The M33 Synoptic Stellar Survey. II. Mira Variables

Wenlong Yuan1, Shiyuan He2, Lucas M. Macri1, James Long2, and Jianhua Z. Huang2

1 George P. and Cynthia W. Mitchell Institute for Fundamental Physics & Astronomy, Department of Physics & Astronomy, Texas A&M University, College Station, TX 77843, USA; lmacri@tamu.edu
2 Department of Statistics, Texas A&M University, College Station, TX 77843, USA

Received 2016 October 19; revised 2017 February 23; accepted 2017 February 27; published 2017 March 22

Abstract

We present the discovery of 1847 Mira candidates in the Local Group galaxy M33 using a novel semi-parametric periodogram technique coupled with a random forest classifier. The algorithms were applied to \( \sim 2.4 \times 10^5 \) \( I \)-band light curves previously obtained by the M33 Synoptic Stellar Survey. We derive preliminary period–luminosity relations at optical, near-infrared, and mid-infrared wavelengths and compare them to the corresponding relations in the Large Magellanic Cloud. 

Key words: methods: statistical – stars: variables: general

Supporting material: machine-readable table

1. Introduction

Mira variables (Miras) are asymptotic giant branch (AGB) pulsating stars that exhibit large cyclical variations in flux at optical wavelengths, typically with periods spanning 100–700 day but in extreme cases going beyond 1500 day. The “canonical” empirical classification requires \( \Delta V > 2.5 \) mag within a pulsation period and spectroscopic confirmation (Kholopov et al. 1985). Recent surveys for these variables (such as Soszyński et al. 2009) have adopted \( \Delta I > 0.8 \) mag as the only requirement for classification, since spectroscopic follow-up of very large samples is not currently feasible. Longer-term variations in the mean flux level of each cycle are typical (Mattei 1997; Whitelock et al. 1997), and visual light curves exhibit a wide range of shapes; Ludendorff (1928) classified Miras into three classes and 10 subclasses based on this attribute.

Since the progenitors of Miras are relatively low-mass stars, they are ubiquitous and present in all types of galaxies. Thousands of Mira candidates have been discovered in the Milky Way and the Magellanic Clouds based on photometry from the Optical Gravitational Lensing Experiment (OGLE; Udalski et al. 1992) and MACHO (Alcock et al. 1993) projects. The Mira period–luminosity relations (PLRs) were initially studied by Gerasimovic (1928); Glass & Lloyd Evans (1981) found the first evidence of a near-infrared (NIR) PLR for Miras based on a small sample of variables in the Large Magellanic Cloud (LMC). Using the MACHO database, Wood et al. (1999) were the first to identify multiple PLRs for AGB stars and to confirm the nature of Miras as radial fundamental mode pulsators. Also using MACHO periods, Glass & Lloyd Evans (2003) determined that the Mira \( K \)-band PLR exhibits a relatively small scatter of \( \sigma \sim 0.13 \) mag, while Whitelock et al. (2008) found similar dispersions for \( K \)-band PLRs separated into O- and C-rich subtypes \( (\sigma = 0.14 \) and \( 0.15 \) mag, respectively). These values are comparable to the intrinsic dispersion of the Cepheid PLR in the same bandpass \( (\sigma = 0.09 \) mag; Macri et al. 2015). Soszyński et al. (2009, 2011, 2013) characterized the NIR Mira PLRs in the LMC, the Small Magellanic Cloud, and the Galactic bulge, respectively, using OGLE and Two Micron All Sky Survey (2MASS) photometry. In just a few years, the Large Synoptic Survey Telescope (LSST) will begin to obtain frequent images of dozens of nearby galaxies that will have the necessary depth to enable the discovery of Miras and the determination of distances to these systems.

In this work, we report the results of a search for Mira candidates in M33 using \( I \)-band observations spanning nearly a decade that were obtained by the DIRECT project (Macri et al. 2001, hereafter M01) and by Pellerin & Macri (2011, hereafter PM11). Traditional periodogram methods, such as Lomb–Scargle (Lomb 1976; Scargle 1982), are not optimal for this search due to relatively sparse temporal sampling, large gaps between observing seasons, and the expected long-term variations in Mira light curves. We developed a novel semi-parametric periodogram technique (He et al. 2016, hereafter H16) based on the Gaussian process method that contains a data-driven component in the model light curve to account for deviations from strict periodicity and gives an overall better performance. We coupled this algorithm to random forest (RF) classifiers, training and testing them extensively on simulated light curves.

The paper is organized as follows. Section 2 introduces the observations and data reduction, Section 3 gives details of the light-curve simulations, Section 4 discusses the methodology used to search for Mira candidates and estimate their periods, and Section 5 presents our results, which include a comparison of the RF classification with other techniques and preliminary PLRs for O-rich Mira candidates.

2. Observations and Data Reduction

We based our search on the observations of M33 obtained by M01 and PM11. These surveys covered most of the disk of this galaxy with a combined baseline of nearly a decade (1996 September to 1999 November for M01; 2002 August to 2006 August for PM11) mainly using the Fred L. Whipple Observatory (FLWO) 1.2 m and the WIYN observatory 3.5 m telescopes with a variety of cameras (see the respective publications for details). While images were obtained in multiple bandpasses (\( BVI \)), our analysis is only based on the \( I \)-band time-series photometry because Mira candidates fall below the detection limit in the bluer bands. Given that both studies had to rely on multiple pointings to cover the area of interest and not all locations were observed on a given night, the sampling pattern varies considerably across the disk.
Figure 1. Top: cadence of M33 observations in I by M01 and PM11. The grayscale levels are linearly proportional to the number of measurements per square arcminute of each epoch. Bottom left: expanded view of the cadence for seasons 1–4. Bottom right: histogram of measurements for stars with \( N > 10 \) and \( I < 21.45 \) mag.

Figure 1 shows the overall sequence of observations, of which only a subset will be applicable at a given position.

We performed new photometric measurements on the preprocessed images from PM11 to mitigate issues arising from geometric distortions and poor image registration at the corners of each field (which corresponds to a single telescope and camera). Unlike the approach of the previous work, we first analyzed the images of a given field and later combined the photometry for matching sources. We obtained aperture and point-spread function (PSF) photometry using the \textsc{DAOPHOT}, \textsc{ALLSTAR}, and \textsc{ALLFRAME} programs (Stetson 1987, 1994) in a quasi-automatic manner by integrating the tasks into an R script pipeline. We defined the PSF for each image using the top 50 bright and isolated stars and selected the one with the sharpest PSF from each image using the \textsc{DAOPHOT} programs (Stetson 1987, 1994) and \textsc{ALLFRAME}. We selected a larger number of secondary standards for image registration and to tie the photometric measurements. These were among the brightest few percent of all sources in a given field and had photometric uncertainties below 0.02 mag. We determined frame-to-frame zero-point offsets, computed mean instrumental magnitudes, and extracted light curves using \textsc{TRIAL} (Stetson 1996).

We obtained astrometric and photometric calibrations for each field using the catalog published by the Local Group Galaxies Survey (LGGS; Massey et al. 2006). We derived the astrometric solution for the reference frame of each field using \textsc{WCSTools} (Mink 1999). We matched the LGGS sources to the star list from the (now astrometrically calibrated) reference frame of each field and solved the transformation equation

\[
I_c = (1 + a) \cdot I_i + b,
\]

where \( I_c \) is the magnitude in the standard system (Kron–Cousins I for LGGS), \( I_i \) is the instrumental PSF magnitude of the reference frame of a given field, \( b \) is the zero-point offset, and \( a \) provides a simple correction for color terms and/or photometric biases due to crowding (given the considerable variation in stellar density across the disk and in image quality among the fields). We were not able to apply a traditional photometric transformation with zero-point and color terms because we only have single-band \( I \) photometry for the vast majority of the sources. We solved for the coefficients using the top 25% and 10% brightest stars in the fields imaged at WIYN and FLWO, respectively, applying an iterative outlier rejection of 3\( \sigma \) and 2.3\( \sigma \), respectively. The median value of \( a \) was 0.001, with 95% of the values falling between \(-0.024 \) and \(+0.015 \).

Given the significant overlap between the fields of M01 and PM11, most objects have multiple light-curve segments that were merged as follows. If two sources in different fields had coordinates that matched to better than 1″, and there were no other sources detected within 1″, then they were considered the same object. If there were neighboring sources within that radius, then the closest object with a magnitude difference less than 0.5 mag was selected (recall that each field was already transformed to the standard system prior to this step). We ensured that at most one source in one field could merge with one source in another field. We tested the photometric precision of the aforementioned steps by reidentifying the local standards of each field and examining the dispersion of the merged light curves relative to the dispersion of individual segments. Figure 2 shows that we reached a photometric precision of 0.03 mag.

We selected 239,907 light curves for the Mira search, rejecting any with fewer than 10 measurements or with mean magnitudes fainter than \( I = 21.45 \) mag. The first cut is based on extensive testing via the simulated light curves (Section 3) of our algorithm (Section 4); the procedure does not yield reliable periods for sparser samplings. The second cut is due to the large photometric uncertainties beyond that magnitude limit, which prevent the detection of light-curve variations of the expected amplitude.

### 3. Simulated M33 Light Curves

We simulated \( 10^5 \) light curves of Miras and the same number of semi-regular variables (SRVs) that accurately reproduce the photometric uncertainties and temporal sampling of the M33 data set. The simulated light curves were used to test our period determination algorithm and to train a classifier to identify Mira candidates. The methods used in the simulation are based on the ones we developed for H16 and rely on very high precision \( I \)-band light curves sampled at hundreds of epochs over 7.5 yr.
by phase III of the OGLE project (Udalski et al. 2008). We also generated an equal number of artificial light curves of “constant” stars, in order to properly balance the training data for the classifier.

3.1. Miras

The procedure used to fit templates to the OGLE-III Mira light curves is explained in detail in Section 4.3 of H16. Briefly, the light curve is decomposed into a mean value, a regular variation of period $P$, a low-frequency (long-term) trend, and a high-frequency/small-scale term. The latter three components are modeled by a Gaussian process with different kernels. The maximum likelihood method is used to obtain the model parameters as a function of the trial value of $k$ for the classiﬁer. We derived the completeness function of the M33 photometry by ﬁtting the observed luminosity function with an exponential, and we applied it to the luminosity function of the OGLE LMC Mira candidates after offsetting the latter by the difference in apparent distance moduli between the two galaxies. We then randomly selected simulated Mira light curves such that we reproduced the observed luminosity function of the M33 photometry.

Figure 3 shows a representative example of a simulated Mira light curve that mimics the cadence and photometric precision of the actual M33 data, while Figure 4 shows the completeness function of the M33 photometry.

3.2. SRVs and “Constant” Stars

The light curves of SRVs share some similarities with those of Miras (cyclic variations), although they tend to be more chaotic and less periodic and usually exhibit smaller amplitudes. Nevertheless, since they outnumber Miras six to one in the catalog of Soszyński et al. (2009), ignoring them could signiﬁcantly bias our classiﬁer. Hence, we included simulated SRV light curves in the training data. We obtained templates by applying a smoothing spline to the OGLE observations and generated artiﬁcial light curves by following the same procedure as for Miras (sampling based on the M33 observing patterns, random shifts of the starting point, convolution with the M33 completeness function, and addition of photometric noise).

Lastly, in order to balance the various types of objects that are used to train the classiﬁer, we simulated the light curves of “constant” stars by randomly shufﬂing the observation times of all light curves in our data set while keeping the original magnitudes and uncertainties. The shufﬂing removes any potential periodicity in the original data and allows the generation of multiple artiﬁcial light curves from the same object.

4. Semi-Parametric Model for Identiﬁcation and Period Determination of Miras

Given the stochastic variations exhibited by Mira light curves at optical wavelengths, traditional algorithms become...
less efficient at discovering these objects and obtaining reliable periods in the limit of the sparsely sampled observations. We showed in H16 that our semi-parametric model provides overall improvement of period recovery in this regime. We applied this model to the M33 observations and coupled it to an RF classifier in order to identify Mira candidates.

We refer interested readers to Section 3 of H16 for a detailed description of our semi-parametric model and its performance, which are only briefly summarized here. The model is based on Gaussian process regression, which has been previously applied to astronomical time-series observations. For example, Faraway et al. (2016) modeled the light curves of several types of transient event from the Catalina Real-Time Transient Survey with a squared exponential kernel, while Aigrain et al. (2016) applied this technique to Kepler data to correct systematic trends in their photometry.

The semi-parametric model we used is a simplified version of the one described in Section 3.1 to account for the quality of the M33 data. Given a set of measurements over $n$ epochs, $\{ (t_i, m_i, \sigma_i) \}_{i=1}^{n}$, with $t$, $m$, and $\sigma$ representing time, magnitude, and measurement uncertainty, respectively, the model is

$$m_i = m + \beta_1 \cos(2\pi f t_i) + \beta_2 \sin(2\pi f t_i) + h(t_i) + \sigma_i \epsilon_i,$$

where $f$ is the frequency (day$^{-1}$), $h(t)$ is modeled by a Gaussian process with the squared exponential kernel

$$k(t_1, t_2) = \theta_1^2 \exp(- (t_1 - t_2)^2 / 2\theta_2^2),$$

and the amplitude of the periodic component is $A_p = 2(\beta_1^2 + \beta_2^2)^{1/2}$. The parameters $m$, $\beta_1$, and $\beta_2$ are assumed to follow Gaussian priors and integrated out of the likelihood function when estimating other parameters. Optimization is performed over hyper-parameters $\theta_1$ and $\theta_2$ for each trial value of $f$, ranging from $5 \times 10^{-5}$ to $10^{-2}$ every $10^{-3}$. The log-likelihood function $Q$ (the “frequency spectrum”) is evaluated at each trial frequency (see Equation (10) in H16).

We applied the model to all simulated and real light curves and adopted the highest peak in the frequency spectrum ($f_1$) as the most likely frequency. We found that the true period was successfully recovered (with a tolerance of $|f - f_{\text{true}}| < 2.7 \times 10^{-4}$, as defined in H16) for 69.4% of all simulated Mira light curves.

We estimated the period uncertainties for all light curves using a nonparametric bootstrap approach followed by error

---

### Table 1

| Feature | Description | Src$^a$ | Rank$^b$ |
|---------|-------------|--------|---------|
| $\Delta Q$ | Difference of log likelihoods $Q_1 - Q_2$ | F | 1 12 |
| $\sigma(R_q)/\sigma(m)$ | Ratio of standard deviations (see below) | L | 2 9 |
| $A_{0.9}$ | Light-curve range from 10th to 90th percentile | L | 3 5 |
| $\Delta Q_{12}$ | Difference in log likelihood between highest and second peak | F | 4 11 |
| $A$ | Light-curve range | L | 5 7 |
| $A_p$ | Amplitude of periodic component | M | 6 2 |
| $\log_2 \theta_2$ | Log of hyper-parameter $\theta_2$ | M | 7 6 |
| $\sigma(\mathbf{m})$ | Standard deviation of residuals about $\mathbf{m}$ | L | 8 4 |
| $f_1$ | Best-fit frequency | M | 9 1 |
| $\sigma(R_q)$ | Standard deviation of residuals from piecewise quadratic fits | M | 10 13 |
| $Q_1$ | Best-fit log likelihood | F | 11 15 |
| $\mu$ | Unweighted mean magnitude | L | 12 3 |
| $Q_b$ | Baseline value of frequency spectrum (10th percentile) | F | 13 14 |
| $\sigma(R_{\text{model}})$ | Standard deviation of best-fit model residuals | L | 14 10 |
| $N$ | Number of measurements | L | 15 18 |
| $\theta_1$ | Hyper-parameter $\theta_1$ | M | 16 8 |
| $\sigma(\beta_1)$ | Posterior uncertainty of $\beta_1$ | M | 17 17 |
| $\sigma(\beta_2)$ | Posterior uncertainty of $\beta_2$ | M | 18 16 |

**Note.**

$^a$ Source of parameter = F—frequency spectrum, L—light curve, M—model.

$^b$ Rank in importance for classification as Mira or discrimination between C- and O-rich subtypes. Rank is determined by the mean decrease in the Gini index.

---

Figure 5. Examples of frequency spectra (top) and corresponding light curves (bottom) for a simulated Mira (left), SRV (middle), and constant star (right). The blue dashed lines and arrows indicate some of the quantities used as classification features.
scaling, as follows. First, we resampled the measurements with replacement and derived new values of \( f_1 \), repeating this procedure 500 times per variable. We used the standard deviation of the results for each object as an initial estimate of the period uncertainty. Next, we carried out the same procedure on the simulated Miras (with 30 iterations per light curve) and calculated \( \Delta P = (P_1 - P_2) / \sigma(P_1) \), where \( P_1 \) and \( P_2 \) are the input and recovered periods and \( \sigma(P_1) \) is the bootstrap-based uncertainty for the latter. Restricting our analysis to the uncertainty estimates of the Mira candidates. This required a rescaling factor of 2.33, which was then applied to the bootstrap-based uncertainty estimates of the Mira candidates.

5. Results

5.1. RF Classification of Miras

A machine-learning technique, RF has already proven to be effective in classifying different classes of variable stars (Dubath et al. 2011; Richards et al. 2011). We built an RF classifier based on the model parameters, the features of the frequency spectra, and information obtained from the simulated light curves, as detailed below. Once trained on the simulated data, it was applied to the M33 observations to select Mira candidates. Our choice of RF is supported by a comparative study reported in Section 5.2, where it is shown to outperform several state-of-the-art classifiers on simulated data. Figure 5 shows the frequency spectra for a representative artificial light curve for each of the three classes (Mira, SRV, and constant star). The frequency spectra of SRVs and constant stars are usually quite different due to their lack of periodicity, which indicates that the shape of this function can be used to identify Mira candidates.

We extracted seven features from the best-fit model parameters, four from the frequency spectra, and seven from the light curves. Table 1 provides a summary of all of the features and their rank in terms of importance for separating Mira candidates from other stars and for separating the former into likely C-rich or likely O-rich. The upper left panel of Figure 5 shows a graphical representation of the features that were extracted from the frequency spectra. Figure 6 shows two features we extracted from the light curves: the standard deviation of the residuals from piecewise quadratic fits (\( \sigma(R_0) \)) and its ratio to the scatter about the overall mean value (\( \sigma(M) \)). This ratio is significantly smaller for Miras than for any of the other classes.

We trained the RF classifier by building 400 decision trees with simulated light curves, each composed of \( 1.2 \times 10^5 \) non-Miras and \( 2.5 \times 10^3 \) Miras. This ratio was chosen to match the estimated fraction of Mira candidates in the actual data, derived from the ratio of Miras to other stars in the OGLE catalog (after applying the M33 completeness function). We then applied the trained classifier to the actual M33 data and obtained the voted probabilities for each star being a Mira (\( P_M \)). Figure 7 shows the resulting histogram. Based on a five-fold cross-validation on the simulated light curves, the Mira recovery rate at \( P_M = 0.5 \) is 75.4%, and the impurity is 0.7%. There are 5480 objects with probabilities above this value, 5145 of which have \( A_P > 0.6 \) mag and were selected for further study. Figure 8 shows three representative light curves for Mira candidates with different values of \( P_M \). The full set of light curves is available in the online edition of this article. The Mira subtype was tentatively inferred by using another RF classifier trained on the same features that yielded the probability of each candidate being O-rich (\( P_0 \)). Using the features of Mira candidates in the LMC bar to classify variables in the inner disk of M33 should be a robust approach, given the similar chemical abundances of both regions (Romaniello et al. 2008;
Figure 9. RF classification of Mira candidates into O-rich (blue) or C-rich (red), plotted as a function of \( P \) and \( A_P \). Left: LMC variables classified by Soszyński et al. (2009). Middle: simulated M33 variables based on the LMC sample but accounting for the shallower depth in absolute magnitude of our survey. Right: Mira candidates in M33 from this work.

Figure 10. Illustration of how multiple attributes help to discriminate O-rich from C-rich Mira candidates. Left: same as right panel of Figure 9 but indicating the area of interest where both subtypes overlap. Middle and right: separation of candidates on other two-dimensional slices of the RF parameters.

Bresolin (2011). Figure 9 shows the separation between subtypes based on the features with the highest rank in terms of discrimination: \( P \) and \( A_P \). The difference in the distribution of variables between the left and middle panels is due to the shallower depth and sparser sampling of the M33 survey relative to the OGLE coverage of the LMC. We caution that this is a limited two-dimensional view of a classification process that is based on 18 features. Figure 10 attempts to provide additional insight into the RF classification process by plotting the distribution of a subsample of candidates in other two-dimensional slices of parameter space. Based on a five-fold cross-validation on the simulated light curves, the O-rich recovery rate at \( P_O = 0.5 \) is 91.4% and the impurity is 12.8%, while the corresponding values for C-rich variables are 82.3% and 12.1%.

5.2. Comparison with Other Classification Methods

Although we chose RF as our classifier, it is insightful to compare its performance against other popular classifiers. We selected three state-of-the-art classifiers: sparse linear discriminant analysis with \( \ell_1 \) penalty (LDA), sparse logistic regression with \( \ell_1 \) penalty (SLR), and a \( \nu \)-classification support vector machine with radial basis kernel (SVM). We used the same input features discussed in Section 5.1, normalized to zero mean and unit variance.

First, we considered the classification task of Mira versus non-Mira. The comparison was in terms of the receiver operating characteristic (ROC) and its summary statistic, area under the curve (AUC). They were computed via repeated splitting of the simulated data set. In each instance, \( 10^5 \) Mira and non-Mira light curves (3.7% of the total) were sampled without replacement to serve as training data, while the rest served as test data. The procedure was repeated 200 times, and the final prediction for each light curve was calculated from the averaged probability across all iterations. The resulting ROC curves, shown in the left panel of Figure 11, are nearly identical, with AUC values of 0.984, 0.979, 0.975, and 0.976 for RF, LDA, SLR, and SVM, respectively.

We carried out a similar comparison for the task of classifying Miras into C-rich versus O-rich, with ROC curves plotted in the right panel of Figure 11. In this case, RF is significantly superior to the other methods, with AUCs of 0.912, 0.793, 0.787, and 0.801 for RF, LDA, SLR, and SVM, respectively.

5.3. Mira Candidates and PLRs

We examined the distribution of the best-fit periods for the selected objects and found a large peak at \( P \sim 340 \) day that is not seen in the LMC samples. Visual examination of the light curves in this period bin showed that they exhibit a significant change in the mean magnitude of segments obtained from
different telescopes. Further examination of the reference images for each field revealed that, for these objects, the poorer angular resolution of the FLWO images resulted in the blending of several sources (clearly separated in the WIYN frames). We visually inspected each light curve and its respective reference images and rejected affected objects. Our final sample consists of 1847 Mira candidates. Table 2 lists the following properties of the candidates (and their uncertainties, when applicable): IAU-standard ID, coordinates, most likely period, mean $I$-mag, amplitude of the periodic component ($A_P$), brightest magnitude of the best-fit model light curve ($I_m$), subtype (O/C), number of light-curve measurements ($N$), range of magnitudes ($A$), times ($\Delta t$) spanned by the light curve, and time of maximum light for the periodic component ($T_0$). We classified 1581 and 266 objects as O-rich and C-rich, respectively.

Figure 12 shows histograms of periods and amplitudes for both subtypes, while Figure 13 shows their deprojected...
Figure 14. PLRs in several bands for Mira candidates classified as O-rich in the LMC. The solid lines show the best-fit quadratic relations to the final LMC samples (large symbols) after iterative $3\sigma$ clipping of outliers (small dots). Dashed lines indicate the dispersion in the fits.
**Figure 15.** PLRs in several bands for Mira candidates classified as O-rich in M33. The solid lines show the LMC-based quadratic relations of Figure 14 shifted by the best-fit relative distance modulus in each band (including blending correction). Small dots indicate variables removed by iterative 3σ clipping. Dashed lines indicate the dispersion of the Gaussian component of the model.
galactocentric distribution. Our survey is limited to the innermost ~5 kpc of the galaxy, and within this limited area we see no statistically significant difference in the distribution of candidates by subtype or period.

We note that we recovered the only spectroscopically confirmed Mira in M33 (Barsukova et al. 2011), for which we found \( P = 578 \pm 32 \) day (in contrast to the previously published estimate of \( P = 665 \) day). Our classifiers yielded \( P_M = 0.89 \) and \( P_O = 0.9 \) for this object.

Figures 14 and 15 show preliminary PLRs for O-rich Mira candidates in the LMC and M33 at wavelengths ranging from 0.8 to 4.5 \( \mu \)m. We emphasize that the following is a simple analysis to demonstrate the validity of our methods for identifying, phasing, and classifying Mira candidates in M33. A complete analysis (including C-rich candidates) will be presented in a future paper.

The \( I_m \) magnitudes of the LMC Mira candidates were determined using the method described in Section 3, while the random-phase magnitudes at longer wavelengths were obtained from the SAGE catalog (Meixner et al. 2006). We chose to plot the \( I_m \) PLR to show a minimally biased comparison of the relations at this wavelength, since the \( V - I \) colors necessary to generate a Wesenheit-corrected mean-light \( I \)-band PLR are not available for M33. We show one example of a relation corrected for interstellar extinction for \( K_s \), using the formulation of Soszyński et al. (2009). We note that this formulation may not be appropriate to correct for the circumstellar dust that is especially prevalent among C-rich and long-period Miras (see Ita & Matsunaga 2011 for a thorough analysis of this issue). We solved for quadratic PLRs,

\[
m = a_0 + a_1(\log P - 2.3) + a_2(\log P - 2.3)^2,
\]

using an iterative 3\( \sigma \) clipping and removing the single largest outlier in each band until convergence. Table 3 summarizes the results of the fits.

The M33 sample was restricted to 1161 candidate variables with \( \Delta P/A < 1.1, \sigma(A_p)/A_p < 0.15, \sigma(P)/P < 0.1 \), and \( P < \Delta t \). These selection criteria were based on an examination of the input and recovered parameters for the simulated M33 Miras.

When the amplitude of the periodic component significantly exceeds the range of magnitudes spanned by the data and/or the best-fit period is longer than the time span of the light curve, the recovered parameters exhibit considerably larger scatter and the fraction of variables with successfully recovered periods (as defined in H16) is noticeably lower. The simulated O-rich Miras that met our selection criteria had input/output ratios of \( A_P \) of 1.02 \( \pm 0.21 \) versus 0.70 \( \pm 0.82 \) for the others. Likewise, the fraction of successfully recovered periods was 86% for the variables meeting the criteria versus only 45% for the others.

The \( I_m \) magnitudes of the M33 Mira candidates are from Table 2, while the random-phase magnitudes at longer wavelengths were taken from Javadi et al. (2015; for \( JHK_s \)) and Thompson et al. (2009; for 3.6 and 4.5 \( \mu \)m). We matched the catalogs using tolerances of 0.03 \( \Delta m \) and 0.01 \( \Delta \alpha \) and \( \Delta \delta \). The larger completeness of the M33 \( JHK_s \) sample and solved for \( a_0 \), applying an iterative 3\( \sigma \) clipping that removed the single largest outlier at a time. Once this procedure converged, we modeled the cumulative distribution of the PLR residuals as the combination of a Gaussian (to account for the finite width of the instability strip) plus an exponential distribution toward brighter values (to account for blends, which can only bias the residuals in one direction). The final values of \( a_0 \) listed in Table 3 include these “blending corrections,” which amount to \( \sim 0.09 \) and \( \sim 0.25 \) mag at near- and mid-infrared wavelengths, respectively. The larger contamination for the two longest bands is likely due to the significantly poorer angular resolution of the Spitzer images. The scatter in the M33 PLRs (after accounting for blended objects) compares favorably with the higher-quality LMC samples, and the mean (error-weighted) LMC-relative distance modulus of 6.31 \( \pm 0.11 \) mag is consistent with previous determinations (Bonanos et al. 2006; PM11).

### 6. Summary

We carried out a search for Mira variables in M33 using sparsely sampled \( I \)-band light curves. We determined periods using a novel semi-parametric Gaussian process model and

---

Table 3: Preliminary O-Rich Mira PLRs

| Galaxy | \( \lambda \) | \( a_0 \) | \( a_1 \) | \( a_2 \) | \( \sigma \) | \( N_i \) | \( N_f \) |
|--------|-------------|-------------|-------------|-------------|-------------|--------|--------|
| LMC    | \( I_m \)   | 13.66 \( \pm \) 0.02 | \( -2.12 \pm 0.14 \) | \( -5.01 \pm 0.37 \) | 0.33 | 427     | 399    |
|        | \( J \)     | 12.71 \( \pm \) 0.01 | \( -3.16 \pm 0.01 \) | \( -3.07 \pm 0.03 \) | 0.27 |         |        |
|        | \( H \)     | 11.87 \( \pm \) 0.01 | \( -3.42 \pm 0.01 \) | \( -2.85 \pm 0.03 \) | 0.27 |         |        |
|        | \( K_s \)   | 11.52 \( \pm \) 0.01 | \( -3.72 \pm 0.01 \) | \( -2.75 \pm 0.03 \) | 0.24 | 390     | 370    |
|        | \( WJK_s \)| 10.72 \( \pm \) 0.01 | \( -4.15 \pm 0.02 \) | \( -2.46 \pm 0.04 \) | 0.25 |         |        |
|        | 3.6         | 11.12 \( \pm \) 0.01 | \( -3.75 \pm 0.01 \) | \( -2.95 \pm 0.03 \) | 0.17 |         |        |
|        | 4.5         | 11.02 \( \pm \) 0.01 | \( -3.63 \pm 0.01 \) | \( -3.24 \pm 0.03 \) | 0.19 |         |        |
| M33    | \( I_m \)   | 19.82 \( \pm \) 0.01 | \( \ldots \)       | \( \ldots \)       | 0.40 | 1161    | 1125   |
|        | \( J \)     | 19.02 \( \pm \) 0.01 | \( \ldots \)       | \( \ldots \)       | 0.33 |         |        |
|        | \( H \)     | 18.29 \( \pm \) 0.01 | \( \ldots \)       | \( \ldots \)       | 0.34 |         |        |
|        | \( K_s \)   | 17.94 \( \pm \) 0.01 | \( \ldots \)       | \( \ldots \)       | 0.26 | 972     | 853    |
|        | \( WJK_s \)| 17.19 \( \pm \) 0.01 | \( \ldots \)       | \( \ldots \)       | 0.32 |         |        |
|        | 3.6         | 17.20 \( \pm \) 0.01 | \( \ldots \)       | \( \ldots \)       | 0.20 | 302     | 282    |
|        | 4.5         | 17.15 \( \pm \) 0.01 | \( \ldots \)       | \( \ldots \)       | 0.23 |         |        |

Note. Here \( N_i \) is the initial number of variables in the sample, \( N_f \) is the final number after iterative outlier rejection, and \( \sigma \) is the Gaussian width. The fits to the M33 samples include a blending correction and use fixed values of \( a_1 \) and \( a_2 \) determined from the LMC samples.
used the RF method to identify Mira candidates and classify them into C- or O-rich subtypes. We identified 1847 likely Mira candidates, most of them O-rich, that exhibit PLRs with dispersions comparable to those seen in the LMC.

W.Y. and L.M.M. acknowledge financial support from NSF grant AST-1211603 and from the Mitchell Institute for Fundamental Physics and Astronomy at Texas A&M University. S.H. was partially supported by the Texas A&M University–NSFC Joint Research Program. JZH was partially supported by NSF grant DMS-1208952. We thank Drs. A. Javadi & J. L. Prieto for providing the M33 photometry at near- and mid-infrared wavelengths, respectively. We thank the anonymous referee for helpful and constructive comments.

This publication has made use of the following resources:

1. Data products from the OGLE, conducted by the Astronomical Institute of the University of Warsaw at Las Campanas Observatory, operated by the Carnegie Institution for Science.
2. The VizieR catalog access tool and the cross-match service provided by the Centre de Données astronomiques de Strasbourg, France.
3. NASA’s Astrophysics Data System at the Harvard-Smithsonian Center for Astrophysics.
4. The NASA/IPAC Extragalactic Database (NED), which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with NASA.
5. Data products from the Two Micron All Sky Survey, which is a joint project of the University of Massachusetts and the Infrared Processing and Analysis Center at the California Institute of Technology, funded by NASA and the NSF.
6. The Texas A&M University Brazos HPC cluster.

Facilities: FLWO: 1.2 m, WIYN.
Software: DAOPHOT (Stetson 1987), ALLSTAR and ALLFRAME (Stetson 1994), TRIAL (Stetson 1996), WCSTools (Mink 1999).

References

Aigrain, S., Parviainen, H., & Pope, B. J. S. 2016, MNRAS, 459, 2408
Alcock, C., Allsman, R. A., Axelrod, T. S., et al. 1993, in ASP Conf. Ser. 43, Sky Surveys, Protostars to Protogalaxies, ed. B. T. Soifer (San Francisco, CA: ASP), 291
Barsukova, E. A., Goranskij, V. P., Hornoch, K., et al. 2011, MNRAS, 413, 1797
Bonanos, A. Z., Stanek, K. Z., Kudritzki, R. P., et al. 2006, ApJ, 652, 313
Bresolin, F. 2011, ApJ, 730, 129
Dubath, P., Rimoldini, L., Sèveges, M., et al. 2011, MNRAS, 414, 2602
Faraway, J., Mahabal, A., Sun, J., et al. 2016, Statistical Analysis and Data Mining, 9, 1
Gerasimovic, B. P. 1928, PNAS, 14, 963
Glass, J. S., & Lloyd Evans, T. 1981, Natur, 291, 303
Glass, J. S., & Lloyd Evans, T. 2003, MNRAS, 343, 67
He, S., Yuan, W., Huang, J. Z., Long, J., & Macri, L. M. 2016, AJ, 152, 164
Ita, Y., & Matsunaga, N. 2011, MNRAS, 412, 2345
Javadi, A., Saberi, M., van Loon, J. T., et al. 2015, MNRAS, 447, 3973
Kanbur, S. M., Hendry, M. A., & Clarke, D. 1997, MNRAS, 289, 428
Kholopov, P. N., Samus, N. N., Kazarovets, E. V., & Perova, N. B. 1985, IBVS, 2681, 1
Lomb, N. R. 1976, ApSS, 39, 447
Ludendorff, H. 1928, HDA, 6, 49
Macri, L. M., Ngeow, C.-C., Kanbur, S. M., Mahzooni, S., & Smitka, M. T. 2015, AJ, 149, 117
Macri, L. M., Stanek, K. Z., Sassdev, D. D., Krockenberger, M., & Kaluzny, J. 2001, AJ, 121, 861
Massey, P., Olsen, K. A. G., Hodge, P. W., et al. 2006, AJ, 131, 2478
Mattei, J. A. 1997, JAVSO, 25, 57
Meixner, M., Gordon, K. D., Indebetouw, R., et al. 2006, AJ, 132, 2268
Mink, D. J. 1999, in ASP Conf. Ser. 172, ADASS VIII, ed. D. M. Mehringer, R. L. Plante, & D. A. Roberts (San Francisco, CA: ASP), 498
Pellerin, A., & Macri, L. M. 2011, ApJS, 193, 26
Richards, J. W., Starr, D. L., Butler, N. R., et al. 2011, ApJ, 733, 10
Romaniello, M., Primas, F., Mottini, M., et al. 2008, A&A, 488, 731
Scargle, J. D. 1982, ApJ, 263, 835
Schlafly, E. F., & Finkbeiner, D. P. 2011, ApJ, 737, 103
Soszyński, I., Udalski, A., Szymański, M. K., et al. 2009, A&A, 59, 239
Soszyński, I., Udalski, A., Szymański, M. K., et al. 2011, A&A, 61, 217
Soszyński, I., Udalski, A., Szymański, M. K., et al. 2013, A&A, 63, 21
Stetson, P. B. 1987, PASP, 99, 191
Stetson, P. B. 1994, PASP, 106, 250
Stetson, P. B. 1996, PASP, 108, 851
Thompson, T. A., Prieto, J. L., Stanek, K. Z., et al. 2009, ApJ, 705, 1364
Udalski, A., Szymański, M., Kaluzny, J., Kubiak, M., & Mateo, M. 1992, A&A, 42, 253
Udalski, A., Szymański, M. K., Soszyński, I., & Poleski, R. 2008, A&A, 500, 69
Whitelock, P. A., Feast, M. W., Marang, F., & Overbeek, M. D. 1997, MNRAS, 288, 512
Whitelock, P. A., Feast, M. W., & van Leeuwen, F. 2008, MNRAS, 386, 313
Wood, P. R., Alcock, C., Allsman, R. A., et al. 1999, in IAU Symp. 191, Asymptotic Giant Branch Stars, ed. T. Le Bertre, A. Lebre, & C. Waelkens (San Francisco, CA: ASP), 151