selected at random for each injection, but we keep track of its value. Where we have extracted the SS light curves using multiple aperture sizes from the Oxford pipeline, we inject the signals into the version that minimizes the scatter for the corresponding, non-injected light curve.

3.3. The Detection Algorithm

Because these injection tests involve only sinusoidal signals, we use a modified version of the widely used Lomb–Scargle (LS) periodogram Scargle (1982) to recover them. The LS periodogram is essentially equivalent to least-squares fitting of sinusoidal signals (Irwin et al. 2006), and is thus, in principle, optimal to recover stable sinusoidal signals in the presence of white Gaussian noise of known variance.

Real stellar rotation signals are not strictly sinusoidal; they contain significant power at harmonics of the true period due to the scattered distribution of the active regions on the stellar surface, evolve over time with the active region evolution, and may contain signals at a range of periods due to differential rotation. For this reason, rotation period searches in *Kepler* and K2 data often use other period search methods that do not depend on the assumption of sinusoidality or strict periodicity (McQuillan et al. 2013a; Angus et al. 2018). These methods have been shown to outperform approaches based on least-squares sine-fitting in real and simulated data sets (Aigrain et al. 2015b). However, if we know the signal of interest is sinusoidal, then a sine-fitting approach is likely to do at least as well as these alternative, more flexible methods.

The model we are considering is of the form:

$$m(t) = m_{\text{dc}} + \alpha \sin(\omega t + \phi), \quad (3)$$

where m_{dc} is the light curve’s vertical offset, α is the amplitude, ϕ is the phase, and ω is the angular frequency. For each value of ω (or period, $P = 2\pi/\omega$), this model can be expressed as a linear basis model with three basis functions: a constant, a sine term, and a cosine term, each with period P . For each trial period, we can solve for the values of m_{dc} , α , and ϕ that minimize the sum of the squared residuals, or χ^2 (in space data, the relative measurement uncertainties can be treated as approximately constant for a given star, so the two are equivalent). We evaluate the relative reduction in χ^2 with respect to a constant model, $S = (\chi^2(0) - \chi^2(P))/\chi^2(0)$, as a function of period P , to construct a periodogram. For this study, we search for 490 periods ranging from 2 to 100 days, using an evenly spaced grid in frequency space.

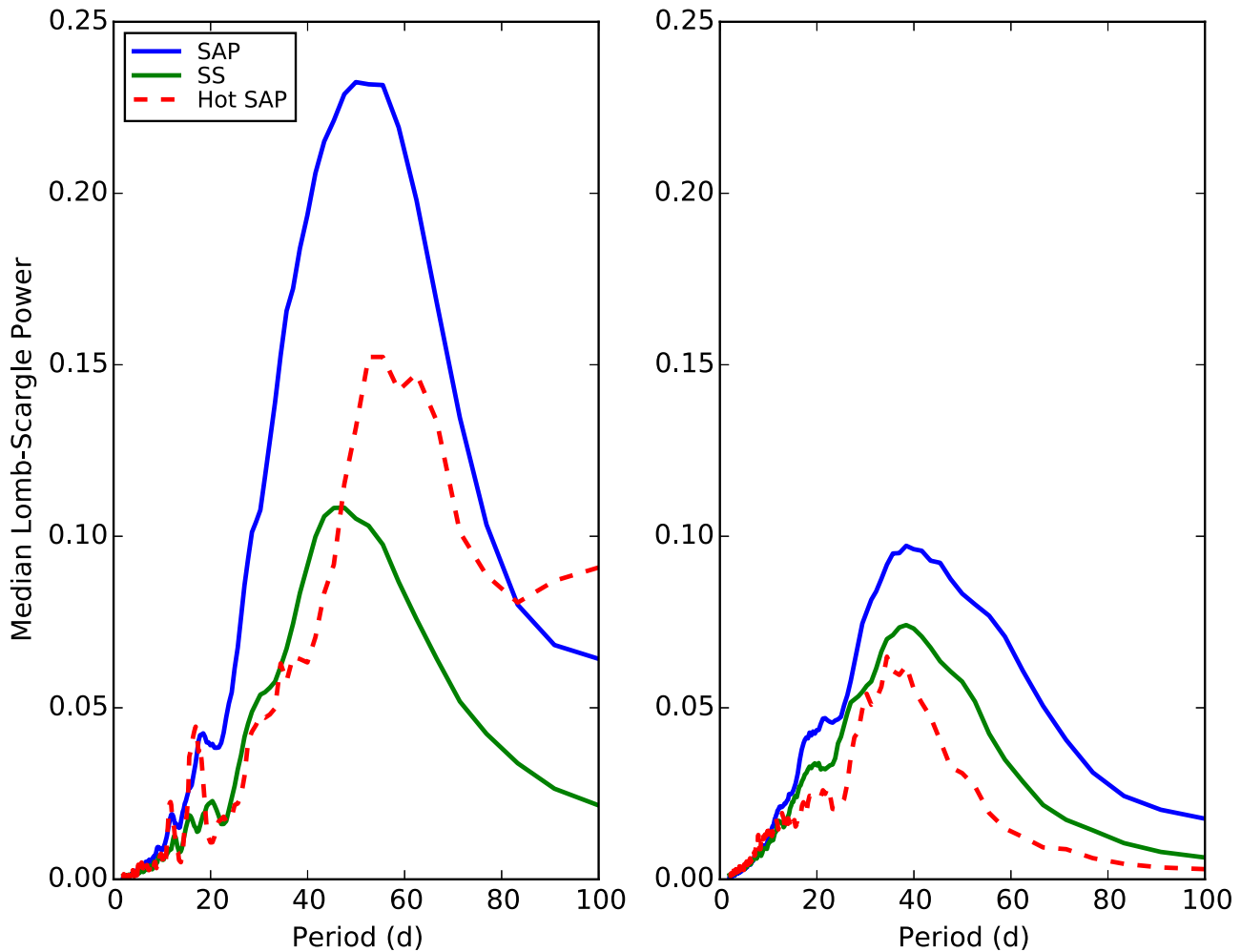


Figure 6. Median Lomb–Scargle periodogram for the non-injected SS (blue) and SAP (green) data sets for both pipelines. The non-injected hot star data set for each pipeline separated from the SAP data set is shown with red dashed lines. The Oxford pipeline is on the left, and the CfA pipeline is on the right. The presence of broad peaks around 30 days and longer in the hot star sample highlights the presence of long-term, non-astrophysical signals in the Campaign 5 light curves.

3.3.1. Periodogram Normalization

If the light curves consisted simply of a stable sinusoidal signal plus white Gaussian noise, locating the peak from the Lomb–Scargle periodogram would give us the best-fit period, and evaluating the significance of the detection would be relatively straightforward (see, e.g., Horne & Baliunas 1986). However, this is not the case, even after our light curves are subjected to the systematics-correction steps of our two pipelines. The residual systematics, which remain in the light curves despite our best efforts, often lead to broad peaks in the range of 30–60 days, which can overwhelm the peak due to the injected signal. To address this, we perform a collective normalization of the periodograms before searching for the period.

To normalize the periodograms, we compute the median periodogram as a function of trial period for stars we expect to share the same systematic noise properties (i.e., separately for the SAP and SS sets, and for the Oxford and CfA pipelines). Each star’s periodogram is then divided by the median periodogram, thereby suppressing the peaks, or excess power, seen in many light curves. The normalized periodogram is thus given by $S'(P) = S(P)/\hat{S}(P)$, where $\hat{S}(P)$ is the median periodogram value for period P . The median periodograms are evaluated from the non-injected versions of the light curves, to

avoid the induced power at fixed periods from their injected counterparts; they can be seen in Figure 6, with the Oxford SS and SAP median periodograms on the left and the CfA on the right. The SAP median periodograms for both data sets include the hot stars, as they were processed in both cases using the same techniques. However, Figure 6 also shows the hot star median periodograms separate from the rest of the SAP set for both pipelines, given by red dashed lines. This emphasizes that the broad peaks in both pipelines at long periods are most likely not a result of true astrophysical signals, as we would not expect periods much longer than about five days in the hot star samples. We also note that the median periodogram for the SAP set from the Oxford pipeline is higher overall and peaks at longer periods than the SS sample, hinting that the former likely contains systematics with larger amplitudes on greater timescales. This is consistent with the relative amplitudes of the PCA trends compared to noise for the two samples, which are smaller for the SS (see Figure 4). This also means that long-period signals will be more heavily suppressed by normalization in the SAP than in the SS sample.

Figure 7 illustrates the effect of the periodogram normalization for an individual star from the SAP set. The top panel shows the light curve of EPIC 211355490 injected with a period of 20 days and amplitude of 0.1%, fully processed with

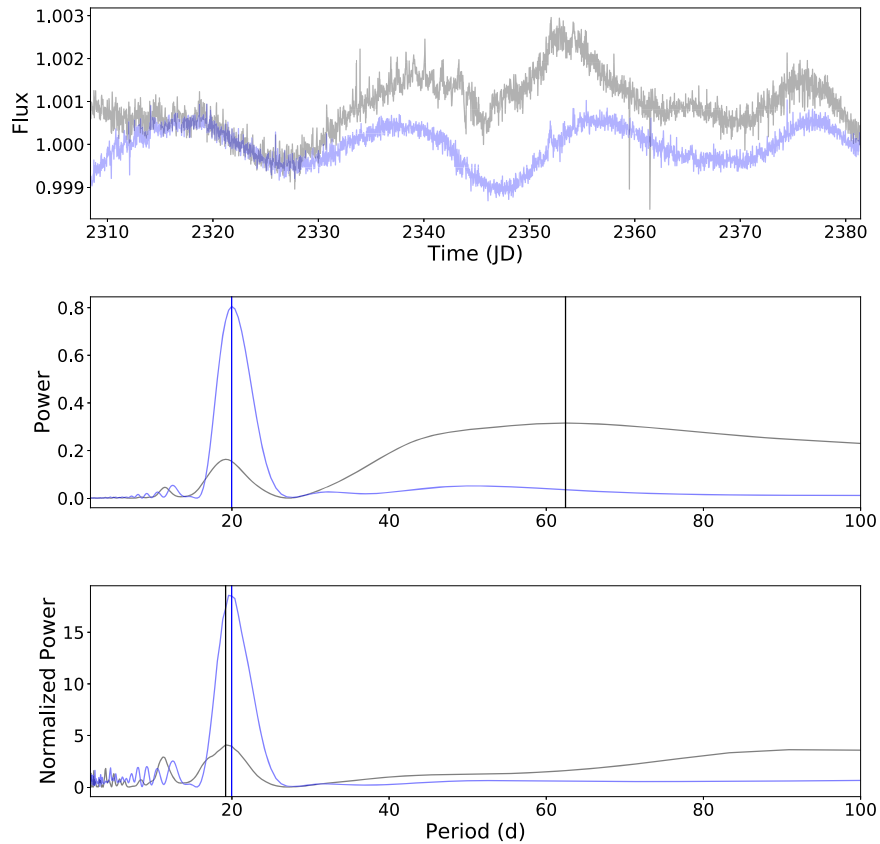


Figure 7. Example of the effect of periodogram normalization using EPIC 211355490 injected with a period of 20 days and amplitude of 0.1%. The top panel shows the systematics-corrected flux from the Oxford (gray) and CfA (blue) pipelines. The middle panel gives the original periodograms for both pipelines, while the bottom panel shows the normalized periodograms. The periods found in each periodogram are indicated by the vertical lines, with the black lines depicting the Oxford periods and the blue lines identifying the CfA periods. The CfA period in both cases is 20.0 days, while the Oxford period improves from about 62 to 19.2 days.

both the Oxford (in gray) and CfA (in blue) pipelines. The middle panel depicts the original periodograms prior to normalization for both pipelines, while the bottom panel gives the normalized versions. The respective periods are indicated by vertical lines. There is a relatively strong, long-period signal in the Oxford pipeline that is not present in the CfA (unsurprising, given the differences in Figure 6), and this is what the Lomb–Scargle finds in the former, settling on a period of 62 days. On the other hand, the normalized Oxford periodogram has a peak at 19.2 days, which closely matches the injected period. The normalization of the periodogram thus allowed us to find the “true” period of Oxford version of this light curve when otherwise it would have been lost by the systematic, long-term trend. The CfA, however, appears to have done a better job of cleaning up systematics in the light curve; it found peaks of 20.0 days in both the original and normalized periodograms.

3.3.2. Detection Threshold

Once we have identified the peak in each normalized periodogram, we must decide whether it is significant. After testing a number of different possible schemes, we opted for the following detection threshold definition:

$$S'_T = \text{MAX}(S'_{90} \times C, S'_{\min}), \quad (4)$$

where S'_T is the threshold value and S'_{90} is the 90th percentile of the normalized periodogram for a given light curve. Here, C and S_{\min} are tuning parameters that we can vary to alter the

relative and absolute components of the threshold, respectively. For example, if $S'_{90} = 4$, and the value for C is set at 3, then the maximum power of the normalized periodogram must be at 12 or greater for the associated period measurement to be considered a detection. S_{\min} comes into play when $S'_{90} \times C$ is so low that most peaks would qualify as a detection; it ensures that there is a minimum value for the threshold. Both C and S_{\min} can then be varied to test their effect on the completeness and reliability of the period search, which we discuss in Section 3.4.4.

If the highest peak in the normalized periodogram passes the threshold defined in Equation (4), then we have a detection and record the corresponding period, as well as the best-fit amplitude and phase. Otherwise, a result of “no detection” is recorded.

3.4. Evaluating the Results

To evaluate the results of the injection tests, we need to quantify how sensitive the period search is, i.e., what fraction of the injected signals are successfully recovered—but also how trustworthy the detections are, i.e., what fraction of the detections are “valid,” meaning that the measured period and phase are within some tolerance of the injected values.

3.4.1. Valid Detections, Completeness, and Reliability

Here, we define several key terms in understanding the statistics from this study:

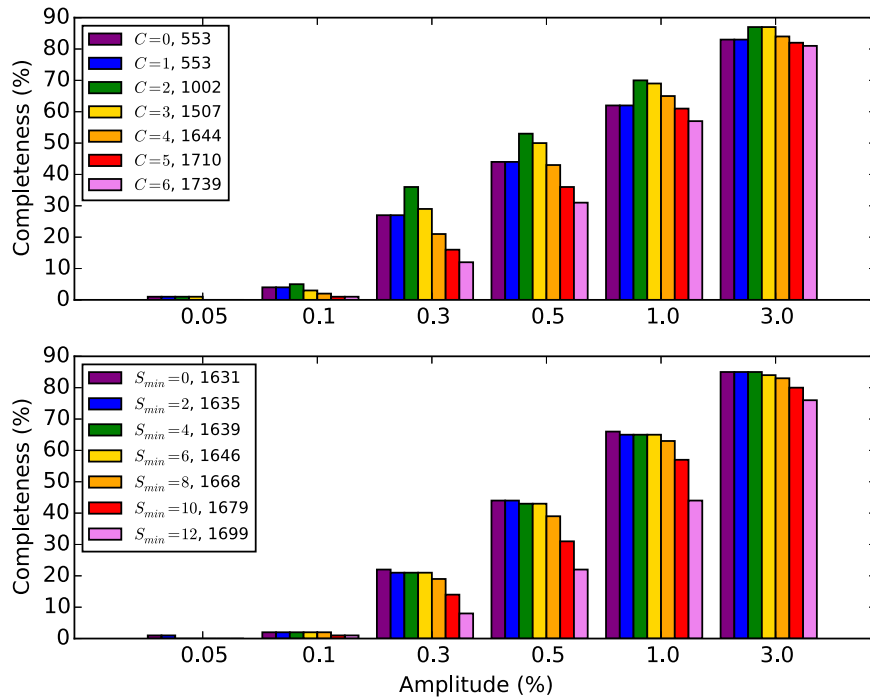


Figure 8. Comparison of completeness values for an injected period of 25 days at all injected amplitudes for different values of C (top row) and S_{\min} (bottom row). For the top row, S_{\min} was fixed at five. For the bottom row, C was fixed at four. The legends are also marked with the number of stars in each injected period and amplitude bin for each value of C and S_{\min} . It is worth noting that, by $S_{\min} = 12$, all sensitivity at periods greater than 25 days was gone. We used the Oxford pipeline for these tests.

1. *Validity.* We consider a detection *valid* if the recorded period and phase are within 20% of the injected values.
2. *Completeness.* For a given range of *injected* period and amplitude, we define the *completeness* as the fraction of injected signals that led to a valid detection. We also calculate a *threshold error* statistic, which records the fraction of cases where the period and phase corresponding to the highest peak in the normalized periodogram were within the valid range, but the peak was not high enough to pass the detection threshold.
3. *Reliability.* For a given range of *recorded*, or measured, period and amplitude, we define the *reliability* as the fraction of recorded detections that were valid.

Another way to understand the definitions of completeness and reliability given above is to look ahead at Figure 10, which shows (for the hot star sample) the detected versus injected periods for different injected amplitudes. To evaluate completeness, we consider a vertical bin in one of these diagrams around one of the injected periods. The completeness is given by the number of valid detections (colored points) in that bin, divided by the total points in the bin. To be valid, a detection must lie within the gray-shaded area, which indicates a $\leq 20\%$ period error. In addition, a valid detection must also have a $\leq 20\%$ phase error (not shown on the figure). We can think of reliability as being similar, but considering a horizontal rather than vertical bin in the same kind of diagram. (This is not quite correct, because reliability is computed for a given detected period and amplitude, whereas the figure shows the injections split by injected amplitude.)

3.4.2. Calculating Uncertainties

To estimate the uncertainties, we assume a Poisson distribution with respect to both completeness and reliability,

and define the uncertainty as:

$$\sigma = \pm \frac{1}{\sqrt{N}}, \quad (5)$$

where N is the number of injections per injected period and amplitude bin (i.e., the number of injected amplitudes multiplied by the number of injected periods multiplied by the number of stars in each test sample) in the case of completeness. For reliability, N is the total number of detections in each detected period and measured amplitude range. Therefore, for each test set within each pipeline, the individual completeness values will have the same uncertainty while the reliability will vary from bin to bin. We recognize this is a little simplistic, but the more sophisticated approach would require repeating the injections many times, which would be far too time-consuming. For our purposes, this is a good first approximation.

3.4.3. Intrinsically Variable Stars

We injected signals into real *K2* light curves in order to faithfully reproduce the noise properties of actual M67 light curves. In doing this, we are implicitly assuming that most of them do not already contain a detectable intrinsic periodic signal. However, some of them do, of course. In such cases, a detection might be caused by the intrinsic rather than the injected signal. This can lead to detections that appear “invalid,” but are in fact correct, biasing the completeness and reliability statistics.

To avoid this, we run the period search on the non-injected versions of the light curves. If a detection occurs in any of these, the star is labeled as intrinsically variable and is excluded from the completeness and reliability calculations. This means

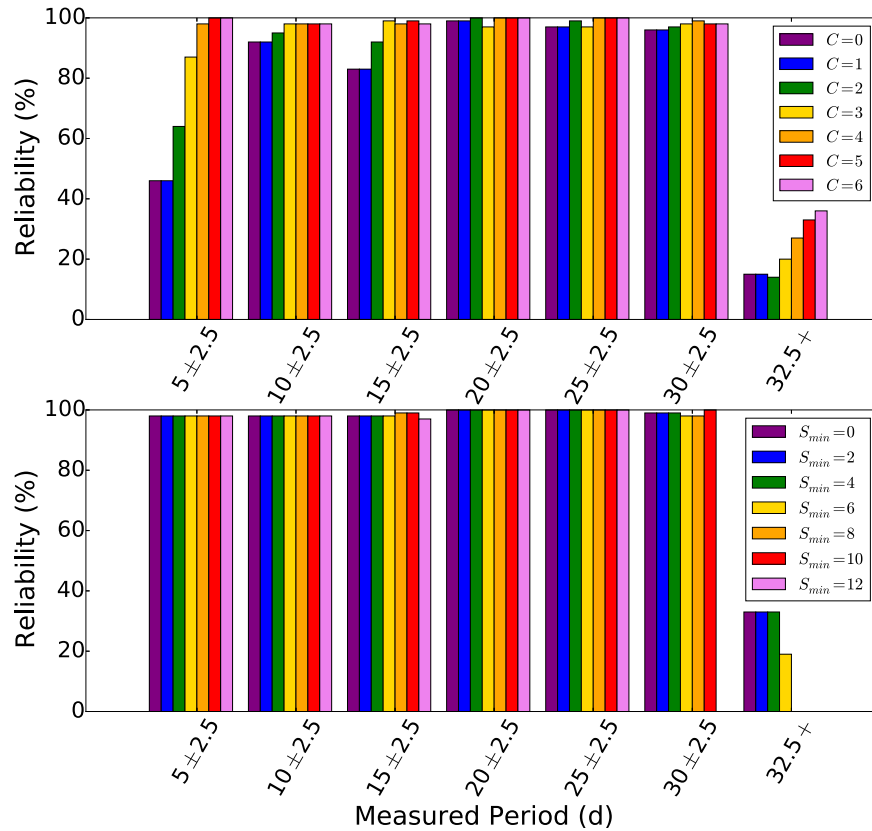


Figure 9. Comparison of reliability values for a measured amplitude between 0.25% and 0.45% across all detected periods for different values of C (top row) and S_{\min} (bottom row). For the top row, S_{\min} was fixed at 5. For the bottom row, we fixed C at 4. It is worth noting that we lost sensitivity for detections greater than 32.5 days at $S_{\min} = 8$, and we lost sensitivity for detections greater than 27.5 days at $S_{\min} = 12$. We used the Oxford pipeline for these tests.

that only a fraction of the total number of injections we performed are actually used in the final results.

This is not entirely satisfactory, as in reality we do not know whether the signal detected in the light curves without injections were indeed due to the intrinsic variability of the star or to systematics. Furthermore, the number of stars excluded as “variables” depends on the detection threshold used. However, we were unable to devise a more satisfactory solution. We note that, for reasonable threshold values, the variable stars represent a relatively small minority of the test light curves.

3.4.4. Varying the Threshold

The detection threshold obviously has an effect on the results of this study. Lower thresholds increase completeness, but also lead to a drop in reliability in the same bins. If measured rotation periods are to be compared with theoretical models, it is preferable to prioritize higher reliability over slight gains in completeness—so long as the completeness itself is well-measured, to account for missed detections. We therefore recorded the normalized periodogram peak S' ; the corresponding period, phase, and amplitude; and the 90th percentile of the normalized periodogram, S'_{90} , for each injected signal in each light curve. It was then trivial to vary the value of C and S_{\min} in (4) and examine the impact that this had on completeness, reliability, and the “variables” excluded from the statistics.

We experimented with $C = 0$ to $C = 6$ while holding S_{\min} at 5, and varying S_{\min} from 0 to 12 while holding C at 4, in order to test the effect that each parameter has on completeness and

reliability. We used the Oxford pipeline to perform these two tests. A sample of the results of these tests can be seen in Figures 8 and 9. Figure 8 shows the effect on completeness for changing both tuning parameters at an injected period of 25 days across all injected amplitudes. The top row of each figure shows the effect of changing C , while the bottom row shows the effect of changing S_{\min} . In both panels, we print the number of stars in each injected period and injected amplitude bin. We can see that, at $C = 0$ and $C = 1$, the results are dominated by the minimum threshold value, but they have relatively few stars per bin. Completeness then peaks at $C = 2$ before starting to decline again. There is also an increase of ~ 500 stars per bin at $C = 2$, followed by another 500 stars at $C = 3$. Beyond $C = 3$, the number of stars per bin still grows, but less rapidly. For S_{\min} , the completeness essentially stays the same until about $S_{\min} = 8$. By $S_{\min} = 12$, the completeness drops to about half the values at $S_{\min} = 0$ at amplitudes of 0.5% and less. Not visible in Figure 8 is the fact that all sensitivity at periods greater than 25 days, at any amplitude, is lost. The number of stars per bin increases with S_{\min} , but only very slightly. Thus, as expected, C has a much greater effect on the number of stars per bin—and more importantly, on completeness—than S_{\min} , as long as S_{\min} is not too high (in the case of completeness).

Likewise, in Figure 9, we can see an illustration of how C and S_{\min} affect reliability across the ranges of measured periods in this study. Here, we have fixed the measured amplitude to the range of 0.25–0.45%, just above the solar range. Again, the top row shows the effect of changing C . As expected, reliability generally increases as C increases, noticeably so until about

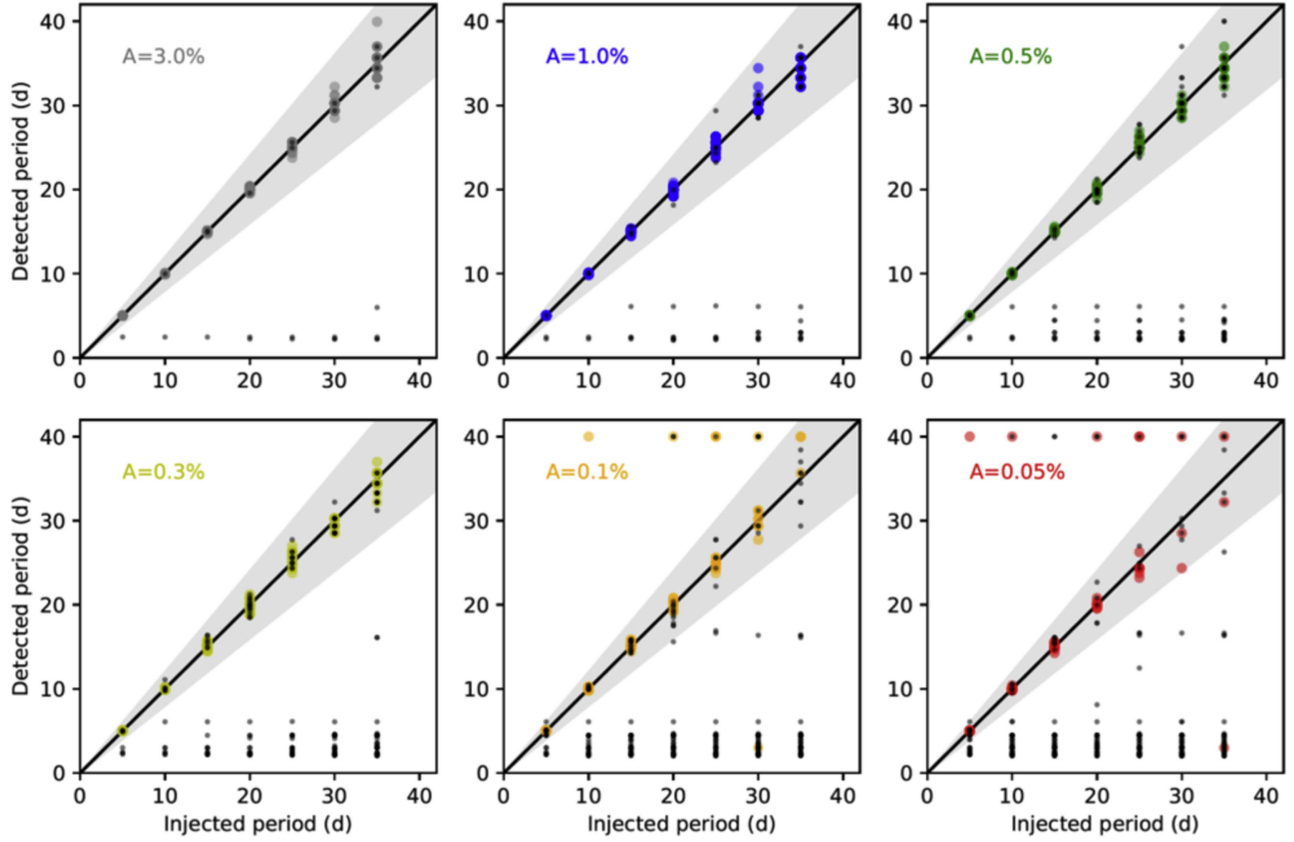


Figure 10. Detected vs. injected period at each of the injected amplitudes for the hot star sample processed with the Oxford pipeline. Colored points represent cases that passed our detection threshold; those that did not are shown as smaller black points. The shaded gray area shows the region where the injected and detected periods are within 20% of each other.

Table 1
Oxford Hot Star Completeness (%)

$a_{\text{inj}} \setminus P_{\text{inj}}$	5 days	10 days	15 days	20 days	25 days	30 days	35 days
3.00%	98	95	95	94	94	92	88
1.00%	95	94	88	88	83	73	59
0.50%	92	89	80	70	65	52	35
0.30%	86	80	71	50	48	26	23
0.10%	65	47	27	17	17	11	3
0.05%	41	24	17	12	9	3	2

Note. Total number of injections per period–amplitude bin: 66; $\sigma = \pm 12.3\%$.

Table 2
CfA Hot Star Completeness (%)

$a_{\text{inj}} \setminus P_{\text{inj}}$	5 days	10 days	15 days	20 days	25 days	30 days	35 days
3.00%	100	99	94	91	87	83	80
1.00%	97	93	91	84	81	77	76
0.50%	93	90	86	83	77	73	60
0.30%	91	87	84	80	69	49	34
0.10%	83	66	56	47	44	34	20
0.05%	66	50	36	16	16	13	6

Note. Total number of injections per period–amplitude bin: 70; $\sigma = \pm 12.0\%$.

$C = 4$. The effect of changing S_{min} , however, is largely negligible except at the long period ranges. There, it decreases until $S_{\text{min}} = 8$. Here, the reliability is zero, but with very few detections, and beyond, the number of detections drops to zero for periods longer than about ~ 30 days.

Figures 8 and 9 demonstrate the clear trade-off between completeness and reliability. Using large values for C and S_{min} , i.e., a stringent detection threshold, leads to low completeness and large threshold error, but excellent reliability. A high detection threshold also means that fewer stars are marked as

Table 3
Oxford SAP Completeness (%)

$a_{\text{inj}} \backslash P_{\text{inj}}$	5 days	10 days	15 days	20 days	25 days	30 days	35 days
3.00%	99	97	92	86	85	79	65
1.00%	93	81	75	65	65	56	28
0.50%	82	71	58	43	43	30	8
0.30%	73	59	40	24	21	12	3
0.10%	44	21	5	3	2	1	1
0.05%	21	4	1	1	0	0	0

Note. Total number of injections per period–amplitude bin: 1639; $\sigma = \pm 2.5\%$.

Table 4
CfA SAP Completeness (%)

$a_{\text{inj}} \backslash P_{\text{inj}}$	5 days	10 days	15 days	20 days	25 days	30 days	35 days
3.00%	97	92	89	82	74	72	68
1.00%	91	81	73	66	61	53	39
0.50%	83	69	63	53	47	33	23
0.30%	74	58	52	41	37	30	18
0.10%	54	33	23	18	14	11	7
0.05%	35	19	9	5	3	2	2

Note. Total number of injections per period–amplitude bin: 1587; $\sigma = \pm 2.5\%$.

Table 5
Oxford Superstamp Completeness (%)

$a_{\text{inj}} \backslash P_{\text{inj}}$	5 days	10 days	15 days	20 days	25 days	30 days	35 days
3.00%	100	98	93	88	89	77	74
1.00%	91	81	74	65	66	50	46
0.50%	78	66	54	46	46	37	34
0.30%	68	50	41	38	37	27	24
0.10%	39	30	23	15	13	2	2
0.05%	28	14	8	3	2	0	0

Note. Total number of injections per period–amplitude bin: 843; $\sigma = \pm 3.4\%$.

intrinsically variable, so that more injected light curves are used to compute the final statistics. As the threshold is gradually lowered, completeness increases and threshold error decreases, at the cost of reduced reliability. The number of variable stars also grows, and almost all stars are excluded by very low detection thresholds, leaving relatively few in the calculation of the results. Therefore, we settled on $C = 4$ and $S_{\text{min}} = 4$ as the best compromise. When using these values, 238 of the 1877 stars in the Oxford SAP set were marked as intrinsically variable, as were 120 of the 976 stars in the SS set and 23 of the 89 hot stars, leaving 1639, 856, and 66 stars in each set, respectively. Using the same tuning parameters for the CfA pipeline, we are left with 1587 SAP stars, 848 SS stars, and 70 hot stars.

4. Results

We restrict ourselves to presenting the results in this section, and defer a more detailed discussion to Section 5. The completeness and reliability results for the CfA pipeline come from the single scale PDC–MAP, or “PDC–ssMAP.” Across the board, PDC–ssMAP performed better, for our purposes, than PDC–msMAP, as expected (see Section 2.3.3).

4.1. Recovered versus Injected Periods

In Figures 10 through 15, for the hot star, SAP, and SS samples from the Oxford and CfA pipelines, each figure is made up of six panels, corresponding to the injected amplitudes of 3.0%, 1.0%, and 0.50% (top row), and 0.3%, 0.1%, and 0.05% (bottom row). Within each panel, the colored circles represent detections, while the cases that did not pass the detection threshold are shown as smaller black dots. The gray-shaded area in each plot shows the period validity range, though recall that a $\pm 20\%$ phase match is also required. Phase considerations aside, a black point in the shaded area is a missed valid detection, whereas a colored circle outside the shaded area is an invalid detection. The stars marked as intrinsically variable were excluded from these figures.

4.2. Completeness and Reliability

The completeness results for the Oxford and CfA pipelines are shown in the top and bottom rows of Figure 16, respectively. We have identified in blue the completeness for the solar case—where the injected amplitude is 0.1% and the injected period is 25 days—for each data set. Tables 1 through 6 in Appendix A record the completeness values and

Table 6
CfA Superstamp Completeness (%)

$a_{\text{inj}} \backslash P_{\text{inj}}$	5 days	10 days	15 days	20 days	25 days	30 days	35 days
3.00%	93	86	83	74	63	56	51
1.00%	82	70	61	52	45	36	27
0.50%	69	57	50	40	34	27	23
0.30%	58	47	41	34	31	26	16
0.10%	42	31	21	15	15	11	7
0.05%	29	15	10	5	4	4	2

Note. Total number of injections per period–amplitude bin: 844; $\sigma = \pm 3.4\%$.

Table 7
Oxford Hot Star Reliability (%)

$2\alpha_{\text{peak}}/p_{\text{peak}}$	5 \pm 2.5 days	10 \pm 2.5 days	15 \pm 2.5 days	20 \pm 2.5 days	25 \pm 2.5 days	30 \pm 2.5 days	32.5 days+
$\geq 2.00\%$	100 \pm 12.4 [65]	98 \pm 12.5 [64]	100 \pm 12.6 [63]	98 \pm 12.6 [63]	100 \pm 12.7 [62]	100 \pm 12.8 [61]	100 \pm 13.1 [58]
0.75–2.00%	100 \pm 12.6 [63]	100 \pm 12.7 [62]	98 \pm 13.0 [59]	100 \pm 13.1 [58]	98 \pm 13.4 [56]	100 \pm 14.1 [50]	100 \pm 16.4 [37]
0.45–0.75%	100 \pm 13.0 [59]	100 \pm 13.5 [55]	100 \pm 13.9 [52]	100 \pm 16.0 [39]	100 \pm 16.7 [36]	100 \pm 17.1 [34]	94 \pm 24.3 [17]
0.25–0.45%	98 \pm 12.9 [60]	100 \pm 13.5 [55]	100 \pm 14.4 [48]	97 \pm 16.0 [39]	100 \pm 16.0 [39]	100 \pm 22.9 [19]	62 \pm 17.7 [32]
0.075–0.25%	93 \pm 14.7 [46]	100 \pm 17.4 [33]	100 \pm 23.6 [18]	100 \pm 26.7 [14]	100 \pm 30.2 [11]	100 \pm 37.8 [7]	100 \pm 70.7 [2]
$< 0.075\%$	93 \pm 18.6 [29]	100 \pm 25.0 [16]	100 \pm 30.2 [11]	100 \pm 37.8 [7]	100 \pm 37.8 [7]	100 \pm 70.7 [2]	... [0]

uncertainties for each injected period–amplitude bin for all test samples from each pipeline.

The reliability results for the Oxford and CfA pipelines are shown in the top and bottom rows of Figure 17, respectively. The bins with red dots are those with relatively few detections. We consider bins to be insignificant if they have 100 detections or less in the SAP and SS samples and 30 or less in the hot star sample. All three samples, from both pipelines, have very few detections in the longest-period, lowest-amplitude bins. The corresponding (generally very high) reliability values therefore have large uncertainties and should not be treated as very meaningful. The actual values with uncertainties are presented in Tables 7 through 12 in Appendix A.

Finally, Figure 18 shows the injected versus detected amplitudes for valid detections from the Oxford (top) and CfA (bottom) pipelines. The detected amplitudes can differ significantly from the injected ones for several reasons. First, the light curves into which the signals were injected may already contain some power at the corresponding period. Second, the intrinsic noise levels of the light curve will affect the amplitude of the flux as a whole. Lastly, the systematics correction steps can alter the injected signal, in some cases leading to measured amplitudes that are smaller than the injected ones (e.g., the Oxford SS).

5. Discussion

We now discuss the results from our injection tests, starting with a brief evaluation of the effects of periodogram normalization in light of the completeness and reliability results in Section 5.1. We then look at the “ideal case” of the hot star results in Section 5.2 before looking at the SAP and SS samples in Section 5.3. Next, we provide an overall comparison of the samples and the pipelines in Section 5.4. Finally, we will

discuss the implications of the injection tests on M67 rotation studies using *K2* data.

5.1. Residual Long-term Trends

The median periodograms shown in Figure 6 of Section 3.3.1 tell us a lot about the residual trends present in the light curves after full processing from both the Oxford and CfA pipelines. For the SAP and SS samples, the median periodogram rises steeply for periods above 20 days, peaking around ~ 50 –55 days and ~ 40 –45 days, respectively, before decreasing again and flattening off for periods > 80 days (to which we expect little or no sensitivity in *K2* light curves anyhow). While evident in both pipelines, this feature is much more dramatic in the Oxford pipeline. Importantly, the power present at periods relevant to a study of M67 in both hot star samples shows that this is largely non-astrophysical, despite our steps for systematic correction. Furthermore, the differences between the SAP and SS samples are significant: our a priori expectation was that the SS sample might be more problematic than the SAP one, due to increased crowding in the densest parts of M67. However, the median periodograms tell a different story, with considerably more residual power in the 30–70 day period range in the SAP than the SS light curves for both pipelines.

If the raw periodograms were used for period detection, these residual trends would lead to numerous false detections at mid-to-long periods (see Appendix B.1). Our normalization procedure designed to avoid this seems successful, as we record consistently high reliability wherever we have a significant number of detections (see Figure 17 and Tables 7–12 in Appendix A). However, normalization also suppresses the detection of real signals at longer periods, as is apparent in Figure 16. While we may avoid the trap of lingering

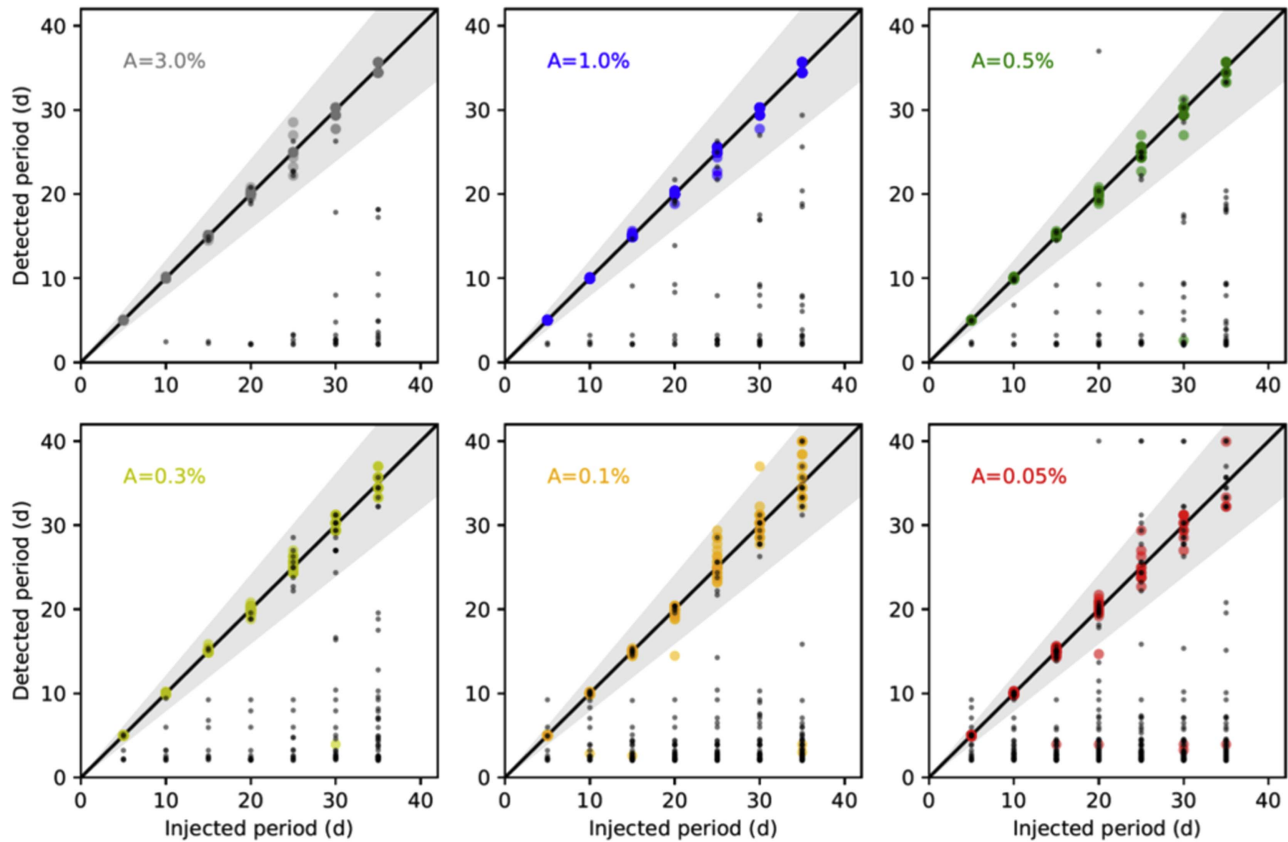


Figure 11. Same as Figure 10, but for the hot star sample processed with the CfA pipeline.

systematics by normalizing the periodograms, this comes at the potential cost of real astrophysical signals.

5.2. The Hot Star Sample

The set of 89 hot stars is an ideal test case, as they should not show significant rotational modulation beyond about five days. Therefore, we can be confident that, if we see a long period where we did not inject one, it is most likely the result of lingering systematics (as we alluded previously). For the Oxford pipeline, we were left with 66 stars after removing stars found to have intrinsic variability (i.e., those likely to have an astrophysical signal that could interfere with the injection test results) using our detection threshold on the non-injected versions of the light curves. Of the 23 stars marked as variable, 19 had detected periods less than 10 days and the rest had periods less than ~ 16 days, apart from one (see Figure 1). In the CfA pipeline, we used 70 stars in the analysis. All 19 variables had periods under 12 days. These numbers further enforce the fact that our normalization works reasonably well for this particular data set in both pipelines, though the presence of flagged variables with periods longer than five days shows that the set of hot stars may also be contaminated by background stars, or that there are a number of cooler F stars in the sample, but it is good enough for our purposes.

As shown in Figures 10 and 11, we do a good job at recovering the injected periods from both pipelines in the hot star sample. Some of the invalid detections likely come from variables—possibly pulsators—not removed from the sample but with low amplitudes. Where these start to become detections at the lowest injected amplitudes, the injected signal

may be “boosting” (i.e., increasing the amplitude of the intrinsic signal) them just enough to be detected. Even for very regularly sampled data like K2, injecting a signal at a given period affects the periodogram at other frequencies in a complex way. However, a number of the invalid detections could also be the result of the normalization, which promotes shorter periods at the expense of longer ones: when dividing out the median periodograms, the much lower power seen at short periods increases short-period significance relative to that of longer periods.

The hot star completeness values for both pipelines can be seen in the leftmost panels of Figure 16 and in Tables 1 and 2 in Appendix A. As expected from simple signal-to-noise arguments, we see completeness decrease with increasing injected period and decreasing injected amplitude for both pipelines. We point out that the best-case scenario for recovering a solar-like signal of 25 days and 0.1% amplitude is $\sim 45\%$ from the hot star data set. However, the limited size of the hot star sample means that the individual completeness values are somewhat uncertain ($\pm \sim 12\%$) for both pipelines.

The reliability in the hot star sample from either pipeline is consistently high ($>90\%$). High reliability means that we can typically trust a detection made within a measured amplitude and period range. We can see from Figure 17 that, where we have a significant number of detections (more than 30 for the hot star samples), we can be fairly confident in our results, though again, the uncertainties are relatively large for the hot star samples. The high reliability also indicates that normalization did not introduce low-frequency signals in light curves that did not have a strong intrinsic signal in the first place, again reinforcing the validity of this step.

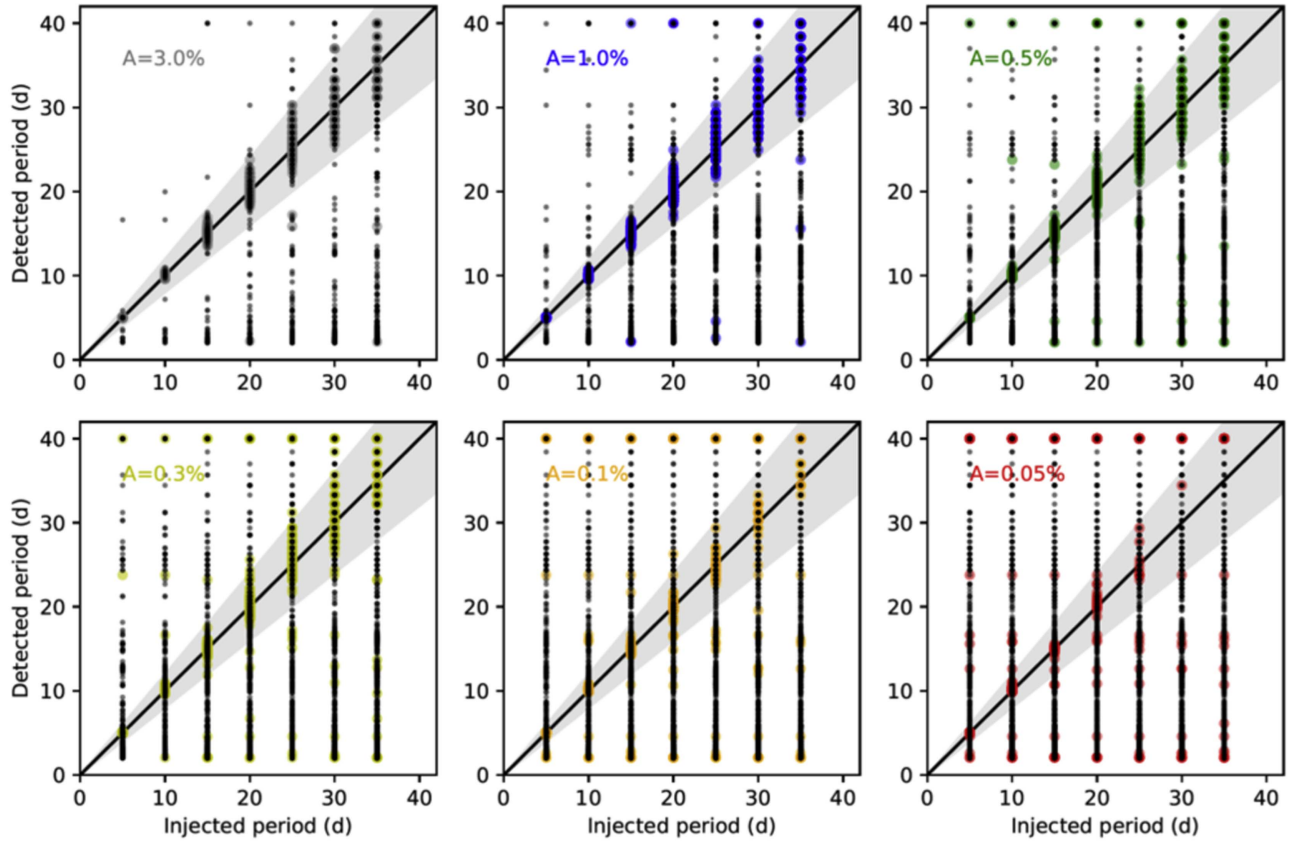


Figure 12. Same as Figure 10, but for the SAP sample processed with the Oxford pipeline.

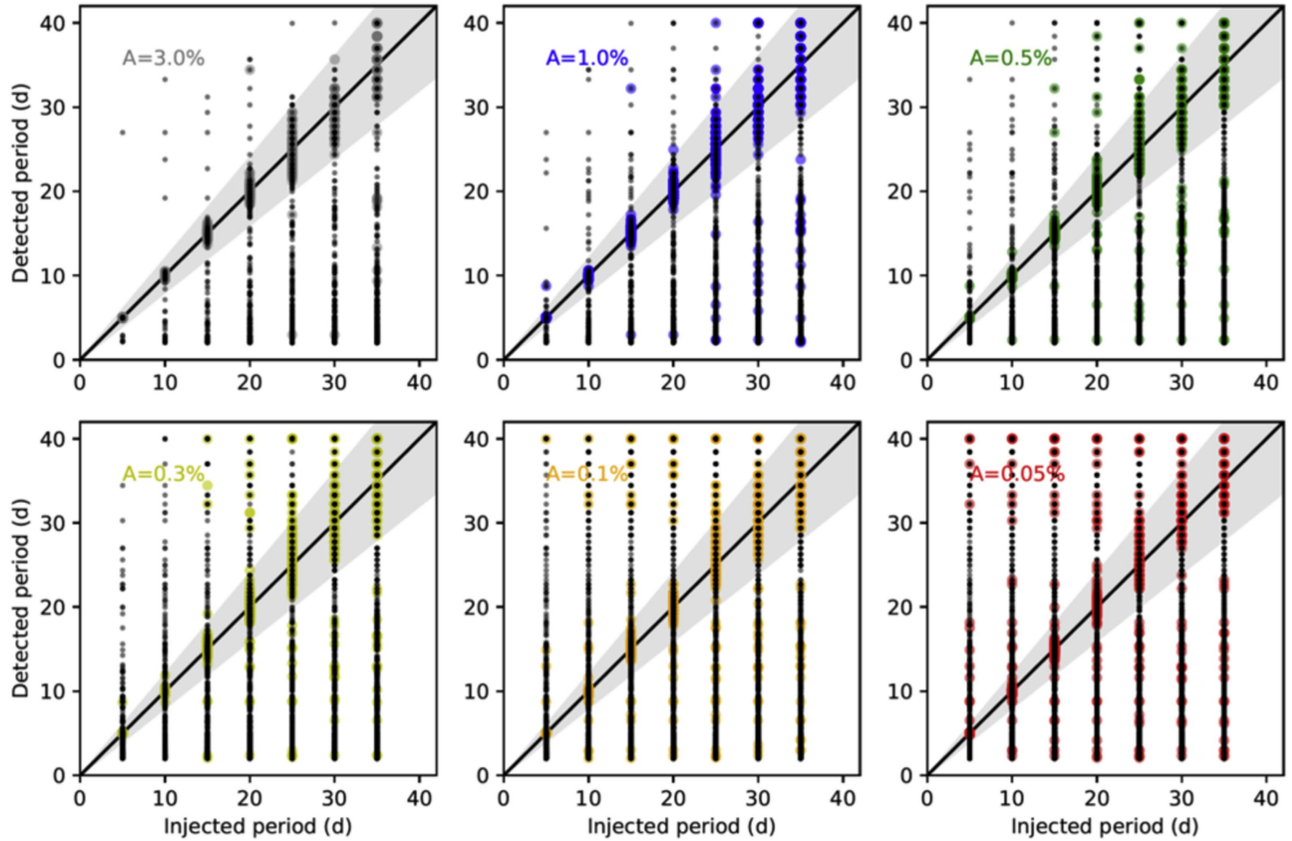


Figure 13. Same as Figure 10, but for the SAP sample processed with the CfA pipeline.

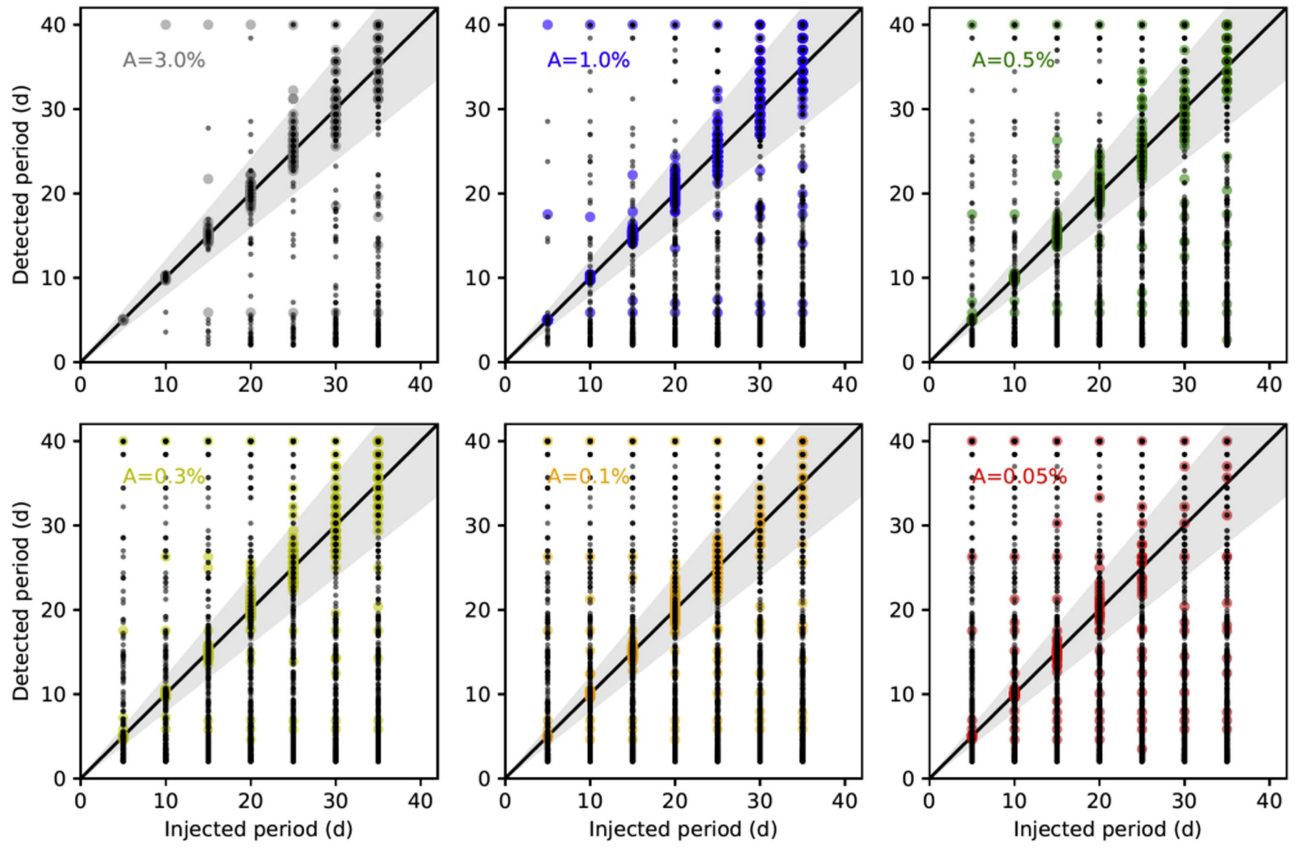


Figure 14. Same as Figure 10, but for the SS sample processed with the Oxford pipeline.

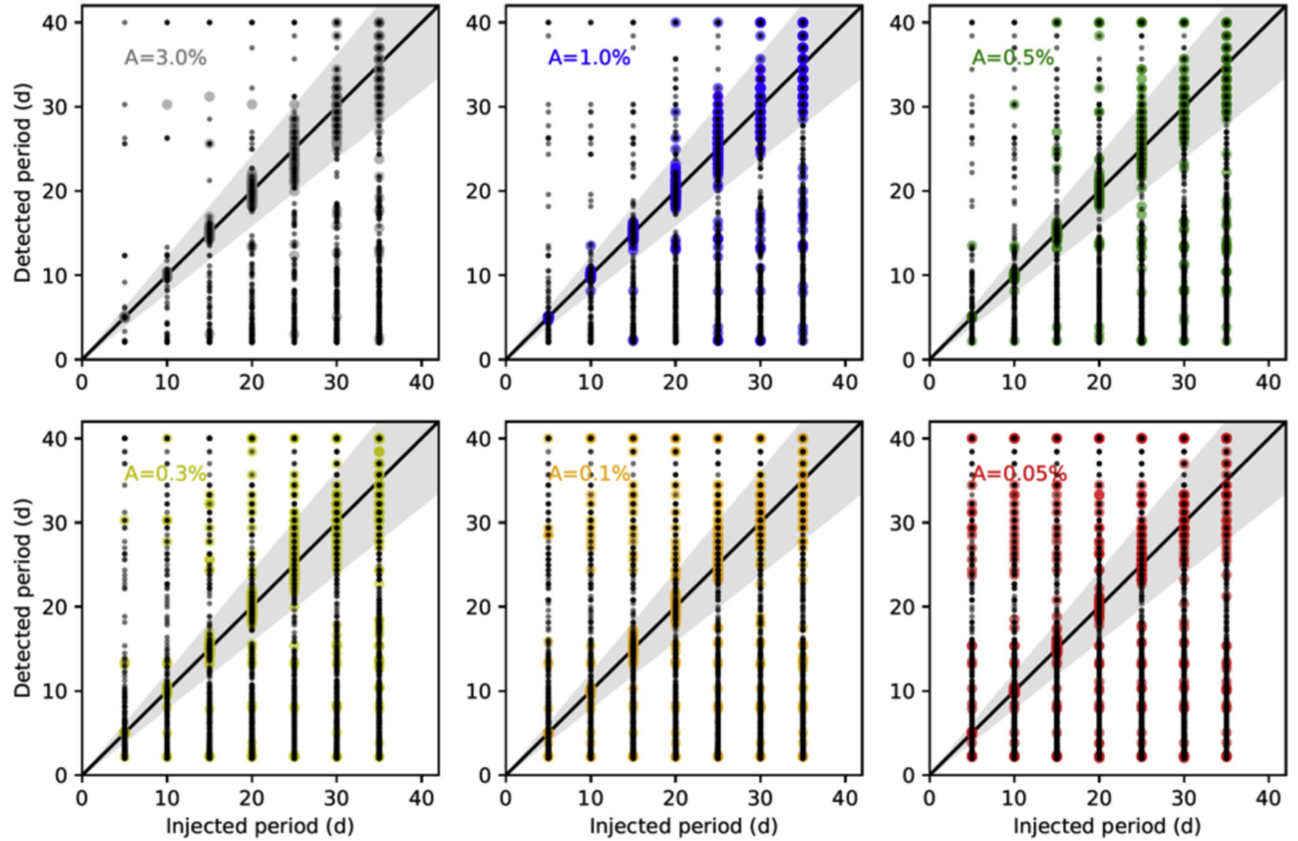


Figure 15. Same as Figure 10, but for the SS sample processed with the CfA pipeline.

Table 8
CfA Hot Star Reliability (%)

$2\alpha_{\text{peak}}/p_{\text{peak}}$	5 \pm 2.5 days	10 \pm 2.5 days	15 \pm 2.5 days	20 \pm 2.5 days	25 \pm 2.5 days	30 \pm 2.5 days	32.5 days+
$\geq 2.00\%$	100 \pm 12.0 [70]	100 \pm 12.0 [69]	100 \pm 12.3 [66]	100 \pm 12.7 [62]	100 \pm 13. [59]	100 \pm 13.4 [56]	100 \pm 13.4 [56]
0.75–2.00%	100 \pm 12.1 [68]	98 \pm 12.3 [66]	100 \pm 12.5 [64]	100 \pm 13.0 [59]	100 \pm 13.6 [54]	100 \pm 13.2 [57]	100 \pm 13.7 [53]
0.45–0.75%	100 \pm 12.5 [64]	100 \pm 12.7 [62]	100 \pm 13.0 [59]	100 \pm 13.2 [57]	100 \pm 13.9 [52]	100 \pm 15.1 [44]	100 \pm 17.1 [34]
0.25–0.45%	100 \pm 12.4 [65]	100 \pm 12.9 [60]	100 \pm 13.0 [59]	100 \pm 13.1 [58]	100 \pm 14.1 [50]	98 \pm 15.6 [41]	100 \pm 18.0 [31]
0.075–0.25%	95 \pm 12.8 [61]	100 \pm 14.3 [49]	98 \pm 15.2 [43]	100 \pm 17.4 [33]	100 \pm 18.9 [28]	100 \pm 18.9 [28]	87 \pm 25.8 [15]
$< 0.075\%$	84 \pm 13.5 [55]	97 \pm 16.9 [35]	96 \pm 20.4 [24]	93 \pm 25.8 [15]	100 \pm 27.7 [13]	100 \pm 28.9 [12]	100 \pm 57.7 [3]

Table 9
Oxford SAP Reliability (%)

$2\alpha_{\text{peak}}/p_{\text{peak}}$	5 \pm 2.5 days	10 \pm 2.5 days	15 \pm 2.5 days	20 \pm 2.5 days	25 \pm 2.5 days	30 \pm 2.5 days	32.5 days+
$\geq 2.00\%$	100 \pm 2.5 [1620]	99 \pm 2.5 [1598]	99 \pm 2.6 [1525]	100 \pm 2.7 [1414]	99 \pm 2.7 [1405]	99 \pm 2.6 [1440]	96 \pm 3.2 [996]
0.75–2.00%	100 \pm 2.6 [1523]	98 \pm 2.7 [1364]	99 \pm 2.8 [1250]	100 \pm 3.0 [1076]	98 \pm 3.0 [1095]	99 \pm 3.1 [1025]	88 \pm 4.7 [456]
0.45–0.75%	100 \pm 2.8 [1267]	98 \pm 3.1 [1065]	98 \pm 3.5 [812]	100 \pm 4.3 [534]	98 \pm 4.0 [623]	98 \pm 4.5 [495]	46 \pm 6.2 [261]
0.25–0.45%	98 \pm 2.8 [1278]	98 \pm 3.1 [1074]	98 \pm 3.6 [768]	100 \pm 4.4 [519]	100 \pm 5.1 [391]	99 \pm 6.8 [214]	33 \pm 7.4 [184]
0.075–0.25%	95 \pm 3.5 [832]	97 \pm 5.1 [386]	85 \pm 7.8 [165]	99 \pm 10.3 [95]	98 \pm 14.4 [48]	95 \pm 22.9 [19]	20 \pm 14.3 [49]
$< 0.075\%$	83 \pm 5.3 [358]	97 \pm 12.4 [65]	100 \pm 22.4 [20]	100 \pm 25.0 [16]	100 \pm 44.7 [5]	100 \pm 100.0 [1]	100 \pm 100.0 [1]

5.3. SAP and SS Completeness and Reliability

We have shown that the hot star sample is important for validating our injection test procedures and highlighting the presence of non-astrophysical power around 25 days and longer in the *K2* Campaign 5 light curves. We now discuss the results from the SAP and SS samples, which specifically illustrate the complexity of a period search in M67. Figures 12–15, which show the detected versus injected periods for the SAP and SS data sets from both pipelines, are a lot messier than their hot star counterparts, most likely due to the larger sample sizes and more diverse stellar populations within. However, like the hot stars, detections become more difficult and less reliable as the injected period increases and the injected amplitude decreases. In addition, lingering power around 10–20 days seems to exist in the SAP and SS light curves. This power was either not originally strong enough to mark the light curves as variables until boosted by a lower-amplitude, injected signal, or it could be the result of half-period harmonic measurements from the Lomb–Scargle periodogram.

The completeness results for both the SAP and SS samples are in the middle and far right panels, respectively, in Figure 16. As with the hot stars, we see the same general trend of diminishing completeness with increasing injected period and decreasing amplitude, but it has been exacerbated. For both samples from either pipeline, collapsing the figures onto either axis, the completeness falls around or below 50% for amplitudes $\leq 0.50\%$ and periods ≥ 20 days, showing just how difficult it is to detect long-period, low-amplitude signals

in either case. Of particular importance is the solar case, highlighted in blue in Figure 16, where the injected period is 25 days and the amplitude is 0.10%. Critically, the completeness here is $\sim 15\%$ or lower for both the SAP and SS samples in either pipeline. Even with the best-case scenario of perfect sinusoidal signals, solar variability, which we expect in M67, is clearly difficult to find with our detection criteria.

The middle and far right panels of Figure 17 present the reliability for the SAP and SS samples. As with the hot stars, where there is a significant number of detections (> 100 for the larger data sets), the reliability for the SAP and SS samples rarely drops below 90% for both pipelines. This is encouraging, as it shows that the procedure we have developed should lead to relatively few false alarms, i.e., where we do get a detection, we can generally trust the period measurement.

There are a few striking features in Figure 17, however. Detecting long periods (particularly at ~ 35 days) in the Oxford SAP sample appears to be not only more difficult than the CfA pipeline, but also less reliable. The low reliability in this period range could either be a result of periodogram normalization or a failure to correct the residual long-term trends in the data, which the CfA pipeline does better. The Oxford SAP sample also has lower reliability in the shortest-period, lowest-amplitude bin. This could be an effect of promoting short-period signals when dividing out the median periodogram. We experimented with adding a “floor” to the Oxford median periodograms, via which we set any value below a given threshold to that floor in an attempt to reduce this effect. We tested floor values of 0.005 and 0.01. While these did improve

Table 10
CfA SAP Reliability (%)

$2\alpha_{\text{peak}}/p_{\text{peak}}$	5 \pm 2.5 days	10 \pm 2.5 days	15 \pm 2.5 days	20 \pm 2.5 days	25 \pm 2.5 days	30 \pm 2.5 days	32.5 days+
$\geq 2.00\%$	98 \pm 2.5 [1567]	95 \pm 2.5 [1539]	99 \pm 2.6 [1426]	100 \pm 2.8 [1293]	97 \pm 2.9 [1194]	98 \pm 2.9 [1172]	100 \pm 3.1 [1070]
0.75–2.00%	100 \pm 2.6 [1470]	98 \pm 2.8 [1316]	96 \pm 2.8 [1245]	99 \pm 3.0 [1078]	99 \pm 3.2 [954]	96 \pm 3.3 [903]	92 \pm 3.9 [667]
0.45–0.75%	98 \pm 2.8 [1313]	97 \pm 3.1 [1057]	93 \pm 3.1 [1011]	95 \pm 3.5 [797]	96 \pm 3.8 [677]	92 \pm 4.2 [567]	90 \pm 5.5 [333]
0.25–0.45%	98 \pm 2.9 [1213]	93 \pm 3.1 [1050]	95 \pm 3.3 [896]	98 \pm 3.7 [721]	96 \pm 4.0 [640]	93 \pm 4.0 [618]	82 \pm 5.2 [365]
0.075–0.25%	96 \pm 3.3 [915]	98 \pm 4.2 [575]	91 \pm 4.8 [426]	98 \pm 5.5 [329]	95 \pm 6.8 [217]	94 \pm 6.5 [237]	75 \pm 7.8 [165]
$< 0.075\%$	95 \pm 4.2 [572]	95 \pm 5.9 [291]	95 \pm 8.6 [134]	100 \pm 13.4 [56]	88 \pm 15.2 [43]	100 \pm 16.7 [36]	93 \pm 26.7 [14]

reliability in the short-period, low-amplitude bin, the floors tended to reduce reliability overall in all three Oxford samples, particularly the hot star and SS sets, and especially around mid-range periods. Thus, we decided to exclude a floor from the study to avoid further artificial effects.

One surprising feature in the SS samples is the lack of detections across all periods at amplitudes of 2.0% and greater for the Oxford pipeline, even though many 3.0% signals were injected. This can be seen in both Figures 17 and 18. If we then compare the number of SS detections from one pipeline to another in the amplitude ranges 0.75–2.00% and 0.075–0.25% in Tables 11 and 12 of Appendix A, there are systematically more detections in the Oxford pipeline for these ranges than in the CfA pipeline. This indicates that the amplitudes of the Oxford SS sample are suppressed, primarily after the PCA. Though the PCA may also suppress amplitudes in the Oxford SAP sample, it is more evident in the SS sample due to the high noise amplitude of the SS PCs compared to the SAP PCs. The amplitude suppression could also indicate that the Oxford pipeline removes some intrinsic variability in addition to the systematics. However, it seems that, despite the suppression, the signals remain intact enough to be detected at a similar rate, just at lower amplitudes.

5.4. Comparing the Samples and Pipelines

In both pipelines, the hot star sample has the highest overall completeness and reliability. This is unsurprising, given its smaller sample size and lack of diversity in terms of variability. Due to the general absence of rotational modulation and other competing signals in the original light curves, the hot stars should also better preserve the injected sinusoids. Finally, the average brightness of the hot star data set is $K_p = 11.0$, while the combined average of the SAP and SS data sets is $K_p = 15.3$, and the number of valid detections generally decreases with increasing magnitude.

The Oxford SS completeness is slightly lower than the Oxford SAP at short periods and large amplitudes, but it is better for periods ≥ 15 days and amplitudes below about 0.50%. In addition, the SS sample is more reliable for both the longest and shortest detected periods. These differences are likely due to the separate light curve extraction methods in the Oxford pipeline. The SAP sample is extracted with pixelated apertures, which may not be optimal for a crowded field, even in the outer regions of the cluster. The SS sample, however, uses deblended, circular apertures, making it better able to deal with crowding.

On the other hand, the CfA SAP completeness and reliability are quite comparable (or slightly higher) across most bins when compared against the CfA SS sample. The slight advantage to the SAP sample is probably due to less crowding, as the two data sets are extracted and processed the same way in the CfA pipeline. In addition, the larger size of the SAP sample means that there is a larger matrix from which to characterize and correct the common-mode trends via PDC-MAP (or PCA, in the case of the Oxford pipeline). This could help to partially explain the generally higher completeness values from the SAP sample in either pipeline, as well as the higher reliability in the case of the CfA pipeline.

With respect to completeness, the CfA pipeline seems to perform slightly better than the Oxford pipeline, particularly in the hot star and SAP samples. Thus, Figure 19 provides the average completeness (and associated reliability in smaller text) from the CfA SAP and SS samples, as a summary of the best we can do with the *K2* Campaign 5 M67 data, with the solar case highlighted in red. The differences between the Oxford and CfA pipelines are rooted in the approaches each takes to produce corrected light curves from raw *K2* data. The moving apertures and deblending procedures during light curve preparation give the Oxford pipeline an edge in certain cases, particularly with the SS sample, but the larger, more optimized aperture selection from the CfA pipeline is superior elsewhere. While the Oxford pipeline typically better removes systematics from strongly variable stars, most M67 targets won't be sufficiently variable on short timescales, so the advantage is minimal. Finally, the more sophisticated PDC-MAP outperforms the relatively crude PCA of the Oxford pipeline in the common-mode systematic removal. While an ideal pipeline would combine the best elements of the two, the performance of the existing pipelines is comparable, and both are valid options regarding the analysis of *K2* light curves. Most critically, however, is that for both pipelines, the completeness around the solar case in the SAP and SS samples is $\sim 15\%$ at best, illustrating that it is very difficult to detect 0.1% amplitude, 25 days signals in *K2* Campaign 5 M67 data.

5.5. Implications for *K2* M67 Rotation Studies

The injection test results indicate that, if the Sun is considered typical, measuring rotation periods for solar-like stars in M67 based on their *K2* light curves is challenging. Specifically, for the best-case scenario of solar-like rotational variability—0.1% amplitude, 25 days non-evolving sinusoidal signal—our completeness maxes out at 14% and 15% for the

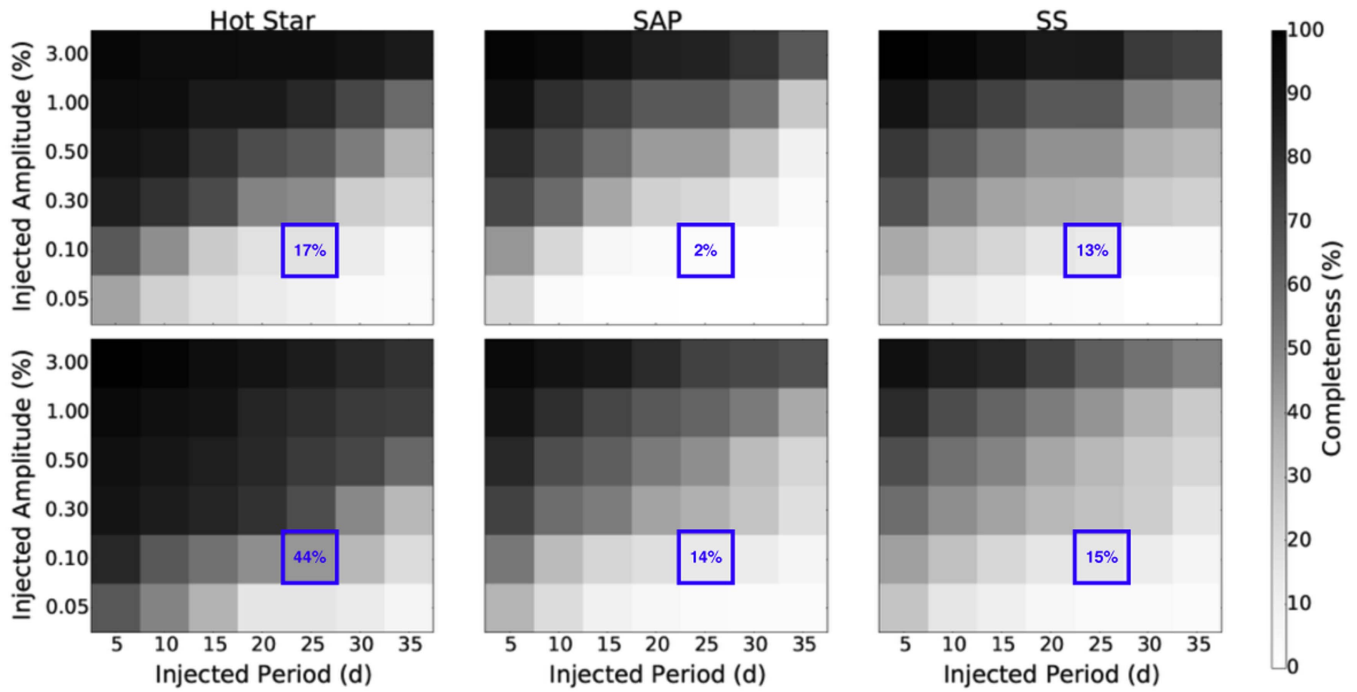


Figure 16. Completeness tables for the hot star (left), SAP (center), and SS samples (right), detrended with the Oxford pipeline (top row) and the CfA pipeline (bottom row). The grids have the injected period bin along the x -axis and the injected amplitude along the y -axis. The completeness for the solar case is highlighted in blue for each data set.

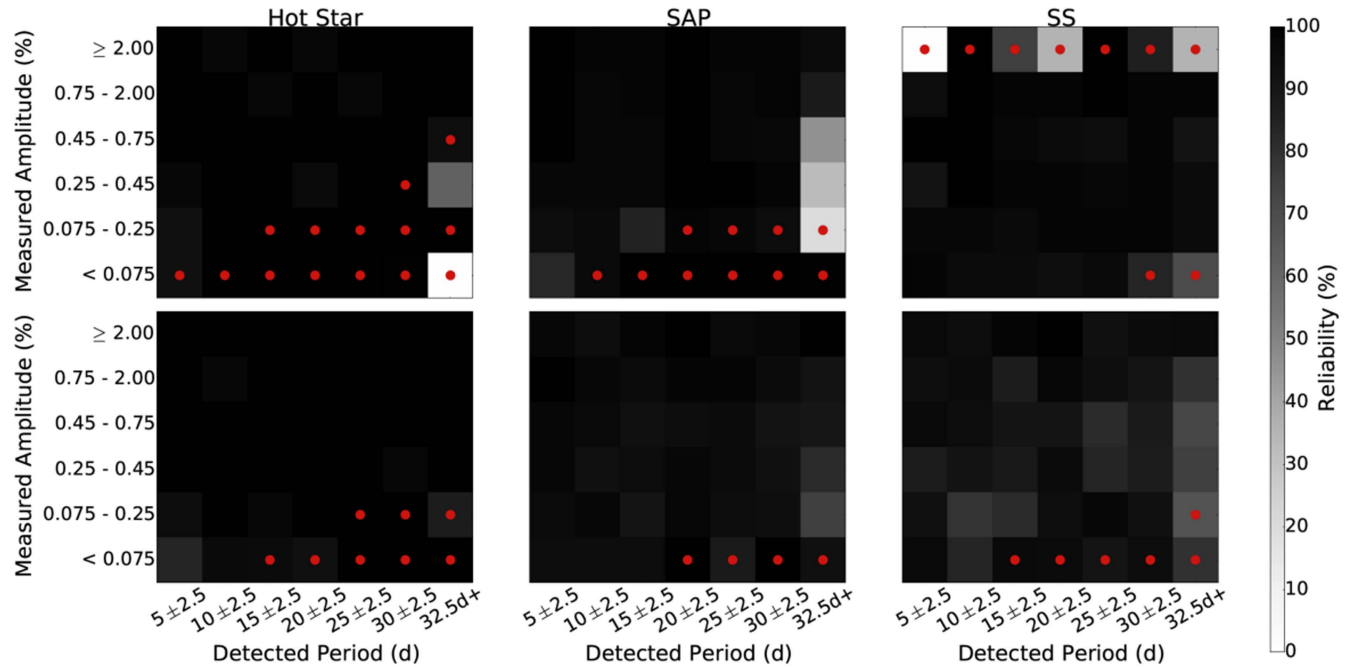


Figure 17. Reliability for the hot star (left), SAP (center), and SS samples (right), detrended with the Oxford pipeline (top row) and the CfA pipeline (bottom row). While the bins shown here roughly match those in Figure 16, here they represent *detected* rather than injected period and amplitude. While the same number of injections was carried out for every injected period and amplitude, the measured and injected values do not necessarily match, so the number of detections in each bin varies. The red circles represent where there were less than 30 detections in the hot star sample and less than 100 in the SAP and SS samples.

SAP and SS samples, respectively. Considering that detecting non-sinusoidal, evolving spot-modulation will be more difficult, these results suggest that both current and future rotation period measurements in M67, based solely on the *K2* Campaign 5 data, must be carefully examined before they are

used to guide models of stellar angular momentum evolution and/or calibrate gyrochronology relations.

We consider the 20 stars with measured rotation periods from Barnes et al. (2016) and the stars from Gonzalez (2016a) where we had detections from either the CfA or Oxford pipeline. Recall

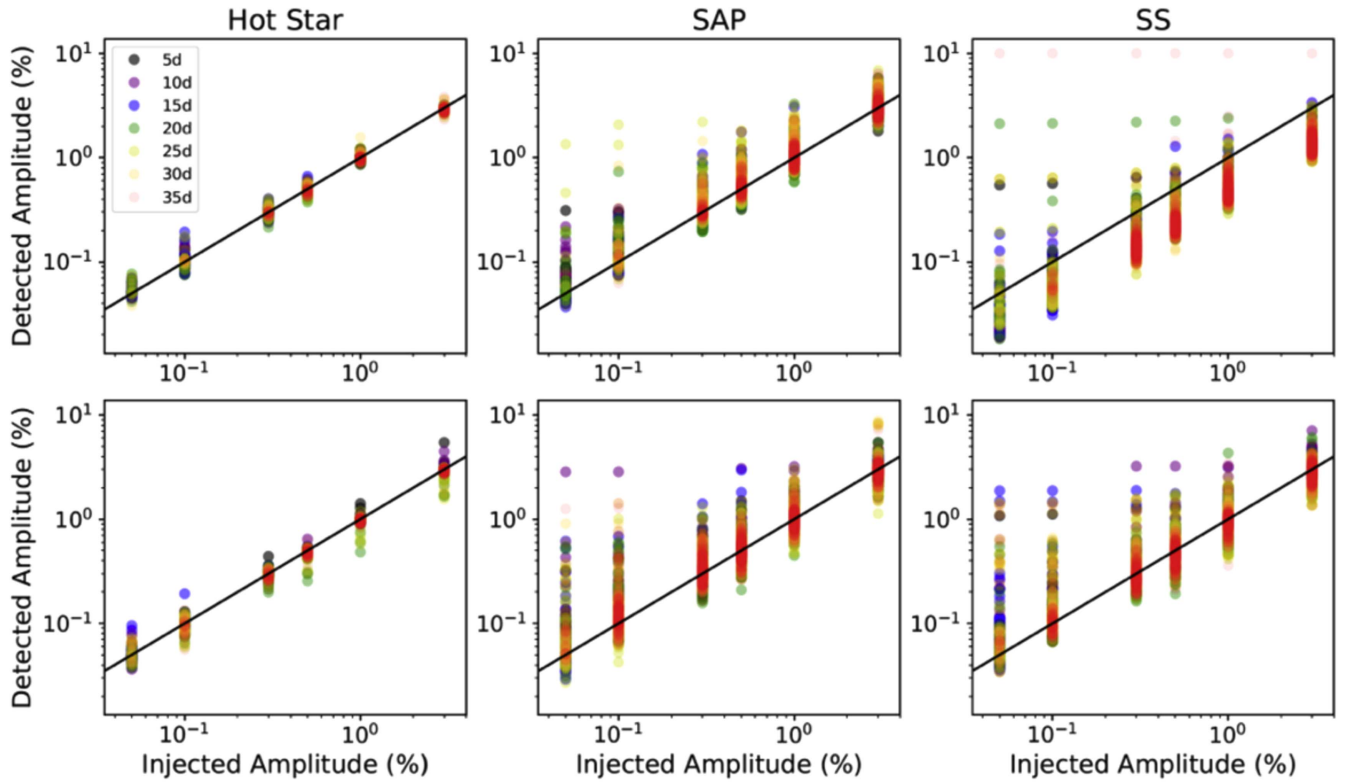


Figure 18. Injected vs. detected amplitude for the valid detections in the hot star (left), SAP (center), and SS samples (right), detrended with the Oxford pipeline (top row) and CfA pipeline (bottom row). The associated injected periods are marked in different colors.

Table 11
Oxford Superstamp Reliability (%)

$2\alpha_{\text{peak}}/p_{\text{peak}}$	5 \pm 2.5 days	10 \pm 2.5 days	15 \pm 2.5 days	20 \pm 2.5 days	25 \pm 2.5 days	30 \pm 2.5 days	32.5 days+
$\geq 2.00\%$...	100 \pm 100.0	75 \pm 25.0	35 \pm 15.8	100 \pm 27.7	86 \pm 37.8	36 \pm 13.4
	[0]	[1]	[16]	[40]	[13]	[7]	[56]
0.75–2.00%	95 \pm 3.4	100 \pm 3.5	99 \pm 3.5	99 \pm 3.6	100 \pm 3.6	99 \pm 3.9	99 \pm 4.0
	[886]	[826]	[797]	[752]	[760]	[657]	[620]
0.45–0.75%	100 \pm 3.7	100 \pm 3.9	98 \pm 4.1	96 \pm 4.5	95 \pm 4.5	99 \pm 6.8	92 \pm 7.2
	[739]	[652]	[584]	[497]	[501]	[215]	[191]
0.25–0.45%	92 \pm 4.9	100 \pm 5.7	99 \pm 6.0	99 \pm 6.6	98 \pm 6.4	99 \pm 6.0	96 \pm 6.2
	[421]	[311]	[279]	[227]	[246]	[274]	[262]
0.075–0.25%	98 \pm 3.3	98 \pm 3.7	97 \pm 4.1	99 \pm 4.2	99 \pm 4.2	99 \pm 4.6	96 \pm 4.8
	[898]	[713]	[601]	[558]	[563]	[478]	[442]
<0.075%	99 \pm 4.2	96 \pm 5.1	96 \pm 6.3	94 \pm 8.3	97 \pm 9.4	84 \pm 20.0	70 \pm 22.4
	[557]	[388]	[253]	[145]	[114]	[25]	[20]

that all of these light curves come from the SAP sample. Because we do not have amplitude measurements for the versions of the light curves used in either paper, we estimated them by binning the flux of each light curve based on the reported period and taking the median of the difference between the 95th and 5th percentiles of each bin. We did this for both the Oxford and CfA versions of each light curve and averaged the values. We then estimated the completeness and reliability based on the corresponding SAP results from this study. We averaged the Oxford and CfA completeness for the injected period and amplitude bins closest to the determined values. If the calculated amplitude was halfway between the injected amplitudes, we used the higher amplitude bin. We also used the average Oxford and CfA reliability from the appropriately ranged bins into which the reported periods and estimated amplitudes fell.

The results are shown in Tables 13 and 14. The completeness for the reported periods from Barnes et al. (2016) is low, averaging at about 27%, with only one period falling into a range above 50% (EPIC 211394185). The corresponding reliability is high, only dipping below 95% in one instance (EPIC 211397512). However, these values are a bit misleading, as they only really apply to those cases where we would actually get a detection according to our criteria, i.e., using a threshold factor of $C = 4$. If we have a detection in an area with low completeness, then the result is generally trustworthy. Therefore, in Table 13, we also show the period from the normalized Lomb–Scargle for the corresponding, non-injected light curves from the Oxford and CfA pipelines where we had a detection based on our threshold (i.e., where the star was classified as a “variable”). For this sample of stars, the Oxford

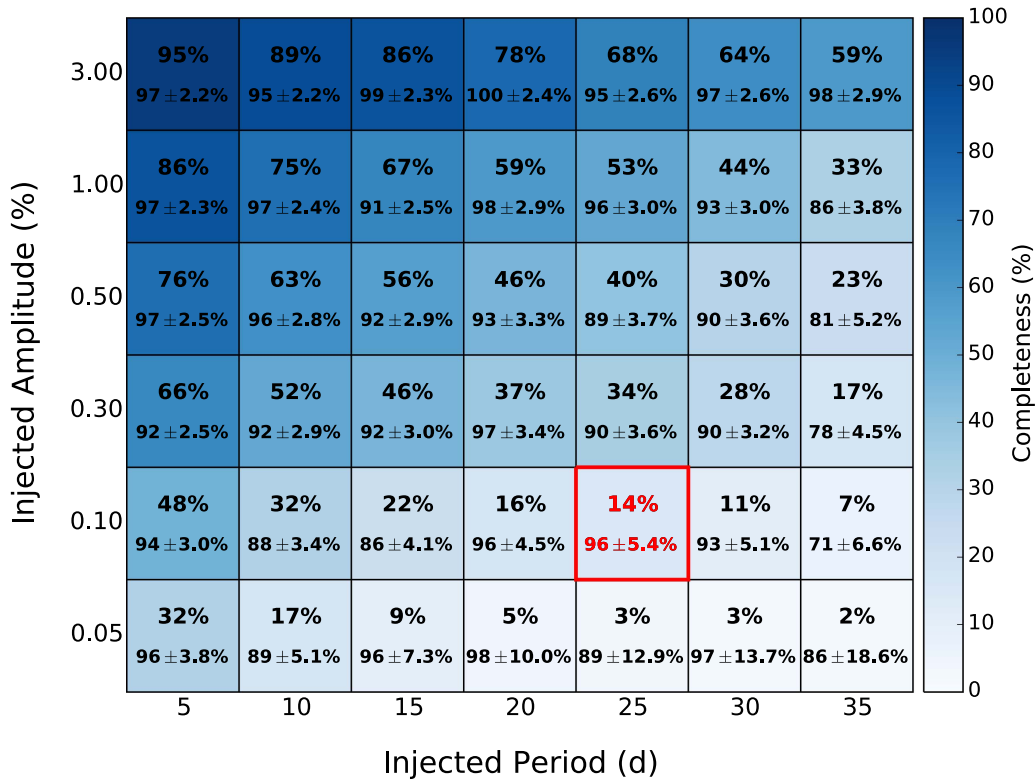


Figure 19. Average completeness (top number) and the roughly associated reliability (bottom number with uncertainty) for the CfA SAP and SS samples, with the solar case highlighted in red. The uncertainty for the completeness values is $\pm 2.1\%$. Note that the reliability values do not perfectly match the injected amplitudes and periods, but are representative of the ranges shown in Figure 17.

Table 12
CfA Superstamp Reliability (%)

$2\alpha_{\text{peak}}/p_{\text{peak}}$	5 ± 2.5 days	10 ± 2.5 days	15 ± 2.5 days	20 ± 2.5 days	25 ± 2.5 days	30 ± 2.5 days	32.5 days+
$\geq 2.00\%$	97 ± 3.5 [810]	95 ± 3.6 [768]	99 ± 3.8 [703]	100 ± 4.0 [620]	93 ± 4.3 [545]	96 ± 4.4 [513]	97 ± 4.9 [411]
0.75–2.00%	94 ± 3.7 [740]	96 ± 4.0 [619]	87 ± 4.0 [613]	98 ± 4.8 [428]	94 ± 5.0 [394]	91 ± 5.0 [405]	80 ± 6.4 [241]
0.45–0.75%	97 ± 4.2 [568]	95 ± 4.6 [477]	91 ± 4.9 [418]	91 ± 5.5 [329]	82 ± 6.3 [256]	88 ± 5.8 [302]	72 ± 8.8 [128]
0.25–0.45%	87 ± 4.1 [587]	92 ± 4.8 [434]	89 ± 4.9 [418]	96 ± 5.7 [312]	84 ± 6.0 [278]	87 ± 5.0 [396]	75 ± 7.4 [183]
0.075–0.25%	93 ± 5.0 [400]	79 ± 5.3 [356]	82 ± 6.6 [232]	94 ± 7.1 [199]	98 ± 8.5 [139]	93 ± 8.0 [158]	67 ± 10.6 [89]
$< 0.075\%$	97 ± 6.4 [241]	84 ± 8.4 [141]	97 ± 11.9 [71]	96 ± 14.9 [45]	91 ± 20.9 [23]	95 ± 21.8 [21]	80 ± 25.8 [15]

pipeline only has three detections, while the CfA has four. Almost all of these detections appear to be rough harmonics of the published result from Barnes et al. (2016), while the CfA period for EPIC 211423010 is close to the Barnes value.

Both the Oxford and CfA pipelines have detections for EPICs 211394185 and 211411621 from the Barnes sample. Figures 20 and 21 show the Oxford and CfA versions of these two light curves, respectively, along with the original periodograms in the middle panels and the normalized versions in the bottom panels. While obviously variable, a clear period is difficult to find by eye in either case, though it appears that the normalization suppressed any power at a period that would be comparable to the Barnes result. The normalization step is important for avoiding false positives, however, as we have

illustrated with the presence of long-term trends in the median periodogram of the hot star samples. We recognize that Barnes et al. (2016) used several different period detection methods other than the Lomb–Scargle periodogram. However, the concern here is obviously the underlying signal. The lack of detections from the Oxford and CfA pipelines, as well as the lack of agreement with the Barnes result where we do have one, show just how challenging it is to measure solar-like signals in the K2 Campaign 5 data.

Looking at Table 14, we have a list of 18 EPICs from either the “PDCSAP” or “K2SC” samples given in Gonzalez (2016a) where we had a detection from at least one of our pipelines. The average completeness is $\sim 60\%$, which is higher than the Barnes sample, and the reliability is $\geq 95\%$ (except in one case,

Table 13
Comparison with Periods from Barnes et al. (2016)

EPIC	Barnes Per (days)	Amp (%)	Comp (%)	Rel (%)	Ox Per (days)	CfA Per (days)
211388204	31.8	0.32	21 ± 1.8	96 ± 3.9
211394185	30.4	0.78	55 ± 1.8	98 ± 2.3	15.4	15.1
211395620	30.7	0.47	32 ± 1.8	95 ± 3.1
211397319	25.1	0.31	29 ± 1.8	98 ± 3.2
211397512	34.5	0.73	34 ± 1.8	68 ± 4.1	16.4	...
211398025	28.8	0.29	21 ± 1.8	96 ± 3.9
211398541	30.3	0.51	32 ± 1.8	95 ± 3.1
211399458	30.2	0.66	32 ± 1.8	95 ± 3.1
211399819	28.4	0.31	21 ± 1.8	96 ± 3.9
211400500	26.9	0.30	29 ± 1.8	98 ± 3.2
211406596	26.9	0.36	29 ± 1.8	98 ± 3.2
211410757	18.9	0.14	11 ± 1.8	99 ± 5.8
211411477	31.2	0.20	21 ± 1.8	95 ± 11.9
211411621	30.5	0.30	21 ± 1.8	96 ± 3.9	15.8	16.1
211413212	24.4	0.22	29 ± 1.8	97 ± 8.0
211413961	31.4	0.26	21 ± 1.8	96 ± 3.9
211414799	18.1	0.17	11 ± 1.8	99 ± 5.8	...	9.0
211423010	24.9	0.29	29 ± 1.8	98 ± 3.2	...	22.2
211428580	26.9	0.33	29 ± 1.8	98 ± 3.2
211430274	31.1	0.40	32 ± 1.8	96 ± 3.9

Notes. The columns, from left to right, are EPIC, the associated period reported from Barnes et al. (2016), our estimate of the peak-to-peak amplitude for the light curve, the estimated completeness and reliability statistics based off the Barnes period and amplitude, and the Oxford and CfA periods where we have detections using $C = 4$.

Table 14
Comparison with Detected Periods from Gonzalez (2016a)

EPIC	Gonzalez Per (days)	Amp (%)	Comp (%)	Rel (%)	Ox Per (days)	CfA Per (days)
211387834	15.2	1.36	74 ± 1.8	98 ± 2.0	14.9	14.9
211390071	25.3	1.42	63 ± 1.8	99 ± 2.2	13.5	...
211393422	28.6	1.53	55 ± 1.8	98 ± 2.3	14.5	...
211397501	12.4, 12.4	1.19	81 ± 1.8	98 ± 1.9	12.2	12.5
211397955	29.0, 29.2	1.36	55 ± 1.8	98 ± 2.3	14.7	14.5
211400662	13.5	2.04	91 ± 1.8	99 ± 1.8	13.7	13.1
211403852	27.3	1.23	63 ± 1.8	99 ± 2.2	13.9	13.3
211404310	26.3	1.17	63 ± 1.8	99 ± 2.2	...	13.0
211405671	25.6, 28.2	2.06, 2.35	77, 76 ± 1.8	98 ± 2.0, 99 ± 1.9	26.3	...
211407277	24.7	2.50	77 ± 1.8	98 ± 2.0	...	25.0
211408116	30.1	0.14	6 ± 1.8	95 ± 11.9	13.5	...
211408874	23.9, 24.5	2.12	80 ± 1.8	98 ± 2.0	...	23.8
211414974	26.8, 29.9	1.91, 1.92	63 ± 1.8	97 ± 2.1, 98 ± 2.3	27.0	...
211424980	15.6, 31.2	1.26, 1.49	74 ± 1.8	98 ± 2.0, 98 ± 2.3	15.6	15.6
211427666	28.8	1.20	55 ± 1.8	98 ± 2.3	14.1	...
211429354	25.9	0.19	8 ± 1.8	97 ± 8.0	...	26.3
211430648	32.6	0.67	16 ± 1.8	68 ± 4.1	...	17.2
211433352	13.7, 13.8	1.21	74 ± 1.8	98 ± 2.0	13.9	13.5

Notes. The columns, from left to right, are the EPICs from Gonzalez (2016a) where either of our two pipelines have detections, the associated periods reported from that paper, our estimate of the peak-to-peak amplitude for each light curve, the estimated completeness and reliability statistics based off the Gonzalez period and amplitude, and the Oxford and CfA periods where we have detections using $C = 4$. Where there is more than one reported value, the first number comes from the PDCSAP sample from Gonzalez (2016a), while the second number is from the K2SC sample.

EPIC 211430648), but the stars listed here are only a small fraction of those published by Gonzalez (2016a). The PDCSAP sample in total has 98 reported periods; we detected 16 from this sample. The K2SC sample had 40 periods; we detected 9. In addition, notice that none of the stars in Table 13 show up in 14, meaning that neither the Oxford nor the CfA pipeline records a detection of the nine stars reported by Gonzalez (2016a) that overlap with Barnes et al. (2016). However, as with the Barnes detections, the CfA and Oxford results agree,

and they are either in agreement with the Gonzalez value, or they are rough harmonics. The harmonics again highlight ambiguity in the light curves, while the general lack of detections from the Gonzalez sample—especially where reported from Barnes—reinforces the conclusions we have already made about finding solar signals in M67 with K2 data.

We can use the completeness and reliability statistics from this work to guide future re-evaluations of M67 periods from

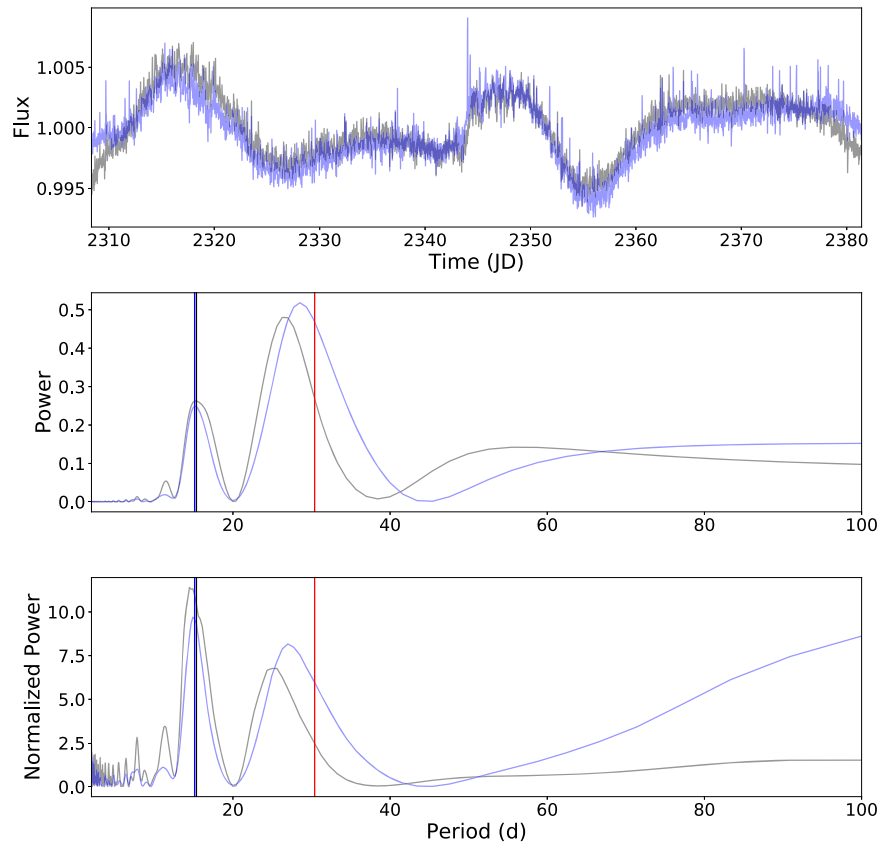


Figure 20. Light curves and periodograms for EPIC 211394185 from the Oxford (gray) and CfA (blue) pipelines. The top panel displays the light curves, while the middle panel gives the original periodogram power for both pipelines, and the bottom panel shows the normalized periodograms. In both periodogram panels, the black lines show the final Oxford period, the blue lines show the location of the final CfA period, and the red lines show the location of the period from the Barnes paper.

K2 data. Reliability helps establish amplitude and period thresholds for detecting true signals in M67. We can probably trust detections in a measured period and amplitude bin with reliability of 80% or higher, suggesting that one or fewer out of every five measurements is likely to be unreliable. However, we need to take completeness into account: we simply cannot say anything about the true number of stars with periods and amplitudes corresponding to a given bin without it, because completeness tells us what fraction of potential detections we miss. For example, if we have three detections in a bin where the completeness was 50%, we know that there were approximately three other stars in that range that we missed.

It is important to recall that these sinusoidal injection tests represent a “best case” scenario: real rotation signals are non-sinusoidal and evolve over time, and both of these factors are likely to affect their detectability in a negative way. Therefore, although it is of course possible that better light curve extraction, detrending, and period search methods might be devised than those we have used here, it seems rather unlikely that the detectability of true rotation signals will improve much beyond what we have shown. In short, the *K2* Campaign 5 data may allow the measurement of some rotation periods for main sequence stars in M67, but only if they are relatively active and display amplitudes somewhat larger than the active Sun.

6. Conclusions and Future Work

The open cluster M67, recently observed in Campaign 5 of the NASA *K2* mission, offers a unique chance to measure

rotation periods for solar-age stars. This means that we have the opportunity to fill in a much-needed gap in the calibration of the age-mass-rotation period relationship that forms the basis for the field of gyrochronology and serves as an age-dating method for main sequence stars. However, our physical understanding behind the empirical observations that have driven gyrochronology is still evolving, and certain studies suggest inconsistencies between the ages that we would predict via gyrochronology and the results of other methods, such as asteroseismology, for older stars. M67, therefore, provides a testing ground both for gyrochronology and new theories exploring the possible weakening of the magnetic fields that induce angular momentum loss in stars. Because of the immense scientific potential of M67, we want to ensure that we have a full understanding of the limits of the data we are using to measure the rotation periods. Though *K2* does have relatively high precision, especially compared to ground-based observations, the data still suffer from stubborn systematic features, and the ~ 75 days observation window means that it will be challenging to identify periods of 25 days or longer, which we are expecting for a cluster the approximate age of the Sun. In addition, the crowded field will make the task even more difficult. It is the goal of this study to understand how these challenges manifest themselves, in order to ultimately acquire a set of M67 rotation periods for which we have confidence.

We devised a series of sinusoidal injection tests with real Campaign 5 data in order to determine the best-case scenario threshold limits for *K2* M67 data. We used a subset of the

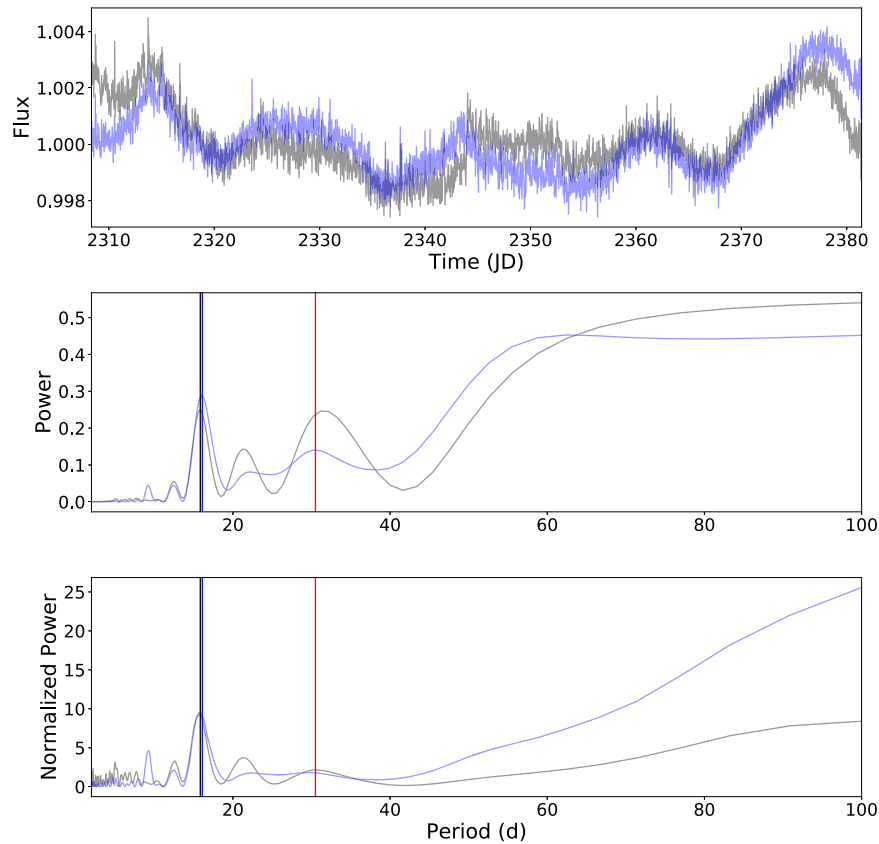


Figure 21. Same as Figure 20 except for EPIC 211411621.

SAP light curves processed via the *Kepler* pipeline that fell on the spacecraft’s module 6 CCD (which encompassed M67), M67 members contained in the superstamp, and a small set of hot A and F stars scattered throughout the Campaign 5 field of view as the three test samples for the injection tests. Into each raw light curve of each sample, we injected sinusoids with six different amplitudes, ranging from 0.05% to 3.0%, and seven periods, ranging from 5 to 35 days. We processed the injected light curves using two different methods based on existing literature, within this paper known as the “Oxford” and “CfA” pipelines. We then defined a detection threshold and ran a Lomb–Scargle periodogram on the injected light curves and their non-injected counterparts. We normalized the periodograms of both the injected and non-injected light curves by dividing out the median periodograms of the latter. Those non-injected light curves whose normalized periodograms met the detection threshold were marked as “variable,” and we removed the corresponding stars from the injected samples. Finally, we analyzed the results of the period search on the remaining sets of injected light curves in terms of completeness and reliability. Completeness describes how sensitive our methods are in recovering the injected signals, while reliability quantifies how trustworthy a detected period is within a measured amplitude range.

The results of the injection tests shed light on the nature of the K2 M67 data. The hot star samples from the Oxford and CfA pipelines highlighted the presence of non-astrophysical trends in the data with power at ~ 25 days and longer, a problematic feature—especially given the periods we expect

in M67—that necessitated the periodogram normalization in the period search. In general, for the hot star, SAP, and SS samples from both pipelines, completeness diminished as the period increased and amplitude decreased. The hot star completeness was higher than the SAP and SS samples due to the brightness of the stars in the sample and the lack of rotational modulation beyond about 10 days for this particular data set. However, for the SAP and SS samples from both pipelines, the completeness falls around or below about 50% at injected periods of 20 days and amplitudes of 0.50%. Crucially, at the solar case—where the injected period is 25 days and the amplitude 0.10%—the maximum completeness is only around 15% for the SAP and SS data sets. Despite the generally low sensitivity, the reliability is typically very high, consistently at values greater than 90%, except in the instance of very long periods (~ 35 days) and a couple cases at detected periods of around 15 days with moderately low measured amplitudes (0.075–0.25%). This means that, while it is very difficult to detect the kind of periods we expect to see in M67 given its age, if we do get a detection, we can generally trust the result.

The CfA completeness is generally greater in both the SAP and SS samples. Except at long periods and the ~ 15 days range, the reliability for the SAP samples from both pipelines are comparable, but the Oxford reliability is slightly higher for the SS, despite amplitude suppression from the PCA. There is a slight trade-off between the two pipelines, and while the CfA performs marginally better overall, both are viable options.

It is important to understand how this study alters our understanding of both future and previously published studies

regarding rotation periods in M67 from *K2* data, namely Barnes et al. (2016) and Gonzalez (2016a). Overall, we urge caution when using the periods from Barnes et al. (2016) and Gonzalez (2016a) because we cannot reproduce the same results here without lowering the detection threshold and sacrificing reliability. However, we can use the injection tests as a basis for the re-evaluation of M67 periods. With completeness, we can quantify just how difficult is to find even the best-case, long-period rotation signals in M67 data, or how many detections we may miss, while reliability gives us confidence in our measurement. To maintain some flexibility while still trusting the result, we decide on a reliability threshold of 80% for a given measured period and amplitude range.

While this study has given us important insights into the *K2* M67 data, we have to remember that the sinusoidal injections offer a best-case scenario. The rotational modulation in most stars does not usually take the form of a simple sinusoid. We ultimately want to do another, smaller round of injection tests using more realistic signals, such as those generated from the star spot models of Aigrain et al. (2015b). This necessarily means that the Lomb–Scargle periodogram may not end up being the optimum rotation detection method. As a result, we will use the star spot model tests to compare the Lomb–Scargle to the autocorrelation function and a Gaussian process. Following this, we hope to present our own rotation periods for M67 from Campaign 5 data, informed by both the sinusoidal and star spot model injection tests, along with a comparison to previously published results and existing theory. While we have shown that it is difficult to find low-amplitude, 25 day periods in the *K2* M67 data from Campaign 5, we should still be able to infer important information regarding the true rotation periods in M67 in a manner similar to the planet occurrence studies conducted by Howard et al. (2012) and Petigura et al. (2013). In addition, Campaign 16, which has just been released at the time of this writing, will greatly enhance our confidence in any rotation periods measured from Campaign 5 data, as the extra observation time will allow us to improve completeness without sacrificing reliability.

We conducted this study specifically in the context of *K2* Campaign 5 M67 light curves, but the caution we have urged based on our conclusions can reasonably be extended to rotation studies with other *K2* campaigns and the upcoming first data release from *TESS*. We have shown that long-term, systematic trends still exist in *K2* data, which could skew low-amplitude period detections of about 25 days and longer, and these trends are likely to be present in other campaigns, especially in the crowded fields of clusters. Finally, while *TESS* will survey nearly the entire sky, most of the coverage will be along the Ecliptic poles as opposed to the Ecliptic itself, where observations will only last about 27 days (Ricker et al. 2015). While this will complement *K2* data, the short observation time and large pixel sizes ($15 \mu\text{m} \times 15 \mu\text{m}$; Ricker et al. 2015) will likely cause the data to suffer from problems similar to those we have seen in this study.

R.E. would like to thank the Rhodes Trust and the U.S. Air Force for helping finance this work. The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Air Force, the Department of Defense, or the U.S. Government. Financial support for this work also comes from the Science and Technology Facilities Council (STFC) consolidated grants

ST/H002456/1 and ST/N000919/1. This work was performed in part under contract with the California Institute of Technology/Jet Propulsion Laboratory, funded by NASA through the Sagan Fellowship Program executed by the NASA Exoplanet Science Institute. Some/all of the data presented in this paper were obtained from the Mikulski Archive for Space Telescopes (MAST). STScI is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS5-26555.

Appendix A Completeness and Reliability Tables

The complete completeness and reliability results for the hot star, SAP, and SS samples from the Oxford and CfA pipelines are given here. Tables 1–6 show the completeness results, while Tables 7–12 show the reliability results; both groups are rounded to the nearest percent. In the reliability tables, the number of detections for each bin is given in brackets below the reliability statistic. Uncertainties are provided.

Appendix B Additional Experiments

Here, we present the results from additional experiments we conducted to further understand the benefits of normalizing the Lomb–Scargle periodogram light curves and to compare our results for the non-injected versions of the light curves with those published in Barnes et al. (2016) and Gonzalez (2016a). In Appendix B.1, Figure 22 shows the completeness and reliability values for the Oxford hot star and SAP data sets without normalization of the Lomb–Scargle periodograms, compared to the results from the normalized versions. In Appendix B.2, we show the measured period from the Oxford and CfA pipelines for the EPICs listed in Barnes et al. (2016) and the associated values of C needed to count them as detections.

B.1. Completeness and Reliability without Normalization

To see the effect of using the regular Lomb–Scargle periodogram (i.e., without normalization) on the completeness and reliability for the *K2* Campaign 5 M67 light curves, we computed the periods for the Oxford hot star and SAP injected data sets without using the normalized power. For a detection threshold, we followed the example of Nielsen et al. (2013) and set the threshold at four times the root mean square (rms) of the zero-mean time series. We removed the same stars marked as variables as in the main set of injection tests. Figure 22 shows the completeness and reliability results for the solar amplitude (0.1% for completeness, 0.07–0.25% for reliability), along with the number of detections with measured amplitudes ranging from 0.07% to 0.25% (i.e., similar to the numbers in brackets in Tables 7 through 12). The blue lines show the results with normalization, while the red lines show the results without normalization. Due to the lower detection threshold criteria and the lingering presence of the long-term trends seen in the median periodograms, the number of detections for the un-normalized samples increases significantly, compared to the normalized hot star and SAP samples. This allows for an increase in completeness, including at the solar case, which improved from 17% with the normalized periodograms to 39% for the regular Lomb–Scargle with the hot stars, and from 2% to 17% for the SAP sample. In addition, completeness begins to

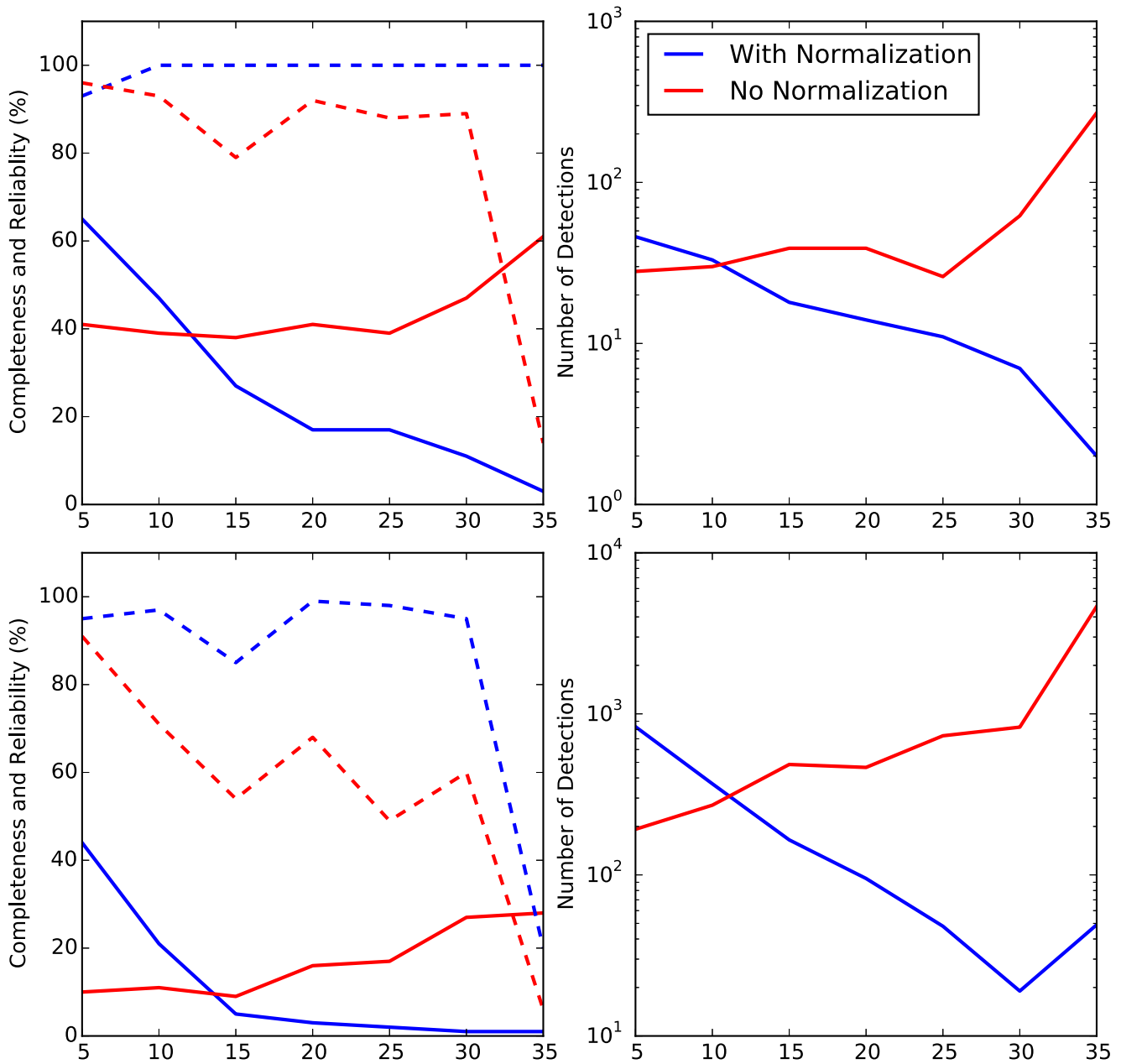


Figure 22. Comparison of completeness and reliability results for the Oxford hot star (top row) and SAP (bottom row) data sets with (blue) and without (red) normalization of the Lomb–Scargle periodogram for the solar amplitude (0.1% for completeness, 0.07–0.25% range for reliability). The left panels show the completeness (solid lines) and reliability (dashed lines) results, while the right panels show the number of detected periods with measured amplitudes of 0.07–0.25% (i.e., similar to the numbers in brackets in Tables 7 through 12).

increase with increasing period within an injected amplitude, starting around 25 days, countering the general trends of the normalized sample. However, as expected, the reliability at periods of 25 days and longer decreases drastically. There is also a disproportionate number of detections at these longer periods, showing just how much of an effect the lingering long-term systematic features have on periodic measurements.

B.2. Detection Thresholds and Periods for Barnes et al. (2016) Stars

Table 15 shows the minimum detection threshold factors for the periods found by the normalized Lomb–Scargle periodogram to count as detections using the light curves from the Oxford and CfA pipelines for the stars in Barnes et al. (2016). We present the

EPICs from Barnes et al. (2016), the published periods, the periods computed from the normalized periodograms of the corresponding non-injected Oxford and CfA pipeline light curves, the threshold factor C required to count the Oxford and CfA periods as detections, and the associated reliabilities. In most cases, the value of C had to be reduced to one or two in order for the measured period from the Oxford and CfA pipelines to be counted as a detection, which means the reliability falls considerably; the average Oxford reliability is 65%, and the average CfA reliability is 61%. In addition, the Oxford and CfA periods rarely match the published values (though in some cases they are harmonics), and they do not always match each other. This further illustrates how difficult it is to find accurate rotation periods in the K2 Campaign 5 M67 light curves.

Table 15
Evaluation of Barnes et al. (2016)

EPIC	Barnes Per (days)	Ox Per (days)	Ox C	Ox Rel (%)	CfA Per (days)	CfA C	CfA Rel (%)
211388204	31.8	2.2	1	6 ± 3.2	2.1	1	26 ± 8.0
211394185	30.4	15.4	4	98 ± 3.6	15.1	4	95 ± 3.3
211395620	30.7	28.5	2	98 ± 4.5	3.5	2	38 ± 3.3
211397319	25.1	7.9	2	96 ± 12.0	47.6	1	54 ± 16.0
211397512	34.5	16.4	4	98 ± 3.6	38.4	3	85 ± 5.7
211398025	28.8	22.2	2	100 ± 28.9	10.4	2	78 ± 4.5
211398541	30.3	5.3	16.9	2	77 ± 5.0
211399458	30.2	7.9	2	86 ± 4.5	4.32	1	17 ± 5.2
211399819	28.4	13.9	2	36 ± 2.4	2.1	1	17 ± 5.2
211400500	26.9	13.5	3	82 ± 6.6	2.0	2	38 ± 3.3
211406596	26.9	31.2	2	85 ± 6.9	2.4	1	26 ± 8.0
211410757	18.9	18.8	9.4	3	93 ± 5.6
211411477	31.2	5.7	1	6 ± 3.2	5.7	1	26 ± 8.0
211411621	30.5	15.8	4	85 ± 7.8	16.1	4	91 ± 4.8
211413212	24.4	5.3	1	6 ± 3.2	7.5	1	61 ± 14.7
211413961	31.4	2.9	1	6 ± 3.2	15.4	3	88 ± 4.5
211414799	18.1	4.6	3	100 ± 20.9	9.0	4	98 ± 4.2
211423010	24.9	20.8	2	92 ± 7.8	22.2	4	98 ± 5.5
211428580	26.9	5.1	1	6 ± 3.2	2.4	1	26 ± 8.0
211430274	31.1	23.2	2	89 ± 11.2	27.0	2	93 ± 6.8

Notes. The columns, from left to right, are EPIC, the associated period reported from Barnes et al. (2016), the period from the Oxford pipeline, the Oxford threshold factor C that results in a detection, the corresponding value for reliability, the period from the CfA pipeline, the CfA detection value for C , and the CfA reliability.

ORCID iDs

Suzanne Aigrain  <https://orcid.org/0000-0003-1453-0574>

Andrew Vanderburg  <https://orcid.org/0000-0001-7246-5438>

Robert Mathieu  <https://orcid.org/0000-0002-7130-2757>

References

- Aigrain, S., Hodgkin, S. T., Irwin, M. J., Lewis, J. R., & Roberts, S. J. 2015a, *MNRAS*, **447**, 2880
- Aigrain, S., Llama, J., Ceillier, T., et al. 2015b, *MNRAS*, **450**, 3211
- Aigrain, S., Parviainen, H., & Pope, B. J. S. 2016, *MNRAS*, **459**, 2408
- Angus, R., Aigrain, S., Foreman-Mackey, D., & McQuillan, A. 2015, *MNRAS*, **450**, 1787
- Angus, R., Morton, T., Aigrain, S., Foreman-Mackey, D., & Rajpaul, V. 2018, *MNRAS*, **474**, 2094
- Armstrong, D. J., Kirk, J., Lam, K. W. F., et al. 2016, *MNRAS*, **456**, 2260
- Armstrong, D. J., Osborn, H. P., Brown, D. J. A., et al. 2014, arXiv:1411.6830
- Balaguer-Núñez, L., Galadí-Enríquez, D., & Jordi, C. 2007, *A&A*, **470**, 585
- Baliunas, S. L., Donahue, R. A., Soon, W., & Henry, G. W. 1998, in ASP Conf. Ser. 154, Cool Stars, Stellar Systems, and the Sun, ed. R. A. Donahue & J. A. Bookbinder (San Francisco, CA: ASP), **153**
- Barnes, S. A. 2003, *ApJ*, **586**, 464
- Barnes, S. A. 2007, *ApJ*, **669**, 1167
- Barnes, S. A. 2010, *ApJ*, **722**, 222
- Barnes, S. A., Weingrill, J., Fritzewski, D., Strassmeier, K. G., & Platais, I. 2016, *ApJ*, **823**, 16
- Carraro, G., Chiosi, C., Bressan, A., & Bertelli, G. 1994, *A&AS*, **103**, 375
- Christiansen, J. L., Jenkins, J. M., Caldwell, D. A., et al. 2012, *PASP*, **124**, 1279
- Cody, A. M., Barentsen, G., Hedges, C., Gully-Santiago, M., & Cardoso, J. V. d. M. 2018, *RNAAS*, **2**, 25
- Collier Cameron, A., Davidson, V. A., Hebb, L., et al. 2009, *MNRAS*, **400**, 451
- Delorme, P., Collier Cameron, A., Hebb, L., et al. 2011, *MNRAS*, **413**, 2218
- Douglas, S. T., Agüeros, M. A., Covey, K. R., et al. 2016, *ApJ*, **822**, 47
- Fan, X., Burstein, D., Chen, J.-S., et al. 1996, *AJ*, **112**, 628
- García, R. A., Ceillier, T., Salabert, D., et al. 2014, *A&A*, **572**, A34
- Geller, A. M., Latham, D. W., & Mathieu, R. D. 2015, *AJ*, **150**, 97
- Gilliland, R. L., Chaplin, W. J., Jenkins, J. M., Ramsey, L. W., & Smith, J. C. 2015, *AJ*, **150**, 133
- Gonzalez, G. 2016a, *MNRAS*, **463**, 3513
- Gonzalez, G. 2016b, *MNRAS*, **459**, 1060
- Hartman, J. D., Bakos, G. Á., Kovács, G., & Noyes, R. W. 2010, *MNRAS*, in press (arXiv:1006.0950)
- Hartman, J. D., Gaudi, B. S., Holman, M. J., et al. 2009, *ApJ*, **695**, 336
- Hempelmann, A., Schmitt, J. H. M. M., & Stępień, K. 1996, *A&A*, **305**, 284
- Hoaglin, D. C., Mosteller, F., & Tukey, J. W. 1983, Understanding Robust and Exploratory Data Analysis (New York: Wiley), <https://www.wiley.com/eng/Understanding+Robust+and+Exploratory+Data+Analysis-p-9780471384915>
- Home, J. H., & Baliunas, S. L. 1986, *ApJ*, **302**, 757
- Howard, A. W., Marcy, G. W., Bryson, S. T., et al. 2012, *ApJS*, **201**, 15
- Howell, S. B., Sobek, C., Haas, M., et al. 2014, *PASP*, **126**, 398
- Irwin, J., Aigrain, S., Bouvier, J., et al. 2009, *MNRAS*, **392**, 1456
- Irwin, J., Aigrain, S., Hodgkin, S., et al. 2006, *MNRAS*, **370**, 954
- Irwin, J., Hodgkin, S., Aigrain, S., et al. 2007, *MNRAS*, **377**, 741
- Jacobson, H. R., Pilachowski, C. A., & Friel, E. D. 2011, *AJ*, **142**, 59
- James, D. J., Barnes, S. A., Meibom, S., et al. 2010, *A&A*, **515**, A100
- Kawaler, S. D. 1988, *ApJ*, **333**, 236
- Kawaler, S. D. 1989, *ApJL*, **343**, L65
- Lang, D., Hogg, D. W., Mierle, K., Blanton, M., & Roweis, S. 2010, *AJ*, **137**, 1782
- Libralato, M., Bedin, L. R., Nardiello, D., & Piotto, G. 2016, *MNRAS*, **456**, 1137
- Loktin, A. V. 2005, *ARep*, **49**, 693
- Luger, R., Agol, E., Kruse, E., et al. 2016, *AJ*, **152**, 100
- Luger, R., Kruse, E., Foreman-Mackey, D., Agol, E., & Saunders, N. 2017, *AJ*, submitted (arXiv:1702.05488)
- Mamajek, E. E., & Hillenbrand, L. A. 2008, *ApJ*, **687**, 1264
- McQuillan, A., Aigrain, S., & Mazeh, T. 2013a, *MNRAS*, **432**, 1203
- McQuillan, A., Mazeh, T., & Aigrain, S. 2013b, *ApJL*, **775**, L11
- McQuillan, A., Mazeh, T., & Aigrain, S. 2014, *ApJS*, **211**, 24
- Meibom, S., Barnes, S. A., Latham, D. W., et al. 2011, *ApJL*, **733**, L9
- Meibom, S., Barnes, S. A., Platais, I., et al. 2015, *Natur*, **517**, 589
- Meibom, S., Mathieu, R. D., & Stassun, K. G. 2009, *ApJ*, **695**, 679
- Meibom, S., Mathieu, R. D., Stassun, K. G., Liebesny, P., & Saar, S. H. 2011, *ApJ*, **733**, 115
- Nardiello, D., Libralato, M., Bedin, L. R., et al. 2016, *MNRAS*, **463**, 1831
- Nielsen, M. B., Gizon, L., Schunker, H., & Karoff, C. 2013, *A&A*, **557**, L10
- Pasquini, L., Biazzo, K., Bonifacio, P., Randich, S., & Bedin, L. R. 2008, *A&A*, **489**, 677
- Petigura, E. A., Howard, A. W., & Marcy, G. W. 2013, *PNAS*, **110**, 19273
- Radick, R. R., Lockwood, G. W., Skiff, B. A., & Baliunas, S. L. 1998, *ApJS*, **118**, 239
- Radick, R. R., Lockwood, G. W., Skiff, B. A., & Thompson, D. T. 1995, *ApJ*, **452**, 332

- Radick, R. R., Thompson, D. T., Lockwood, G. W., Duncan, D. K., & Baggett, W. E. 1987, [ApJ](#), **321**, 459
- Rebull, L. M., Stauffer, J. R., Bouvier, J., et al. 2016a, [AJ](#), **152**, 113
- Rebull, L. M., Stauffer, J. R., Bouvier, J., et al. 2016b, [AJ](#), **152**, 114
- Reinhold, T., & Gizon, L. 2015, [A&A](#), **583**, A65
- Reinhold, T., Reiners, A., & Basri, G. 2013, [A&A](#), **560**, A4
- Ricker, G. R., Winn, J. N., Vanderspek, R., et al. 2015, [JATIS](#), **1**, 014003
- Sanders, W. L. 1977, [A&AS](#), **27**, 89
- Sarajedini, A., Dotter, A., & Kirkpatrick, A. 2009, [ApJ](#), **698**, 1872
- Scargle, J. D. 1982, [ApJ](#), **263**, 835
- Schatzman, E. 1962, [AnAp](#), **25**, 18
- Skumanich, A. 1972, [ApJ](#), **171**, 565
- Smith, J. C., Stumpe, M. C., Cleve, J. E. V., et al. 2012, [PASP](#), **124**, 1000
- Stauffer, J., Rebull, L., Bouvier, J., et al. 2016, [AJ](#), **152**, 115
- Stumpe, M. C., Smith, J. C., Catanzarite, J. H., et al. 2014, [PASP](#), **126**, 100
- Stumpe, M. C., Smith, J. C., Cleve, J. E. V., et al. 2012, [PASP](#), **124**, 985
- Taylor, B. J. 2007, [AJ](#), **133**, 370
- Van Cleve, J. E., Howell, S. B., Smith, J. C., et al. 2016, [PASP](#), **128**, 075002
- Vanderburg, A., & Johnson, J. A. 2014, [PASP](#), **126**, 948
- Vanderburg, A., Latham, D. W., Buchhave, L. A., et al. 2016, [ApJS](#), **222**, 14
- van Saders, J. L., Ceillier, T., Metcalfe, T. S., et al. 2016, [Natur](#), **529**, 181
- Yadav, R. K. S., Bedin, L. R., Piotto, G., et al. 2008, [A&A](#), **484**, 609