The Astronomical Journal, 152:44 (27pp), 2016 August
doi:10.3847/0004-6256/152/2/44

REPEATABILITY AND ACCURACY OF EXOPLANET ECLIPSE DEPTHS MEASURED WITH POST-CRYOGENIC *SPITZER*

James G. Ingalls[1], J. E. Krick[1], S. J. Carey[1], John R. Stauffer[1], Patrick J. Lowrance[1], Carl J. Grillmair[1], Derek Buzasi[2], Drake Deming[3], Hannah Diamond-Lowe[4], Thomas M. Evans[5], G. Morello[6], Kevin B. Stevenson[4], Ian Wong[7], Peter Capak[1], William Glaccum[1], Seppo Laine[1], Jason Surace[1], and Lisa Storrie-Lombardi[1]

[1] Spitzer Science Center, California Institute of Technology, 1200 E California Boulevard, Mail Code 314-6, Pasadena, CA 91125, USA; ingalls@ipac.caltech.edu
[2] Department of Chemistry and Physics, Florida Gulf Coast University, Fort Myers, FL 33965, USA
[3] Department of Astronomy, University of Maryland, College Park, MD 20742-2421, USA
[4] Department of Astronomy and Astrophysics, University of Chicago, 5640 S Ellis Avenue, Chicago, IL 60637, USA
[5] School of Physics, University of Exeter, EX4 4QL Exeter, UK
[6] Department of Physics and Astronomy, University College London, Gower Street, WC1 E6BT, UK
[7] Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA 91125, USA

Received 2016 February 19; revised 2016 May 23; accepted 2016 May 26; published 2016 August 4

ABSTRACT

We examine the repeatability, reliability, and accuracy of differential exoplanet eclipse depth measurements made using the InfraRed Array Camera (IRAC) on the *Spitzer Space Telescope* during the post-cryogenic mission. We have re-analyzed an existing 4.5 μm data set, consisting of 10 observations of the XO-3b system during secondary eclipse, using seven different techniques for removing correlated noise. We find that, on average, for a given technique, the eclipse depth estimate is repeatable from epoch to epoch to within 156 parts per million (ppm). Most techniques derive eclipse depths that do not vary by more than a factor 3 of the photon noise limit. All methods but one accurately assess their own errors: for these methods, the individual measurement uncertainties are comparable to the scatter in eclipse depths over the 10 epoch sample. To assess the accuracy of the techniques as well as to clarify the difference between instrumental and other sources of measurement error, we have also analyzed a simulated data set of 10 visits to XO-3b, for which the eclipse depth is known. We find that three of the methods (BLISS mapping, Pixel Level Decorrelation, and Independent Component Analysis) obtain results that are within three times the photon limit of the true eclipse depth. When averaged over the 10 epoch ensemble, 5 out of 7 techniques come within 60 ppm of the true value. *Spitzer* exoplanet data, if obtained following current best practices and reduced using methods such as those described here, can measure repeatable and accurate single eclipse depths, with close to photon-limited results.

Key words: infrared: planetary systems – methods: data analysis – methods: statistical

1. INTRODUCTION

1.1. Exoplanet Measurements and Correlated Noise

Measurement of relative flux variations is one of the chief means of characterizing transiting exoplanetary systems. At infrared wavelengths, secondary eclipses are a powerful tool for studying the atmospheres of giant exoplanets, with their depths approximately equaling the dayside planet-to-star flux ratio. Extracting information about atmospheres, however, is extremely challenging due to the small differential signals produced by transits, secondary eclipses, and phase curves. The relevant signals are often at the level of 100 parts per million (ppm) or smaller, and require the removal of significant instrumental systematics in the two infrared instruments currently capable of providing information at this precision: the Wide Field Camera 3 (WFC3) on the *Hubble Space Telecope* (*HST*) and the InfraRed Array Camera (IRAC, Fazio et al. 2004) on board the *Spitzer* Space Telescope (Werner et al. 2004). For the IRAC 3.6 and 4.5 μm InSb detectors that remain active on post-cryogenic *Spitzer*, the systematics are due to the interplay of residual telescope pointing fluctuations with intra-pixel gain variations in the moderately undersampled camera.

Over the past decade, a suite of techniques for removing time-correlated noise in IRAC data has been developed. Due to the known coupling between pointing variations and the intra-pixel gain, the earliest methods for correcting cryogenic data used either a simple radial function from a pixel's center (Reach et al. 2005) or fit a second-order polynomial to the observed flux variations as a function of the source centroid position (e.g., Charbonneau et al. 2008). It soon became clear, however, that a single polynomial surface does not sufficiently describe the intra-pixel gain variations. To measure flux decrements with precision less than ~1%, a more responsive approach is necessary to track the small-scale structure in the gain (Ballard et al. 2010). Furthermore, after *Spitzer* entered its post-cryogenic stage in mid-2009, the amplitude of the variations doubled at the current detector temperature of about 28.7 K.[8]

Thus, more flexible non-parametric approaches were developed to measure and remove the systematics. The earliest such methods used some form of nearest neighbor kernel regression to map the intra-pixel gain as a function of centroid position, using a weighted sum of the measured fluxes instead of a predetermined function of centroid (Ballard et al. 2010). A special case of nearest neighbor kernel regression is BiLinearly Interpolated Subpixel Sensitivity (BLISS) mapping (Stevenson et al. 2012). Additional promising techniques that have appeared in recent years include regression via Gaussian Processes (GP; Gibson et al. 2012; Evans et al. 2015); Independent Component Analysis (ICA; Morello 2015); and

[8] http://irsa.ipac.caltech.edu/data/SPITZER/docs/irac/calibrationfiles/pixelphase/

Pixel Level Decorrelation (PLD; Deming et al. 2015). See Appendix B for a detailed review of these techniques.

1.2. Repeatability of Spitzer/IRAC Relative Flux Measurements

As multi epoch monitoring data have accumulated, investigators have begun to quantify the repeatability and reliability of exoplanet differential flux measurements made with *Spitzer* and other observatories. A growing body of evidence is showing that modern IRAC correlated noise-removal techniques obtain consistent results from one measurement to the next, and obtain consistent results between techniques.

One indicator of stability is that the individual measurement uncertainties approximately equal the scatter (standard deviation (SD)) in independently measured transit or eclipse depths. For example, Fraine et al. (2013) analyzed 14 transits of GJ 1214b measured at 4.5 μm with IRAC using a kernel regression decorrelation technique (KR/Data—see Appendix B.4), yielding a scatter in transit depths within 50% of the average reported uncertainty in the individual depths. Wong et al. (2014) also used KR/Data to process data for 12 eclipses of XO-3b, yielding individual uncertainties that were equal to the scatter in the ensemble. The XO-3b data set features prominently in this paper, as a main component of the *Spitzer* 2015 Data Challenge (see below).

Older data often benefit from reanalysis with modern methods. Four GJ 436b transits were reprocessed using ICA by Morello et al. (2015), who determined that the transit depth did not vary by more than 100 ppm, contrary to earlier estimates computed using polynomial fitting (Beaulieu et al. 2011; Knutson et al. 2011). The ICA technique was also used to establish a repeatable (within 200 ppm) transit depth for HD 189733b (Morello et al. 2014), after many conflicting prior values led to questions of stellar variability. BLISS mapping (Diamond-Lowe et al. 2014) and GP (Evans et al. 2015) were both used to reanalyze four eclipses of HD 209458b, including one taken under non-optimal observing conditions (see Section 4.3). Both teams concluded that the group of measurements was self-consistent (scatter 30% less than uncertainties for Evans et al. 2015), and that the earlier estimate of a much deeper occultation, which resulted in claims of a possible temperature inversion layer in the planet's atmosphere (Knutson et al. 2008), was unwarranted.

1.3. Goals of this Paper

Because of the high relative precision required for eclipse depth and other exoplanet measurements, it is important to characterize the ability of an instrument—together with the chosen method of systematics removal—to return consistent results. This is especially crucial when comparing data to models (see Burrows 2014, for a discussion of the difficulty of spectral retrieval from data with low signal-to-noise ratio (S/N)) or measuring atmospheric variability (e.g., see Demory et al. 2016, who found evidence for eclipse depth changes of ~140 ppm over 1 year in 55 Cnc e). Despite the growing number of analyzes of multi epoch transit or eclipse measurements, all have thus far focused on at most two methods of removing correlated noise (Fraine et al. 2013; Diamond-Lowe et al. 2014; Wong et al. 2014; Evans et al. 2015; Morello et al. 2015; Demory et al. 2016), or only considered two epochs per target (Hansen et al. 2014).

This paper examines the repeatability of *Spitzer*/IRAC eclipse depths in the post-cryogenic mission, with an eye toward answering the questions: how stable can we reasonably expect IRAC eclipse depth measurements to be; and how close are they to the truth? We aim to establish limits on both the IRAC instrument and the best modern techniques for removing correlated noise and measuring eclipse depths, using both real and simulated data. Recently, participants undertook a Data Challenge consisting of the measurement of 10 secondary eclipses of XO-3b (Wong et al. 2014), and a complementary analysis of a synthetic version of the XO-3b data. In Section 2, we describe the Data Challenge. We introduce the real XO-3b data set, give an overview of the *Spitzer*/IRAC simulator and the creation of the simulated data set, and outline seven techniques used to decorrelate the photometry. In Section 3, we report on the results of the data challenge, estimating the single eclipse depth repeatability and the reliability or precision of the results when reduced by the different methods. We compare the variability between methods, as well as the accuracy of the techniques when applied to simulated data. In Section 4, we discuss the implications of our results for post-cryogenic exoplanet measurements with *Spitzer*. We also evaluate a recent proposal to inflate IRAC eclipse depth uncertainties (Hansen et al. 2014), and suggest application of our approach to future space observatories. We conclude in Section 5 by summarizing our key results.

2. METHODOLOGY

2.1. The IRAC 2015 Data Challenge

To assess the repeatability, reliability, and accuracy of post-cryogenic observations with IRAC, the *Spitzer* Science Center (SSC) in conjunction with active exoplanet researchers from the astronomical community has performed an analysis of the removal of systematics and measured the repeatability of warm IRAC observations. The SSC made available to the public both a real data set as well as synthetic data (where the eclipse depth is an input) on the IRAC Data Challenge 2015 website.[9] Contributions were solicited, and preliminary results were presented at the IRAC 2nd Workshop on High Precision Photometry, held during the 2015 International Astronomical Union meeting in Honolulu, HI, USA.[10] In this section, we describe the real and simulated data and the decorrelation techniques used.

2.2. Real XO-3b Observations

The XO-3b data used for the Data Challenge consisted of 10 individual secondary eclipse measurements originally analyzed by Wong et al. (2014), and summarized in Table 1. All measurements were made with post-cryogenic *Spitzer* in 2012 and 2013, and were taken as part of Program ID (PID) 90032 (PI: H. Knutson). This program also contains two full phase curve measurements of XO-3b at 4.5 μm, but we confine our analysis in this paper to the eclipse–only data sets. The first six epochs took place within about 30 days of each other; the last four occurred about one-half year later and also spanned 30 days. Each epoch consisted of two Astronomical Observation Requests (AORs): an 11 exposure, 30 min "Pre" AOR to allow short-term pointing drift to settle; and a 233 exposure,

[9] http://irachpp.spitzer.caltech.edu/page/data-challenge-2015

[10] http://irachpp.spitzer.caltech.edu/page/IRAC_IAU_2015

Table 1
Real *Spitzer* XO-3b Eclipse AORs and Positions

Start Time[a]	AOR Number[b]		$\langle X \rangle$[c]	σ_X[d]	$\langle Y \rangle$[c]	σ_Y[d]	σ_{XY}[e]	No.[f]
(JD-2455000)	Pre	Main	(px)	(px)	(px)	(px)	(10^{-4} px^2)	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1242.2402	[46467072]	[46471424]	15.17	0.03	15.00	0.05	−8.56	2
1248.6482	[46467840]	[46471168]	15.10	0.04	15.14	0.06	9.53	3
1251.8187	[46470144]	[46470912]	15.23	0.06	15.03	0.06	−25.63	4
1255.0166	[46467584]	[46470656]	15.17	0.04	15.13	0.05	4.25	5
1264.5897	[46469376]	[46470400]	15.19	0.03	14.99	0.06	−4.31	6
1270.9776	[46466816]	[46469632]	15.13	0.06	15.12	0.05	9.22	7
1405.0165	[46468864]	[46469120]	15.21	0.03	14.92	0.04	−3.93	8
1430.5523	[46469888]	[46468608]	15.15	0.03	15.01	0.05	−4.45	10
1433.7433	[46467328]	[46468352]	15.21	0.03	14.99	0.05	−4.04	11
1436.9273	[46471680]	[46468096]	15.23	0.04	14.96	0.05	−2.80	12
		Column Means[g]	15.18	0.04	15.03	0.05	−3.07	…
		All Data[h]	15.18	0.06	15.03	0.09	−26.63	…

Notes.
[a] Start time of first exposure of initial AOR.
[b] Electronic versions of this table contain links to these data sets in the *Spitzer* archive.
[c] Mean centroid over all measurements in the two AORs.
[d] Standard deviation in centroid over all measurements in the two AORs.
[e] (x, y) covariance in centroid over all measurements in the two AORs.
[f] Eclipse number as listed in Table 1 of Wong et al. (2014; not all eclipses analyzed by Wong et al. were part of the Data Challenge).
[g] Mean, standard deviation, and (x, y) covariance of centroid averaged along the table column.
[h] Mean, standard deviation, and (x, y) covariance of centroid over all AORs.

8.5 hr "Main" AOR that contained the secondary eclipse. Each exposure produced a FITS format image file, containing a cube of 64 32 × 32 pixel images taken 2 s apart with the source in the subarray field of view on the 4.5 μm array. The measurements were taken in staring mode (no repositioning within an AOR), and used PCRS Peak-Up to establish the position of XO-3b at the beginning of each AOR.[11] Table 1 gives the observation start time, *Spitzer* AOR numbers, and the eclipse number (for comparison with Wong et al. 2014, Table 1, which also includes two full phase curve data sets).

2.3. Synthetic XO-3b Observations

Observed variations in eclipse depths are caused by a combination of variations in *Spitzer* pointing, IRAC detector charge trapping, and possible evolution of the planetary system, as well as the limitations and biases of the technique for reducing correlated noise. We can analyze the data using different techniques to assess differences in the methods, but it is often difficult with real data to completely separate pointing from instrumental or planetary variations. This is one reason we have included synthetic data as part of the Data Challenge, for which both the exoplanet and IRAC are given constant properties. We had originally considered using eclipsing binary stars observed with *Kepler* as a truth set which could then be observed with *Spitzer*. Unfortunately, using stellar atmosphere models to extrapolate *Kepler* eclipse depths to *Spitzer* wavelengths are as fraught with potential uncertainty as the planetary eclipse depths themselves, suggesting simulated data are the only reasonable path to estimating accuracy. In the simulations, any measured variations in eclipse depth are due solely to (1) random noise and (2) residual correlated noise not removed by decorrelation analysis. This should give us a better insight into the capabilities of the decorrelation methods than real data alone.

To produce the simulated XO-3b observations used for the Data Challenge, we used `IRACSIM`, a package built in the IDL programming language. The program uses a model of the *Spitzer*/IRAC system to create synthetic IRAC point source measurements, outputting FITS image (or image cube) files similar to those produced by the IRAC basic calibrated data (BCD) pipeline. We give an overview of the model in Appendix A.

Table 2 gives the simulated observation start times and AOR numbers of the synthetic observations. The simulations followed closely the design of the real observations, with each observing "epoch" containing two AORs, a similar number of exposures per AOR, and the same integration parameters. We set the start times for each synthetic epoch at slightly different phases of different actual XO-3b orbits, as often occurs in real observations. This allows for different proportions of samples before and after eclipse for each epoch to minimize biases in fitting.

Table 5 lists the range of inputs to the pointing model used in simulating the XO-3b data. We chose not to duplicate exactly the pointing fluctuations as observed in the real data set, but attempted to simulate a range of possible *Spitzer* observing conditions (drawn roughly from the distribution of observed cases), and thus a range of possible decorrelation situations. In practice, this resulted in generally larger pointing fluctuations and drifts than found in the real data.

We used the `IRACSIM` exoplanet wrapper to model the light curve of XO-3b, obtaining values for the system's stellar, orbital, and transit parameters from the exoplanets.org database (as of 2015 July 2) and simulating the planet's thermal phase variations using the model of Cowan & Agol (2011). Since the goal was to understand IRAC data, not XO-3b, we set somewhat arbitrary values of the planetary parameters: (1)

[11] http://irachpp.spitzer.caltech.edu/page/Obs%20Planning

Table 2
Synthetic *Spitzer* XO-3b AORs and Positions

Start Time[a]	AOR Number[b]		$\langle X \rangle$	σ_X	$\langle Y \rangle$	σ_Y	σ_{XY}
(JD-2455000)	Pre	Main	(px)	(px)	(px)	(px)	(10^{-4} px^2)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
2206.1459	20150000	20150001	15.16	0.02	15.01	0.05	−5.93
2209.2927	20150002	20150003	15.12	0.02	15.08	0.03	−4.57
2212.5079	20150004	20150005	15.16	0.03	15.13	0.05	−10.33
2215.7214	20150006	20150007	15.10	0.02	15.06	0.10	−16.34
2218.9047	20150008	20150009	15.20	0.03	15.17	0.07	−10.95
2222.0547	20150010	20150011	15.12	0.02	15.17	0.04	−5.19
2225.2356	20150012	20150013	15.18	0.03	15.05	0.09	−20.87
2228.4898	20150014	20150015	15.17	0.02	15.09	0.10	−10.61
2231.6296	20150016	20150017	15.14	0.02	15.17	0.05	−7.41
2234.8406	20150018	20150019	15.10	0.03	15.09	0.06	−6.71
		Column Means	15.15	0.03	15.10	0.06	−9.90
		All Data	15.15	0.04	15.10	0.09	−8.64

Notes.
[a] Simulated start time of first exposure of initial AOR.
[b] Data may be downloaded from http://irachpp.spitzer.caltech.edu/page/data-challenge-2015.

albedo, $A = 0$; (2) radiative timescale, $\tau_{\rm rad} = 1$ day; and (3) net rotational angular velocity of the cloud layer, $\Omega_{\rm rot} = 1$ (in units of the orbital angular velocity at periastron). The resulting phase curve gives a non-flat appearance to the flux outside of eclipse and sets the depth of the eclipse, which we define in terms of the stellar flux. In this case, the model eclipse depth for XO-3b is 1875 ppm, about 16% larger than the actual depth published by Wong et al. (2014). The model light curve for the 10th epoch is shown in Figure 1.

2.4. Decorrelation Techniques

The best hypothesis for the source of IRAC time-correlated noise is the coupling of pointing fluctuations with intra-pixel quantum efficiency variations on the InSb detector arrays. When *Spitzer* is commanded to continuously observe an inertially fixed target ("staring" mode), a source position will undergo "jitter" and "wobble" with a net amplitude of about 0.08 detector pixels (px) per hour, while also incurring a slow linear drift of about 0.01 px per hour (see Appendix A.1 for an analytical model of these fluctuations). These telescope motions have been described in detail by Grillmair et al. (2012), and the physical causes of some are known. For example, the wobble is caused by a battery heater cycling on and off with period of ~40 minutes, and the long term drift (y pixel direction only) is caused by the discrepancy between the instantaneous velocity aberration of the spacecraft and the on board aberration correction that occurs only at the start of an AOR. A map of the photometric gain of a point source on the central pixel of the 4.5 μm subarray is displayed in Figure 2, showing that correlated noise due to pointing fluctuations can be as much as 1%–2%, a factor of 10 larger than the XO-3b eclipse depth.

As part of the Data Challenge, exoplanet experts used a total of seven different data reduction techniques to remove correlated noise from the *Spitzer*/IRAC photometry and assess the eclipse depth repeatability. We review the seven techniques in Appendix B, including notes on implementation for the XO-3b data sets. Among these are the most commonly used techniques in the current literature to date (BLISS, KR/Data), as well as a group of more recently developed methods (GP, ICA, KR/Pmap, PLD, Segmented Polynomial (K2 pipeline)). Note that each expert was free to use any approach to centroiding, photometry, and eclipse depth fitting. Thus, any mention of a method by name in this paper refers to the *entire data reduction pipeline*, not just the correlated noise-removal algorithm.

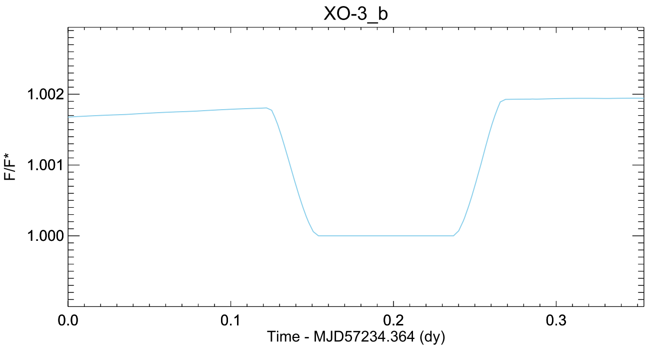

Figure 1. Light curve for simulated XO-3b observations, AOR number 20150019. The simulation uses known values of the system's stellar, orbital, and transit parameters, in addition to a thermal phase model with albedo, $A = 0$; radiative timescale, $\tau_{\rm rad} = 1$ day; and net rotational angular velocity of the cloud layer, $\Omega_{\rm rot} = 1$, in units of the orbital angular velocity at periastron. The eclipse depth is 1875 ppm.

3. RESULTS

3.1. XO-3b Centroids and Photometry

We begin with an overview of the data characteristics. Figure 2 plots all centroid positions for the individual measurements on the subarray center pixel for the real data, and Figure 3 does the same for the simulated data. Due to the dependence of correlated noise on pointing fluctuations, most noise-removal techniques use source centroid as a primary decorrelation variable. Most techniques described in Section 2.4 use either two-dimensional (2D) Gaussian fitting or the flux-weighted "center of light" method to determine a point source center on the undersampled *Spitzer* arrays. In this paper,

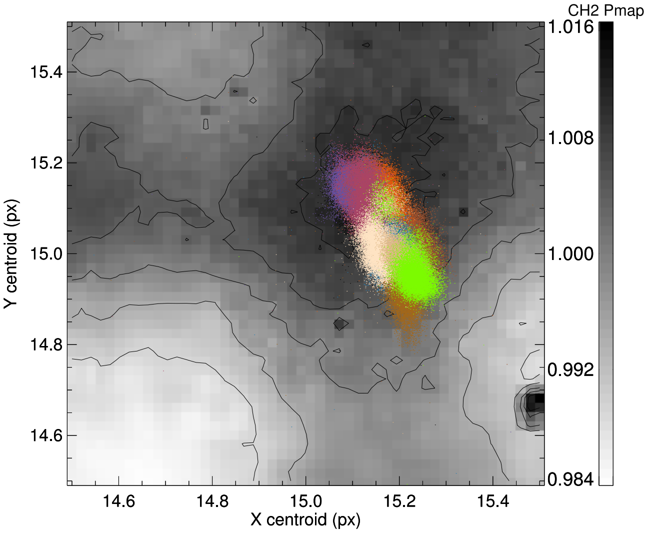

Figure 2. Centroid positions (derived using the center-of-light method) of XO-3 on the IRAC 4.5 μm subarray, from the real data set. Each colored group of points indicates a separate epoch of observation (see Table 1 for details on the epochs). The background grayscale and contours shows the intra-pixel photometric gain map ("Pmap"), as measured using kernel regression on a calibration star (Appendix B.4). The geometric center of the pixel is located at coordinates (15.0, 15.0).

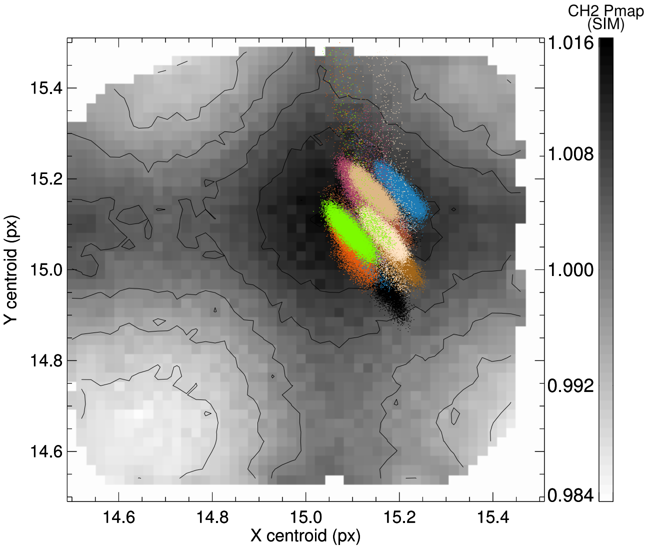

Figure 3. Centroid positions (derived using the center-of-light method) of XO-3 on the IRAC 4.5 μm subarray, from the simulated data set. Each colored group of points indicates a separate epoch of observation (see Table 2 for details on the simulated epochs). The background grayscale and contours shows the intra-pixel photometric gain map ("Pmap"), as measured using kernel regression on a simulated calibration star (Appendix B.4).

if not stated otherwise, we only report center of light centroids:

$$x_c = \frac{\sum_i i\,(f_{ij} - f_{\rm BG})}{\sum_i (f_{ij} - f_{\rm BG})}; \tag{1}$$

$$y_c = \frac{\sum_j j\,(f_{ij} - f_{\rm BG})}{\sum_j (f_{ij} - f_{\rm BG})}. \tag{2}$$

Here, (i, j) is the pixel number, f_{ij} is the image value at that pixel, and $f_{\rm BG}$ is the background flux in surrounding pixels. The sums are over a (7×7) pixel region surrounding the expected position of the source. This centroiding method is sufficiently precise for decorrelation, resulting in positional distortions of at most 0.05 pixels (Ingalls et al. 2014). A detailed discussion of different centroiding techniques is beyond the scope of this paper; see Lust et al. (2014) for analysis of the accuracy of three centroiding methods.

Columns 4–8 of Tables 1 (real) and 2 (simulated) summarize the centroid "clouds" for each epoch, giving the means and SD in x and y position, as well as the xy covariances in centroid. As the real data were all pointed using PCRS peak-up, the mean positions are all within 0.4 pixel of one another, and cluster near the peak of the intra-pixel gain. Negative covariances for most of the real AORs indicate that the clouds are aligned such that y decreases when x increases, which is a common direction for *Spitzer*/IRAC short-term drift. The bottom two rows of Tables 1 and 2 list the column means and the statistics for all data taken together. For the real data, the full data set has a much higher negative covariance than the individual clouds, suggesting that separate pointings fall preferentially along a $\sim -45°$ axis. The simulated data feature a much stronger initial drift, as well as a more pronounced y component to the jitter, wobble, and drift than the real data. Some individual xy covariances in the simulated data are much higher than both the mean and the aggregate covariance for the group, due to this exaggerated elongation. This "stretching" of the positions along y reduces the positional redundancy and, as we will see, challenges the ability of most reduction methods to decorrelate the data.

We display the real XO-3b photometric and other measurements as a function of orbital phase in Figure 4, and those for the synthetic data in Figure 5. As mentioned, some decorrelation methods use the Noise Pixels parameter

$$\tilde{\beta} = \frac{\left(\sum_{ij} f_{ij}\right)^2}{\sum_{ij} f_{ij}^2}, \tag{3}$$

which approximates the effective area (in square pixels) of a point source. The sums are over the same (7×7) pixel region over which the centroid is derived. We display $\tilde{\beta}$ as a function of phase in the third panel of Figures 4 and 5. This parameter partly measures observing geometry: given constant total flux —the numerator of Equation (3) does not change—moving a source from the center to the edge of a pixel will spread the light to more pixels, decreasing the denominator and increasing $\tilde{\beta}$. Thus we see in Figures 4 and 5 how $\tilde{\beta}$ is correlated with the centroids to an extent. But $\tilde{\beta}$ also can measure the smearing of the IRAC PSF due to changes in the amplitude of high-frequency jitter. *Spitzer* is known to have normal modes of oscillation with period less than the detector sample time (see Appendix A.3 for a discussion of IRAC sampling). When the amplitude of oscillation changes, the centroid might not vary markedly but the integrated PSF will change its apparent size, altering $\tilde{\beta}$.

We display the mean background in the fourth panel, and the aperture flux in the fifth panel of Figures 4 and 5. We normalized the values in these last two panels to the mean value over all AORs, allowing us to notice relative shifts between AORs. Fluxes are estimated using the IDL photometry program `aper.pro`, with a 2.25 pixel radius circular aperture and a 3–7 pixel background annulus. Backgrounds are the

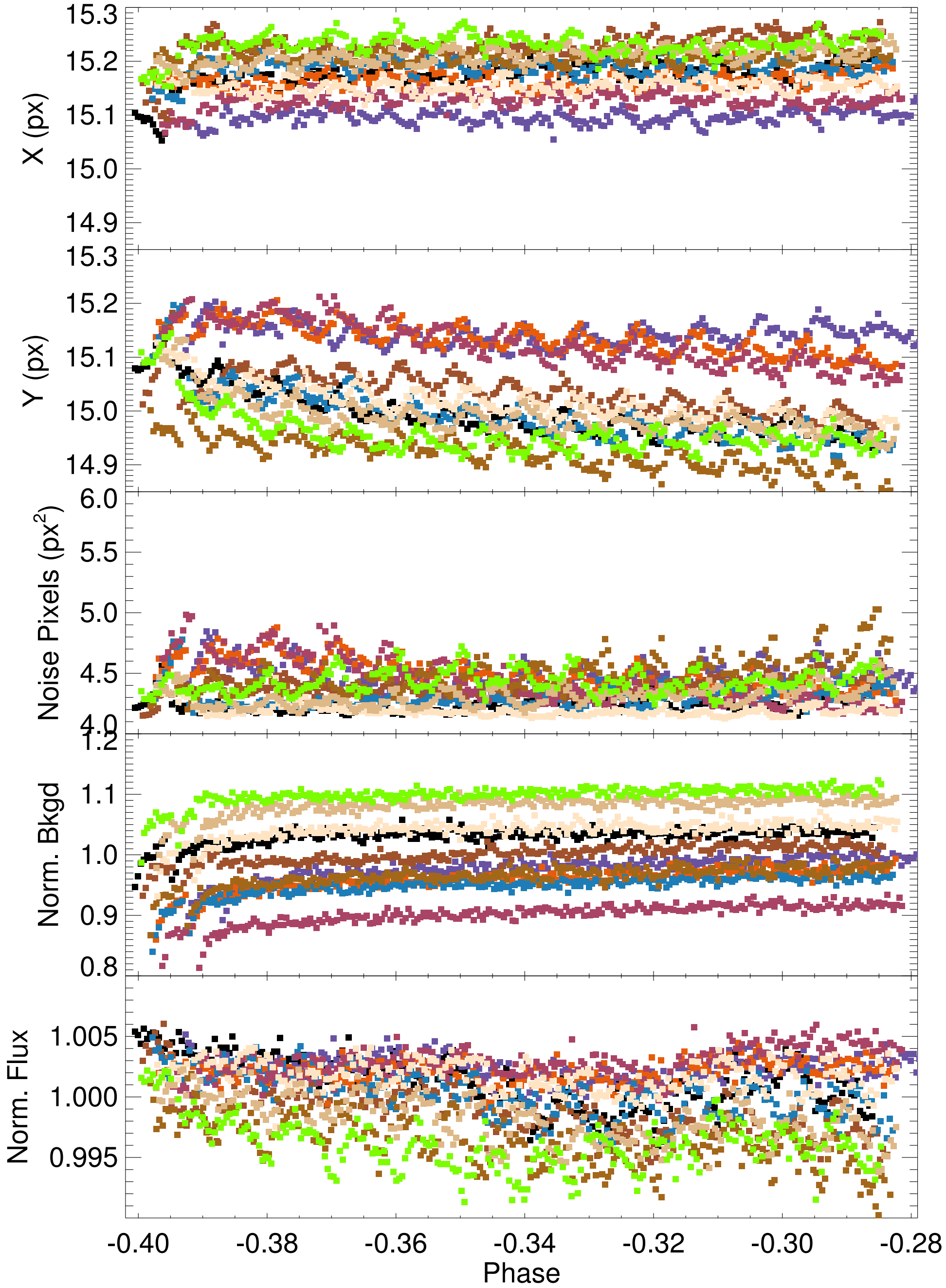

Figure 4. Real XO-3b photometric and other measurements as a function of orbital phase (fraction of orbit since transit). From top to bottom: x and y centroid positions; Noise Pixel parameter, $\tilde{\beta}$; photometric background in 3–7 pixel radius annulus surrounding centroid, normalized to the mean over all AORs; and photometric flux in 2.25 pixel radius aperture, normalized to the mean over all AORs. Each point on the plots is the average of 63 measurements, or ~2 min of integration. We drop the first frame of each 64 frame subarray cube, to minimize residual bias pattern effects (the "first frame effect"; see http://irsa.ipac.caltech.edu/data/SPITZER/docs/irac/features/#1). Each colored group of points indicates a separate epoch of observation (see Table 1 for details on the epochs). The time span is about 9 hr.

mean value per pixel in the annulus, scaled to the area of the aperture. The net aperture flux is thus the integrated intensity per pixel weighted by the fraction of each pixel lying inside the aperture, minus the background value. (Each team participating in the Data Challenge may have used a different method for measuring the flux, including different aperture sizes or background definitions.)

Various known features of *Spitzer* data can be seen in Figure 4. The short-term pointing drift, as well as the sawtooth-shaped "wobble," can be seen in the x and y centroids (and to a lesser extent in $\tilde{\beta}$; pointing effects are stronger in the y direction). The aperture fluxes for some epochs show very clearly the correlated noise signature due to these telescope motions. The eclipse can also be seen in the flux between phase -0.35 and -0.31.

The background values show a quick ramp in the beginning of each epoch, and settle into a much slower increase with time for the final eight or so hours. This behaves similarly to the "flux ramp" seen by many who work on 4.5 μm staring mode IRAC data (e.g., Knutson et al. 2012; Lewis et al. 2013). In this

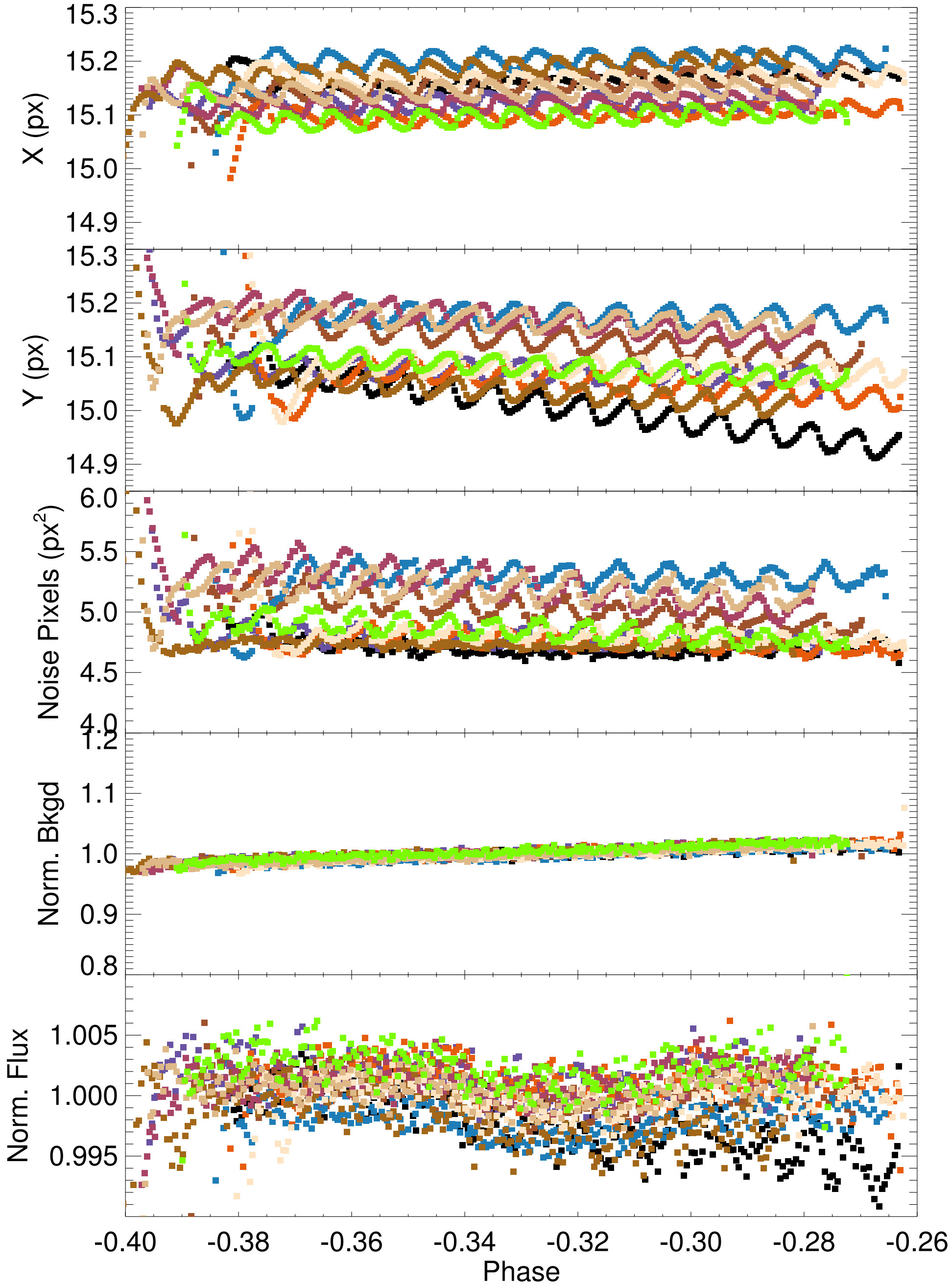

Figure 5. Simulated XO-3b photometric and other measurements as a function of orbital phase. See caption to Figure 4 for description. Each colored group of points indicates a separate epoch of observation (see Table 2 for details on the epochs). Vertical scales for each panel are identical to those in Figure 4.

case, however, the ramp disappears after background subtraction so the background ramp is probably caused by a relaxation in detector bias (the IRAC dark bias has a significant well-known offset that changes with time based on the history of readouts and array idling over the previous several hours), not changing responsivity. The background curves in Figure 4 are all normalized to the same value; the fact that they are separated suggests a different mean background between epochs. This can be attributed to fluctuations in the mean detector dark bias or changes in the residual sky subtraction (or a combination of the two).

The simulation data (Figure 5) show many of the same features, with a few differences. First, noise on timescales shorter than the wobble period averages quite cleanly to near zero in the binned measurements of centroid and noise pixels, as compared to the same plots for the real data set. This suggests the presence in real data of a jitter signal that does not integrate to zero in 64 samples, perhaps with a steeper spectrum (more power at low frequencies) than the $1/f$ signal currently included. Second, the magnitudes of short and long term pointing drift and the amplitude of pointing fluctuations are all larger than in the real data, as seen in the (x, y) centroids and $\tilde{\beta}$. This is also visible in Figure 3 when compared with Figure 2. Third, the simulated backgrounds are much more uniform from one AOR to another because we commanded the same linear increase with time, with constant mean and no offsets between epochs. Fourth, the larger spread in position has increased the overall noise in the light curve. This will have consequences for

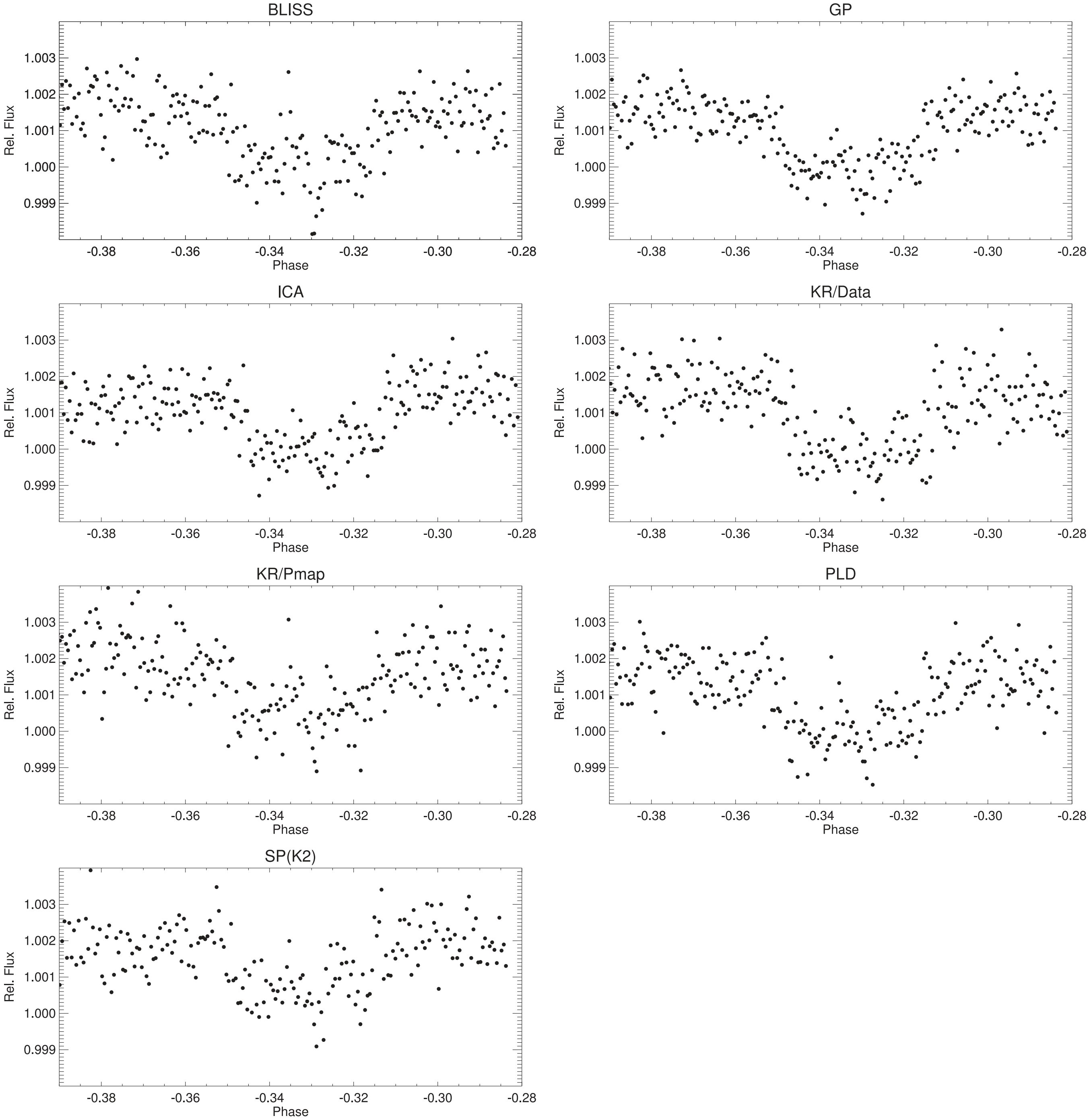

Figure 6. Decorrelated light curves for real XO-3b measurements, 5th epoch (*Spitzer* AOR number 46470400). Fluxes have been binned 64×.

the decorrelation of the measurements and the estimation of eclipse depths.

We display in Figures 6 and 7 light curves decorrelated using different techniques, for the 5th epochs of the real and simulated observations.

3.2. Eclipse Depths

All of the seven Data Challenge participants estimated eclipse depths and uncertainties from decorrelated light curves, for each set of 10 epochs from the real and simulated data sets. Figure 8 plots the measured depths for the real data and Figure 9 plots the results for the simulated data. We define the eclipse depth in terms of the stellar flux:

$$D = \frac{F_{\rm out} - F_{\rm in}}{F_{\rm in}}, \tag{4}$$

where $F_{\rm in}$ is the average photometric flux in eclipse (i.e., the stellar flux) and $F_{\rm out}$ is the flux out of eclipse, interpolated to the center of occultation. We plot weighted average eclipse depths, $\overline{D}$, for each of the seven data reduction methods on the right hand side of Figures 8 and 9. The averages are weighted

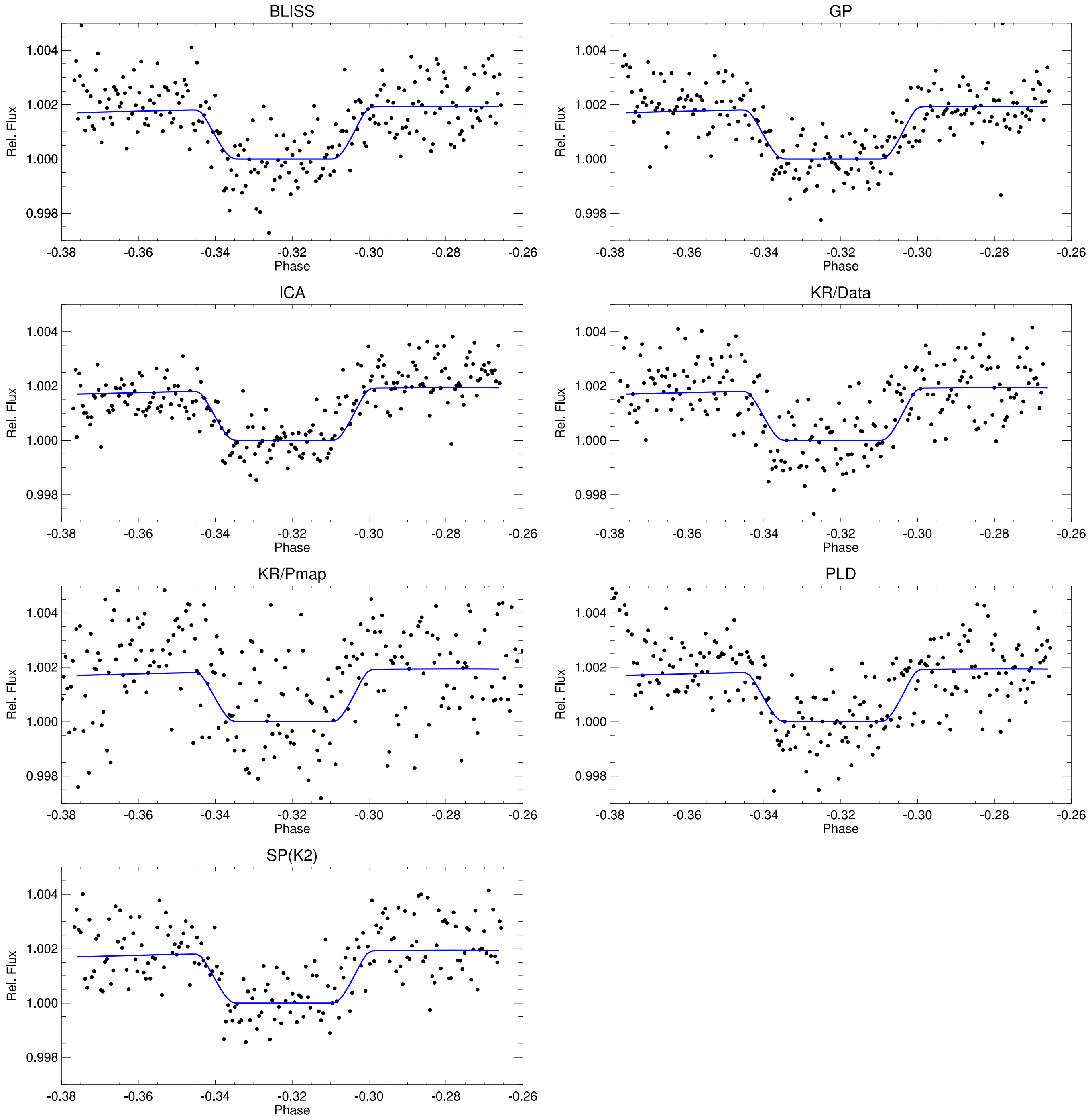

Figure 7. Decorrelated light curves for simulated XO-3b measurements, 5th epoch (simulated AOR number 20150009). Fluxes have been binned 64×. We overlay the input model light curve (not a fit) as a blue solid line.

sums of the individual eclipse measurements:

$$\overline{D} = \frac{\sum_{i=1}^{N} w_i\, D_i}{\sum_{i=1}^{N} w_i}; \tag{5}$$

where the weights consist of the usual inverse variances, but multiplied by an "overdispersion" factor (see Lyons 1992):

$$w_i = \frac{1}{\sigma_i^2\, f_{\rm dis}^2}. \tag{6}$$

The factor $f_{\rm dis}$ allows for the possible underestimation of the individual uncertainties, using the scatter in the group of measurements as an additional constraint. We derive it using the χ^2 equation for the mean value (assuming the D_i values are distributed normally about $\overline{D}$):

$$\chi^2 = \sum_{i=1}^{N} \frac{(D_i - \overline{D})^2}{\sigma_i^2\, f_{\rm dis}^2} = N - 1. \tag{7}$$

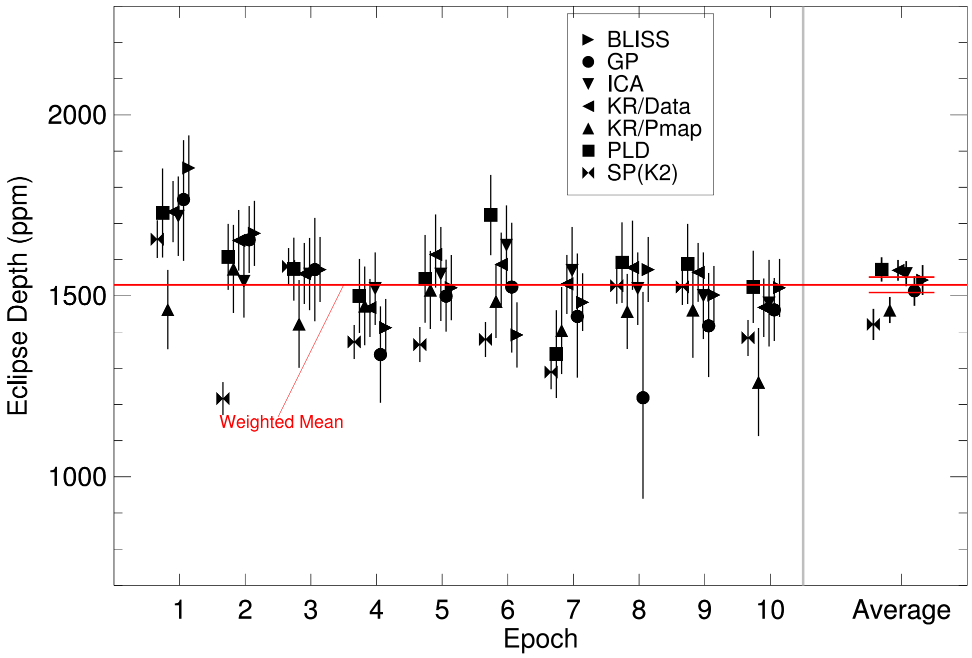

Figure 8. Eclipse depths for 10 real visits to XO-3b, as computed via various methods. The group of points for each epoch is separated to minimize confusion. Error bars in this plot are symmetric; in cases where the technique returned asymmetric uncertainties, we used the largest of the two values. We show the results for the separate visits to the left of the gray vertical line, and the average results to the right. Error bars on the separate visits are the uncertainties reported by the technique. Error bars on the averages are the uncertainties in the weighted mean, adjusted for "underdispersion" by a factor $f_{\rm dis}$ (see text). The horizontal red lines display the grand mean for all results, ± its uncertainty.

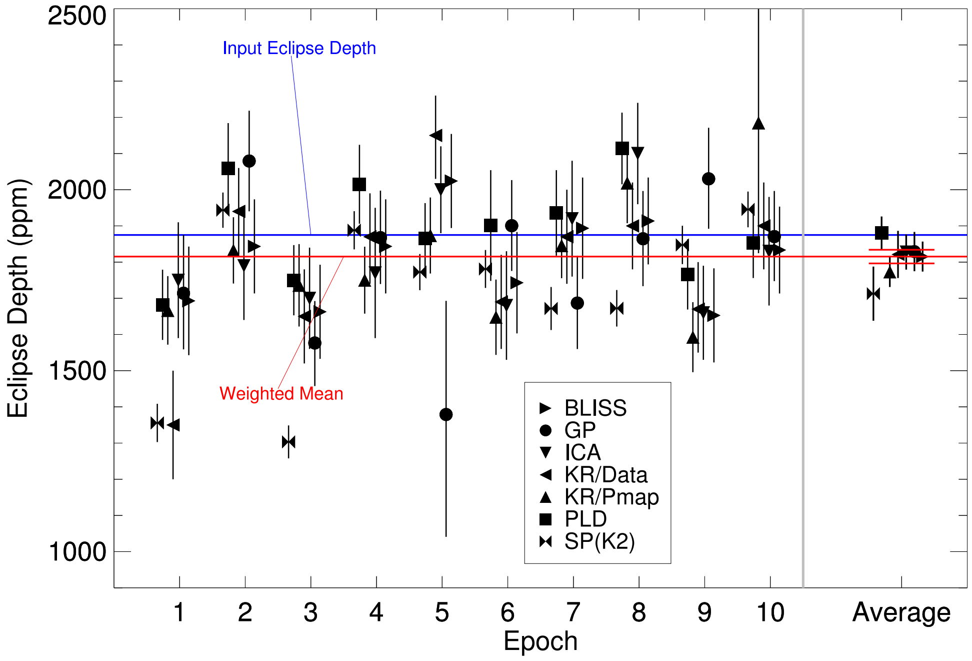

Figure 9. Eclipse depths for 10 simulated visits to XO-3b, as computed via various methods. A blue horizontal line indicates the eclipse depth input to the simulations, 1875 ppm. See caption for Figure 8 for further description.

This can be inverted to solve for $f_{\rm dis}$ (note that since Equation (5) contains $f_{\rm dis}^{-2}$ in both the numerator and denominator, $\overline{D}$ does not depend on $f_{\rm dis}$):

$$f_{\rm dis}^2 = \sum_{i=1}^{N} \frac{(D_i - \overline{D})^2}{\sigma_i^2 (N-1)}. \qquad (8)$$

The total variance in the mean is given by the inverse sum of weights:

$$\sigma_{\rm TOT}^2 = \frac{1}{\sum_{i=1}^{N} w_i} = \frac{1}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2 f_{\rm dis}^2}} = f_{\rm dis}^2\, \sigma_{\rm orig}^2, \qquad (9)$$

Table 3
Eclipse Depth Statistics: Real Data ($\sigma_{\rm phot} \approx 53$ ppm)

Method	$\overline{D}$[a]	$\overline{\sigma}$[b]	SD[c]	$\sigma_{\rm orig}$[d]	$f_{\rm dis}$[e]	$\sigma_{\rm TOT}$[f]	R[g]	r[h]	Closest Match[i]
	(ppm)	(ppm)	(ppm)	(ppm)		(ppm)	(ppm)		
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
BLISS	1543	85	133	27	$1.5^{+0.5}_{-0.3}$	40	189	0.40	KR/Data: (−25 ± 86)
GP	1513	152	155	40	1.0	40	220	0.34	BLISS: (−60 ± 121)
ICA	1560	111	71	34	1.0	34	101	0.74	KR/Data: (−14 ± 56)
KR/Data	1570	94	79	28	1.0	28	113	0.66	ICA: (14 ± 56)
KR/Pmap	1460	117	81	36	1.0	36	116	0.65	SP(K2): (21 ± 172)
PLD	1573	107	111	33	1.0	33	158	0.48	KR/Data: (−3 ± 86)
SP(K2)	1421	48	137	15	$2.8^{+1.0}_{-0.5}$	43	195	0.39	KR/Pmap: (−21 ± 172)
Average[j]	1520	102	110	30	1.3	36	156	0.52	⋯

Notes.
[a] Weighted mean eclipse depth over the 10 AOR measurements of XO-3b.
[b] Mean eclipse depth uncertainty reported for the 10 AOR measurements.
[c] Sample standard deviation in eclipse depth over the 10 AORs.
[d] Weighted uncertainty in the mean eclipse depth, based only on the originally reported uncertainties.
[e] "Dispersion factor" that multiplies the uncertainties, required to make $\chi^2_\nu = 1$ (see text).
[f] Total uncertainty in the mean, after being corrected for dispersion, $\sigma_{\rm TOT} = f_{\rm dis}\ \sigma_{\rm orig}$.
[g] The "repeatability," ie., the standard deviation in differences between pairs of eclipse depth measurements.
[h] The "reliability" of the technique, $\sigma_{\rm phot}$/SD.
[i] Technique with the closest range in eclipse values to this one, followed by (Mean ± SD) difference.
[j] Straight averages along the columns.

Table 4
Eclipse Depth Statistics: Simulated Data ($\sigma_{\rm phot} \approx 53$ ppm)

Method	$\overline{D}$	$\overline{\sigma}$	SD	$\sigma_{\rm orig}$	$f_{\rm dis}$	$\sigma_{\rm TOT}$	R	r	Closest Match	RMSE[a]	$\overline{B}$[b]	a[c]
	(ppm)	(ppm)	(ppm)	(ppm)		(ppm)	(ppm)			(ppm)	(ppm)	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
BLISS	1815	131	120	41	1.0	41	171	0.44	ICA: (−9 ± 76)	131	−59	0.40
GP	1829	154	211	43	$1.2^{+0.4}_{-0.2}$	54	300	0.25	BLISS: (−13 ± 277)	215	−45	0.25
ICA	1827	148	144	45	$1.1^{+0.4}_{-0.2}$	48	205	0.37	BLISS: (9 ± 76)	148	−47	0.36
KR/Data	1821	128	217	40	$1.6^{+0.6}_{-0.3}$	65	309	0.24	BLISS: (−11 ± 129)	219	−53	0.24
KR/Pmap	1772	125	180	32	$1.3^{+0.4}_{-0.2}$	41	256	0.29	ICA: (−5 ± 137)	181	−102	0.29
PLD	1880	108	140	33	$1.4^{+0.5}_{-0.2}$	45	199	0.38	BLISS: (83 ± 114)	134	5	0.39
SP(K2)	1712	51	226	16	$4.6^{+1.6}_{-0.8}$	74	322	0.23	KR/Data: (−80 ± 191)	266	−162	0.20
Average	1808	121	177	36	1.7	53	252	0.32	⋯	185	−66	0.30

Notes.
[a] Root mean square deviation of the 10 individual measurements from the input eclipse depth of 1875 ppm.
[b] The mean bias, or deviation of $\overline{D}$ from the input eclipse depth.
[c] Accuracy of technique, $\sigma_{\rm phot}$/RMSE.

where $\sigma_{\rm orig}$ is the original uncertainty in the mean derived from $w_i = 1/\sigma_i^2$ (i.e., $f_{\rm dis} = 1$).

In Tables 3 (real) and 4 (simulated), we list the values for $\overline{D}$, the mean uncertainty $\overline{\sigma}$, the SD in depth, as well as $\sigma_{\rm orig}$, $f_{\rm dis}$, and $\sigma_{\rm TOT}$ for each technique. Wherever SD $\gtrsim \overline{\sigma}$, one expects that the uncertainties have been underestimated, and indeed in all instances where this holds, $f_{\rm dis} > 1$. For the real data, only two techniques had underestimated uncertainties ($f_{\rm dis} > 1$), and for both real and simulated data only one technique, SP(K2), which was not developed for Spitzer data, has $f_{\rm dis} > 2$.

Since the sum in Equation (7) defines a χ^2 probability distribution with N–1 degrees of freedom, we derive a 68% confidence interval on $f_{\rm dis}$ as those values for which the distribution obtains 16% and 84% of its integrated area. The resulting intervals are specified on Tables 3 and 4 as positive and negative error bars on $f_{\rm dis}$.

3.3. The Photon Limit

Because our goal in this paper is to assess the potential variability in eclipse depth measurements we must first calculate the noise floor for the real and simulated data sets, i.e., the intrinsic variability due to photoelectron counting statistics and readout noise.

We estimate the S/N for a single 2 s data frame based on aperture photometry. Combining Equations (5) and (13) of Garnett & Forrest (1993), the variance in Fowler-sampled electron counts (including the effects of readout noise) is

$$\sigma_e^2 = \sigma_{\rm int}^2\left[\frac{2\sigma_{\rm rn}^2}{\sigma_{\rm int}^2\cdot({\rm FN})} + 1 - \frac{2({\rm FN})}{3n_{\rm max}} + \frac{1}{6({\rm FN})\cdot n_{\rm max}}\right], \tag{10}$$

where $\sigma_{\rm int}^2$ is the equivalent shot noise variance in electron counts accumulated over the *integration* time ($t_{\rm int} = n_{\rm max}\Delta t$), FN is the Fowler number, $\sigma_{\rm rn}$ is the SD of the readout noise (per read), $n_{\rm max} = 2({\rm FN}) + {\rm WT}$ is the total number of Fowler samples per integration, WT is the number of wait ticks, and Δt is the sample time (see Appendix A.3 for more information on Fowler sampling with IRAC). For 2 s subarray measurements, ${\rm FN} = 8$, ${\rm WT} = 184$, $\Delta t = 0.01$ s, and $\sigma_{\rm rn} = 9.4$ e. The shot noise variance has the same value as the total electron counts accumulated over the entire integration, $\sigma_{\rm int}^2 \approx F_e(t_{\rm int}/t_{\rm exp})$. (The scale factor $t_{\rm int}/t_{\rm exp}$ is necessary because Fowler sampling returns F_e, the accumulated charge per *exposure* time, $t_{\rm exp} = ({\rm FN} + {\rm WT})\Delta t$.)

To estimate σ_e we average over the entire multi epoch photometric data set of XO-3b to obtain values for $\overline{F}_{\rm ap-bg}$, the number of electrons measured in the source aperture after background subtraction (i.e., the signal); $\overline{F}_{\rm ap}$, the number of electrons in the aperture *before* background subtraction (from which we derive the noise in the aperture); and $\overline{F}_{\rm bg}$, the number of electrons in the background annulus (from which we derive the noise in the background). For real data, we obtain $\overline{F}_{\rm ap-bg} = 70858$ e, $\overline{F}_{\rm ap} = 70917$ e, and $\overline{F}_{\rm bg} = 463$ e; for the simulations, $\overline{F}_{\rm ap-bg} = 73246$ e, $\overline{F}_{\rm ap} = 73358$ e, and $\overline{F}_{\rm bg} = 881$ e.

The first term in square brackets of Equation (10), when divided by the remaining three terms, gives the relative contribution of readout noise to σ_e^2. For our integration parameters, this term equals $21.9/F_e$, which is much less than $1 - 2({\rm FN})/(3n_{\rm max}) + 1/[6({\rm FN})\cdot n_{\rm max}] = 0.98$ if $F_e \gg 22.3$ e. Thus, readnoise is insignificant for XO-3b, where $F_e \sim 70{,}000$ e.

Substituting $\overline{F}_{\rm ap}$ and $\overline{F}_{\rm bg}$ into Equation (10) (using $\sigma_{\rm int}^2 = F(t_{\rm int}/t_{\rm exp})$) yields noise variances for the aperture and background, $\sigma_{\rm ap}^2$ and $\sigma_{\rm bg}^2$. Their sum equals the noise variance for an aperture photometry measurement: $\sigma_{\rm ap-bg}^2 = \sigma_{\rm ap}^2 + \sigma_{\rm bg}^2$. We obtain $\sigma_{\rm ap-bg} = 268$ e for both real and simulated data. Dividing these into $\overline{F}_{\rm ap-bg}$ gives the expected S/N for a single photometric data point: $({\rm S/N})_{\rm single}^{\rm real} = 264$ and $({\rm S/N})_{\rm single}^{\rm sim} = 268$. These numbers are extremely close to the square roots of the background-subtracted aperture fluxes, which means that neither the backgrounds nor readout noise are significant determinants of S/N for XO-3b. From this point on, we refer to the intrinsic variability as *photon* noise.

We now propagate the expected photon noise error in a single photometric measurement to that for the entire eclipse depth measurement. Recall Equation (4) for the eclipse depth, which can be rewritten:

$$D = \frac{F_{\rm out}}{F_{\rm in}} - 1. \qquad (11)$$

The photon noise variance in the eclipse depth is the variance in $F_{\rm out}/F_{\rm in}$:

$$\sigma_{\rm phot}^2 = (1 + D)^2\left[\left(\frac{\sigma_{\rm out}}{F_{\rm out}}\right)^2 + \left(\frac{\sigma_{\rm in}}{F_{\rm in}}\right)^2\right]. \qquad (12)$$

Since $F_{\rm out}$ and $F_{\rm in}$ are the average fluxes inside and outside eclipse, we have

$$\left(\frac{\sigma_{\rm in}}{F_{\rm in}}\right)^2 = \frac{1}{N_{\rm in}}\left(\frac{\sigma_{\rm single}}{F_{\rm single}}\right)^2 \qquad (13)$$

and similarly for the out-of-eclipse flux. We define $N_{\rm in}$ and $N_{\rm out}$ as the total number of frames in and out of eclipse. Let $N_{\rm in} = f_{\rm in}\ N$, where the total number of measured frames is $N = 14{,}912$ (real) and 15,232 (simulated). Keep in mind that the flux outside of eclipse, $F_{\rm out}$, is a factor $D + 1$ larger than $F_{\rm in}$. Also, substitute $(\sigma_{\rm single}/F_{\rm single})^2 = 1/(S/N)_{\rm single}^2$.

The photon noise variance in the eclipse depth consequently becomes

$$\sigma_{\rm phot}^2 = \frac{(1 + D)^2}{N(S/N)_{\rm single}^2}\left[\frac{1}{(1 - f_{\rm in})(1 + D)} + \frac{1}{f_{\rm in}}\right]. \qquad (14)$$

If we use $f_{\rm in} = 1/3$ and assume eclipse depths of $D_{\rm real} = 1520$ ppm (average measured value) and $D_{\rm sim} = 1875$ ppm (actual input value), we find that the expected variability in the eclipse depth due to photon noise is $\sigma_{\rm phot} = 53$ ppm, for both real and simulated data.

3.4. Repeatability, Reliability, and Accuracy

A substantial literature exists in other scientific fields discussing techniques for estimating the repeatability, reliability, and accuracy of a set of measurements (see, for example, Altman & Bland 1983; Bartlett & Frost 2008, for discussions of repeatability and reliability). We review and adapt these terms below.

3.4.1. Repeatability

We define the repeatability, R, to be the value below which we can expect the difference between two eclipse depth measurements to lie 68% of the time, for a given data reduction method. For our purposes, R equals the SD of the differences in separate measurements made with the same method, $R = {\rm SD}(\Delta_{ij})$. Repeatability has the same units as the measurements themselves (e.g., ppm). Note that the repeatability is not the SD of the measurements, which indicates the spread in depths around the mean value, but of their differences.

One way to assess repeatability visually is with "mean/difference" plots (Altman & Bland 1983), which we show in Figures 10 (real) and 11 (simulated). The plots display, for all pairs of measured eclipses, the difference in depth (Δ_{ij}) as a function of the pair average eclipse depth. (To obtain a statistically valid estimate for this comparison, each pair must be counted twice, with the order of the indices reversed.) Mean/difference plots often show more clearly the limits of variability of the difference between sets than, for example, correlation plots where the variables are plotted against each other. In mean/difference plots, the horizontal spread of the data (spread in average values in paired epochs) is related to the precision of the measurements (when the overall scatter in values is large, the midpoint between pairs of values will have a relatively large spread). The vertical spread in mean/difference plots indicates the repeatability, i.e., how far apart we expect two separate measurements to be. Specifically, we compute R from the SD of each group of paired differences, labeled "SD" on the bottom left of each frame of Figures 10 and 11.

Patterns in mean/difference plots can sometimes elucidate patterns in the data, but they need to be examined carefully because of the inherent correlation between the data axes. If we define $x \equiv (D_1 + D_2)/2$ (the horizontal axis) and $y \equiv (D_1 - D_2)$ (the vertical axis), then it is apparent that the

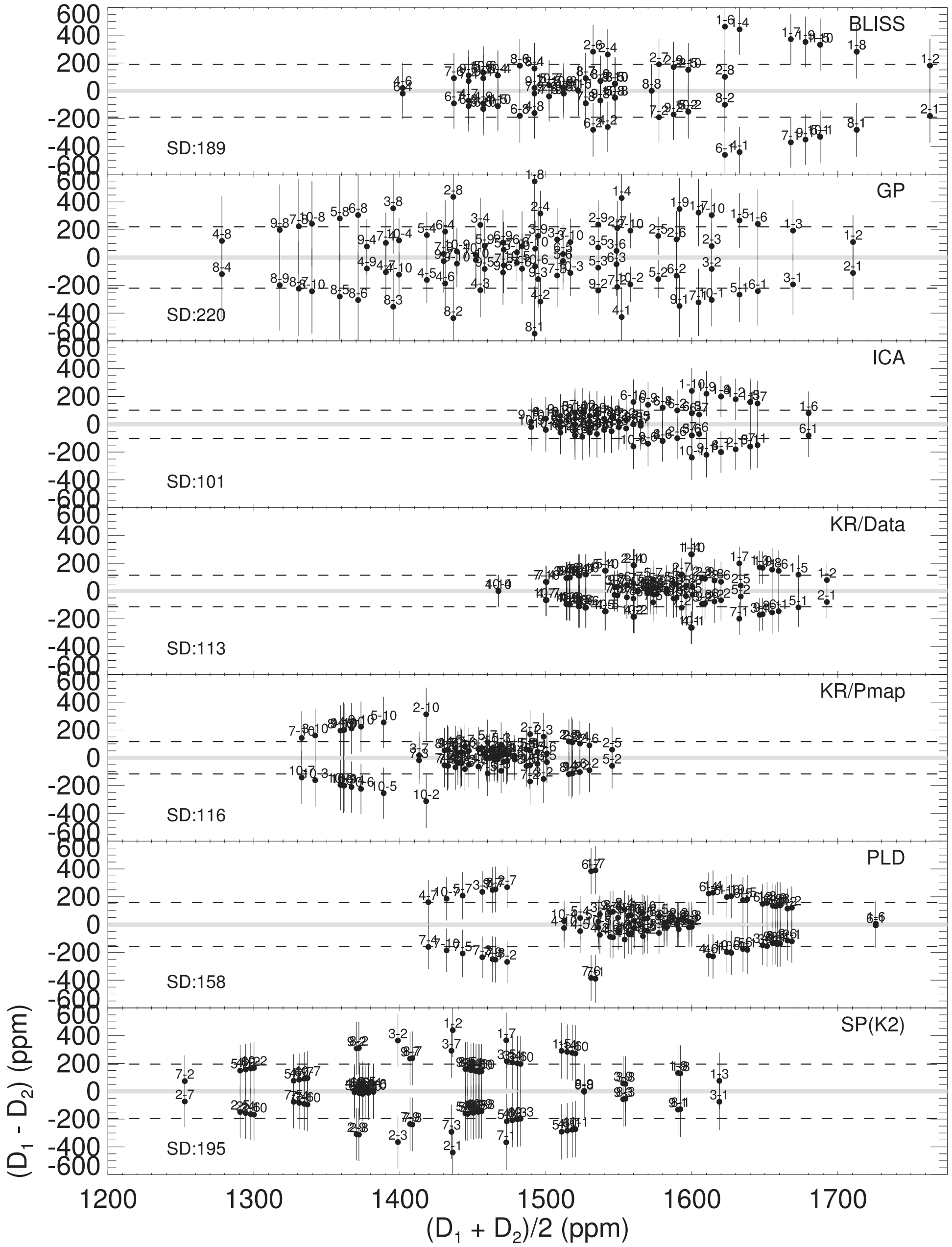

Figure 10. Mean/difference plots for repeated visits to XO-3b, real data. Each panel shows the difference between all pairs of eclipse depth measurements for a given reduction method, as a function of the average of the two depths. Each point is labeled with the two epochs being compared. Two horizontal dashed lines indicate ± one standard deviation of the differences (repeatability), also labeled in the lower left corner of the panel. A gray line indicates $(D_1 - D_2) = 0$. The horizontal spread of the data relates to the precision of the set of measured depths, whereas the vertical spread indicates their repeatability.

two axes are not independent: the relationship between y and x can be written either $y = 2(D_1 - x)$ or $y = 2(x + D_2)$. Thus, for a given D_1 or D_2, the inter epoch difference (y) is expected to follow a linear trend as a function of the inter epoch average (x). This trend is indeed visible in Figures 10 and 11 if we group by epoch. It is most visible when either (1) D_1 or D_2 is significantly different from the average depth, or (2) the inter epoch average, x, has a large spread. For example, in the real data four of the methods (BLISS, GP, ICA, KR/Data, and PLD) show an inverse linear relationship between x and y for paired differences involving epoch 1 (labeled "1–2," "1–3," etc.). This is because the epoch 1 depth is systematically high for each of these methods (as one might also guess from Figure 8).

The values of R for each technique are listed in column 8 of Tables 3 and 4. The real XO-3b results show a repeatability of better than 220 ppm in all cases, with an average value of $\overline{R} = 156$ ppm. The simulations are less repeatable, with $\overline{R} = 252$ ppm. This is probably due to the presence of more noise in the eclipse depth measurements for the simulations, as expected from the greater pointing scatter (Section 3.1). To confirm that the repeatability as computed is consistent with our definition above, we have constructed cumulative distributions of each set of eclipse depth differences. As expected, at least 68% of measured differences for nearly all techniques are less than R for both real and simulated data. In only a few cases, the 68th percentile is as much as 20 ppm larger than R.

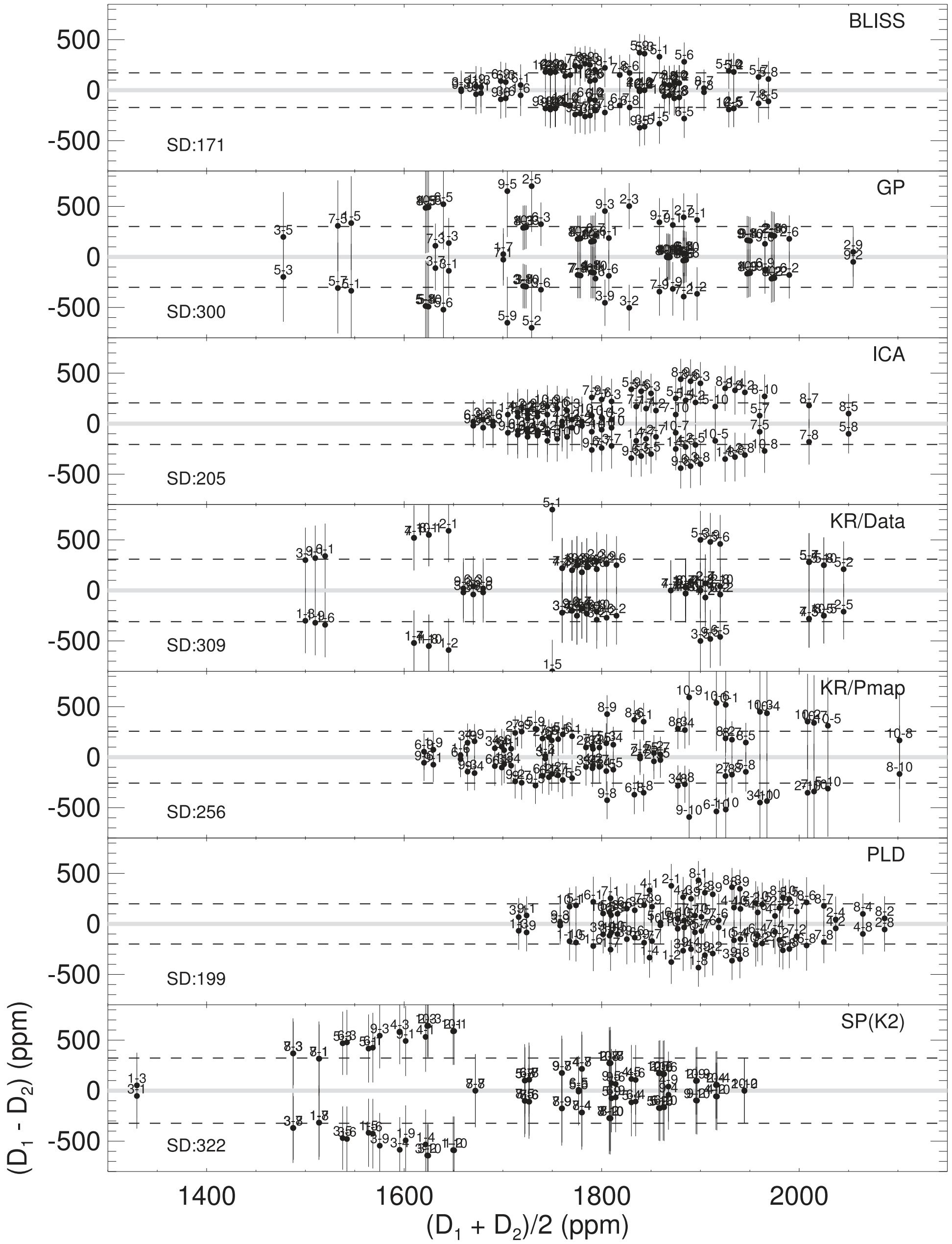

Figure 11. Mean/difference plots for repeated visits to XO-3b, simulated data. See caption to Figure 10 for further description.

Strictly speaking, in earthbound experiments repeatability is usually assessed on consecutive measurements under *identical* conditions.[12] This is not possible for eclipses, since they cannot be repeated at will. In the time between eclipses, the experimental situation will likely change: a new pointing center and different pointing jitter can change the correlated noise properties; exposure of the detector arrays to other sources of photons may produce latent charge on the pixels of interest, or existing latent charge may decay; the planetary phase curve and eclipse timing and depth may not be the same from one orbit to the next due to stellar variability, perturbations of the planet's orbit, or atmospheric evolution. However, for consistency with the astrophysics literature, we will continue to refer to the spread in eclipse depth differences as repeatability.

3.4.2. Reliability

We define the *reliability*, r, to be the ratio between the intrinsic variability of a set of measurements (in the absence of astrophysical variation) to their observed variability, for a given method. In the context of eclipse depth measurements, the intrinsic variability is the SD in the depth due only to photon noise, σ_{phot} (Equation (14)), and the observed variability is the measured SD in the depth. The measured variance combines both the photon noise and the variance due to "measurement error," caused by residual correlated noise. (Here we assume no variability in the planetary system, but its presence would add to the *measured* variance and decrease the reliability.) The value of $r \equiv \sigma_{\mathrm{phot}}/(\mathrm{SD})$ is unitless and can range from 0 (all

[12] The measurement of differences under changing conditions is often called *reproducibility*.

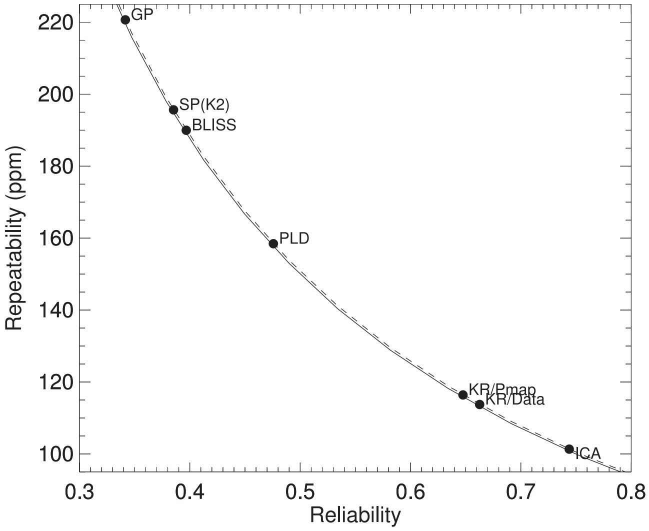

Figure 12. Repeatability as a function of reliability, for the real XO-3b eclipse depth measurements. The dashed curve displays the fit $R = 75.4\ r^{-1}$ ppm, and the solid curve shows the expected behavior $R = \sqrt{2}\,\sigma_{\rm phot}\, r^{-1} = 75\ r^{-1}$.

scatter due to measurement error) to 1 (no measurement error). Reliability is essentially a normalized measure of precision, and is inversely related to repeatability (we demonstrate this relationship below).

We list the computed values of r for each method in column 9 of Tables 3 and 4. For the real data, the reliability is quite high in most cases, with an average of $\bar{r} = 0.52$, suggesting that half of the scatter is due to intrinsic photon noise. The ICA and kernel regression (KR/Data and KR/Pmap) techniques appear to have the least amounts of correlated noise (scatter in eclipse depths consistent with more than half photon noise). For the simulated data, however, the values are lower, with an average reliability of $\bar{r} = 0.32$.

Figures 12 and 13 are scatterplots of repeatability versus reliability for the real and simulated eclipse depths, respectively. These data appear inversely correlated, which is not surprising. If two values are drawn from the same parent population, then the variance in the difference between the values should be twice the variance of the original distribution, which means that for large enough samples $[{\rm SD}(\Delta_{ij})]^2 \;=\; 2({\rm SD})^2$. Thus by the definition of r, we expect $R = \sqrt{2}\,\sigma_{\rm phot}\, r^{-1}$. We overlay this theoretical curve, as well as linear fits to R as a function of r^{-1} on Figures 12 and 13. The two curves for each plot are practically identical, with the fit factors multiplying r^{-1} within 1% of the theoretical values, indicating statistical self-consistency between $[{\rm SD}(\Delta_{ij})]$ and SD. This implies that the repeatability and reliability derived from 10 element samples are robust.

Figure 14 plots the reliability for simulated data as a function of that for real data, for the seven decorrelation methods, with lines of different slope overlaid. There seems to be no relationship between the reliability measures for real and simulated eclipses, except that the simulated values are nearly all lower than their real counterparts. Only BLISS has a similar reliability for both real and simulated data ($r = 0.40$ and 0.44, respectively). The kernel regression techniques both show the largest decrease, with $r_{\rm sim} \approx 0.4\ r_{\rm real}$. We conclude that BLISS is most robust to increases in positional dispersion, the main source of additional correlated noise between the simulated and real data sets. The (Gaussian) kernel regression methods seem to be least robust to such changes.

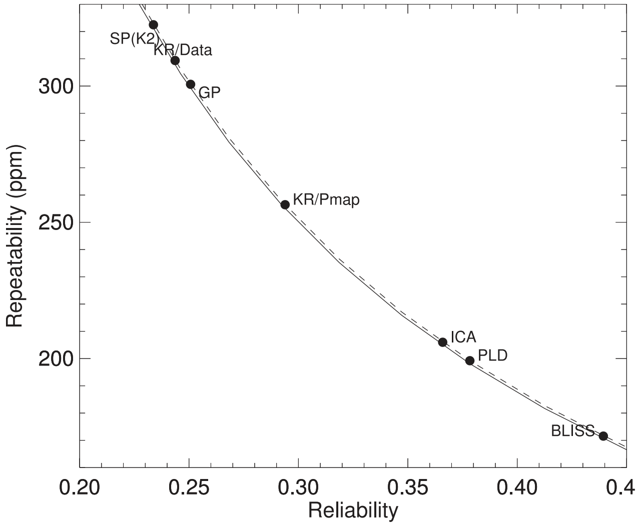

Figure 13. Repeatability as a function of reliability, for the simulated data. The dashed curve displays the fit $R = (75.4\ \ r^{-1})$ppm, and the solid curve shows the theoretical behavior, $R = 75\ \ r^{-1}$.

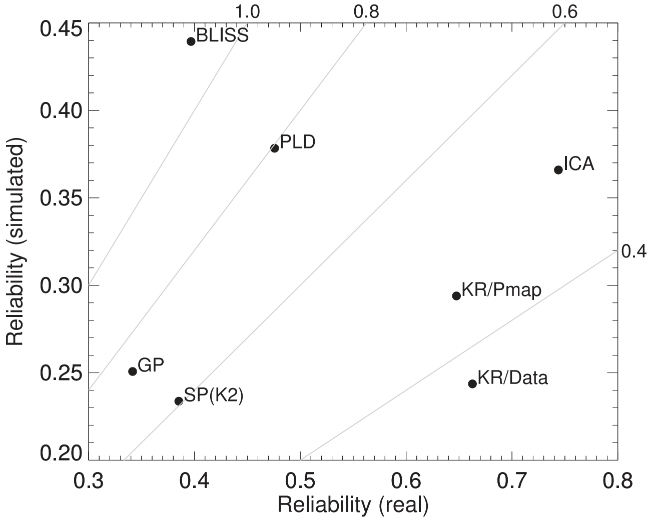

Figure 14. Reliability comparison between simulated and real eclipse depths. Gray lines indicate $r_{\rm sim}/r_{\rm real} = 0.4$, 0.6, 0.8, and 1.0.

3.4.3. Accuracy

The *accuracy* of a technique is a quantitative estimate of how well the technique measures a given characteristic of a system. Earlier definitions of accuracy were synonymous with what is now called *trueness*, the proximity of the *mean* of a set of measurements to the true value. Current definitions of accuracy, however, encompass both random *and* systematic error. That is, accuracy is limited by precision.[13] Even if the mean of a set of measurements is extremely close to the truth (bias is low and trueness is high), if the reliability (precision) is low (the scatter in results is large), the result is still considered to have low accuracy.

Assume an exoplanet system is observed N times, and a given technique j yields a set of measurements of the eclipse

[13] *ISO*5725-1: 1994, “Accuracy (trueness and precision) of measurement methods and results.”

depth, $\{D_{ij}\}$ ($i = 1, \ldots, N$), with average value, $\overline{D_j}$. Let the true depth be D_t. We can think of a measurement of eclipse depth as being the sum of the true value, any bias in that measurement (systematic error), B_{ij}, and two random noise terms:

$$D_{ij} = D_t + B_{ij} + \epsilon_{ij}^{\rm phot} + \epsilon_{ij}^{\rm meas}. \tag{15}$$

Here, $\epsilon_{ij}^{\rm phot}$ is the error in measurement ij due to photon noise and $\epsilon_{ij}^{\rm meas}$ is the random measurement error (e.g., a random component of residual correlated noise). These error terms can be thought of as samples of random variables with means of 0 and SD equal to $\sigma_{\rm phot}$ and $\sigma_{\rm meas}$. Taking the mean of D_{ij} gives

$$\overline{D} = D_t + \overline{B}. \tag{16}$$

Thus the average measured value is approximately the sum of the true value and the average bias. Alternately, if we know D_t (as we do for the simulations), we can estimate the mean bias as

$$\overline{B} = \overline{D} - D_t. \tag{17}$$

The scatter in the data about the true value is measured by the mean square error:

$$\begin{aligned} {\rm MSE} &= \frac{1}{N}\sum_{i=1}^{N}(D_{ij} - D_t)^2. \\ &\approx \overline{(B^2)} + \sigma_{\rm phot}^2 + \sigma_{\rm meas}^2. \end{aligned} \tag{18}$$

We now define accuracy using the square root of MSE, analagous to using SD for reliability:

$$a \equiv \sigma_{\rm phot}/{\rm RMSE}. \tag{19}$$

This has the desired limiting behavior: if the bias is minimized ($\overline{B} \rightarrow 0$; $\overline{D} \rightarrow D_t$), MSE approaches $\sigma_{\rm phot}^2 + \sigma_{\rm meas}^2 = ({\rm SD})^2$ and the accuracy approaches the reliability; but as the bias increases, $a \rightarrow 0$.

Columns 11–13 of Table 4 list the root mean square error (RMSE), the average bias, $\overline{B}$, and the accuracy of each technique applied to the simulated XO-3b eclipses.

Figure 15 plots a as a function of r. This figure shows how well a technique (1) can be relied on to give the *same* eclipse depth over multiple epochs where the true depth is constant (reliability: ratio of intrinsic to measured scatter: bottom axis); and (2) can be expected to give the *correct* eclipse depth over multiple epochs (accuracy: ratio of intrinsic to measured error: left axis). It is better to be on the upper right of the plot (lower scatter, lower error) than on the lower left.

The majority of the methods have RMSE values similar to their SD values, and thus accuracy nearly equal to reliability. We plot a fit to the data in Figure 15, $a = 1.02\, r - 0.02$, which confirms that *on average* the limiting value $a \approx r$ is reached for these techniques. In other words, the bias is within one SD of zero.

In detail the ratio a/r, which equals (SD)/RMSE, is not unity but varies by 20% among the techniques. We display a/r as a function of mean absolute bias in Figure 16. The ratio is (roughly) inversely proportional to $|B|$. This can be understood theoretically if we write

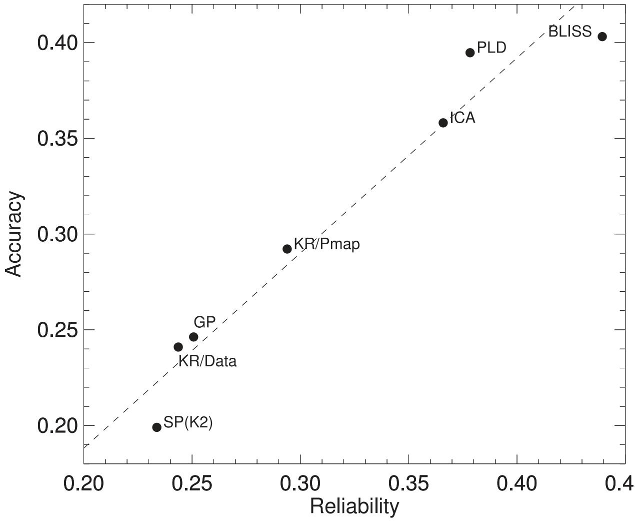

Figure 15. Accuracy vs. reliability, as defined in the text, for the simulated eclipse depth measurements. It is better to be on the upper right of the plot (lower scatter, lower error) than on the lower left. The dashed line displays the fit $a = 1.02\, r - 0.02$, which confirms that on average, the techniques have minimal bias.

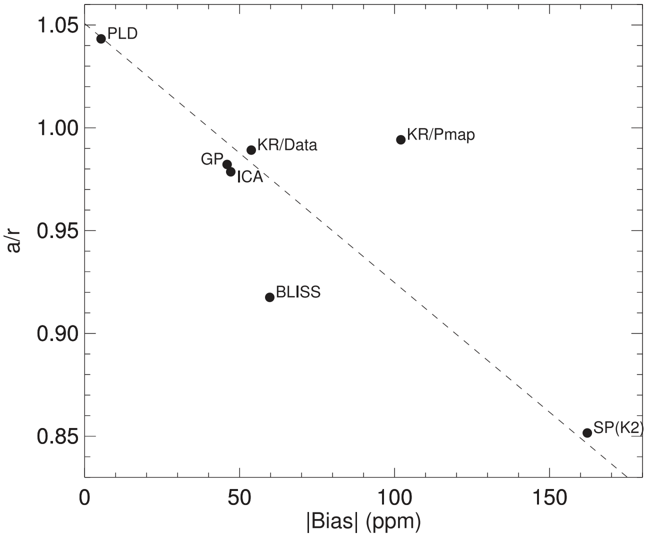

Figure 16. The accuracy/reliability ratio as a function of mean absolute bias for the simulated eclipse depth measurements. The dashed line displays the fit $a/r = 1.1 - 0.0013\, |\overline{B}|$.

$$(a/r)^2 \approx \frac{\sigma_{\rm phot}^2 + \sigma_{\rm meas}^2}{\overline{(B^2)} + \sigma_{\rm phot}^2 + \sigma_{\rm meas}^2} \tag{20}$$

$$= \left[\frac{\overline{(B^2)}}{({\rm SD})^2} + 1\right]^{-1}. \tag{21}$$

3.5. Comparison Between Methods

The repeatability, reliability, and accuracy are all measures applied to the results of a single decorrelation method. We can

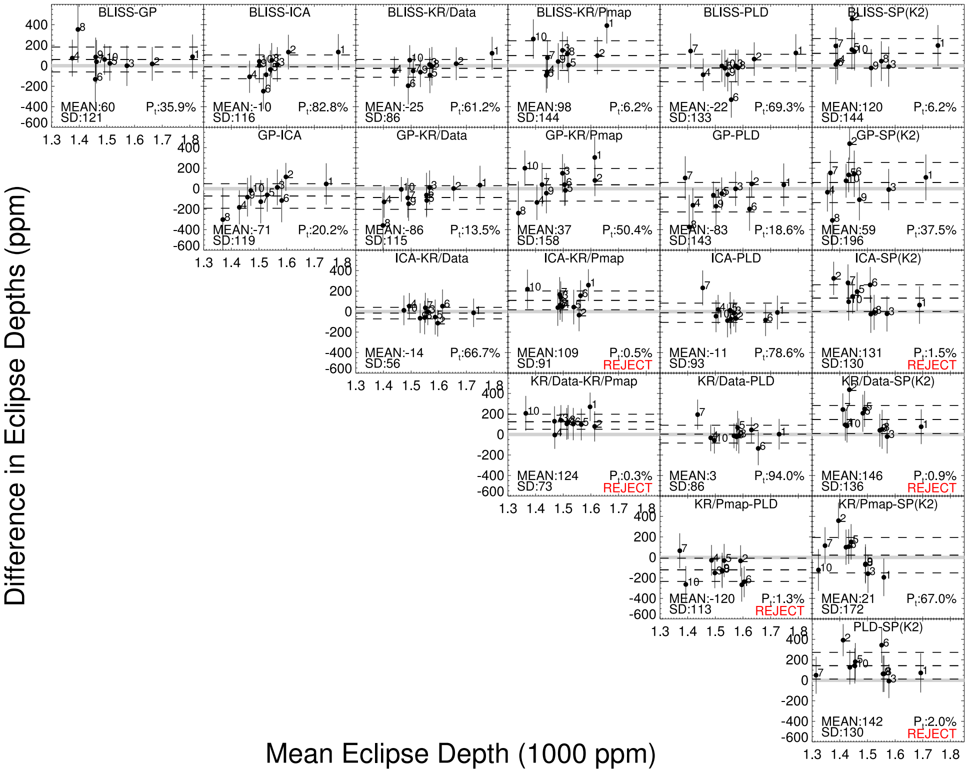

Figure 17. Mean/difference plots comparing decorrelation techniques to each other for real data. Each panel plots differences in XO-3b eclipse depths for each epoch for the two techniques given, as a function of the mean depth for the epoch of the pair of techniques. The epoch is labeled on each point. Three horizontal dashed lines display the mean difference, or relative bias between the methods, ± one standard deviation, and the bottom left of each panel prints these numbers. Horizontal gray lines indicate zero difference. The bottom right of each panel displays the *t*-test *p* value, giving the probability that the *t* parameter is larger than the measured value if the null hypothesis is true (see text). If $p < 5\%$, then the null hypothesis is rejected, which we take to mean that the two techniques are not measuring the same mean eclipse depth.

also conduct a more direct comparison of methods. First, we use the mean/difference plotting method of Altman & Bland (1983) to make a visual comparison. Figures 17 (real) and 18 display these plots for each pair of methods. Dashed lines show the mean of the differences, $\overline{\Delta}$, which estimate the relative bias between techniques; and $\pm\mathrm{SD}(\Delta)$, which bounds the limits of variability. Column 10 of Tables 3 and 4 list the method that gives the closest match to each of the methods of Column 1. This was chosen as the method giving the smallest range of eclipse values, $\min(|\overline{\Delta}| + \mathrm{SD}(\Delta))$.

Another way of comparing two approaches is to use the Student's t-test to assess whether the results are drawn from a distribution with the same mean. The test posits the null hypothesis that both sets of data have the same mean and attempts to reject it. We use the unpaired version of the test to compute the *t* statistic (the difference in average values divided by the combined variance) and compare with the t-distribution for the number of degrees of freedom. The bottom right corner of each panel of Figures 17 and 18 displays the probability that *t* is larger than the computed value if the null hypotheses were true. The null hypothesis is rejected if $p < 5\%$, i.e., the measured statistic is in the tail of the distribution. In all of the comparisons for simulated data and most comparisons for real data, the hypothesis is not rejected. However, for real data, both KR/Pmap and SP(K2) are likely not to have the same mean as ICA, KR/Data, or PLD.

We can also do a global comparison of methods using analysis of variance (ANOVA) F-test, which posits the null hypothesis that *all* sets of eclipse depths have the same mean. This analysis assumes that the group of eclipse depths for each method follows a normal distribution, and that each group has approximately the same variance (usually taken to be within a factor of two of each other, which our measurements satisfy). Similarly to the t-test, it computes a statistic and compares it with the expected distribution under the null hypothesis. In this case the statistic is *F*, the ratio of the average variability among groups (the dispersion of group means) to the average variability within groups (average group variance). The comparison distribution is the F-distribution (also known as the Fisher–Snedecor distribution), which gives the probability of measuring *F* for the applicable degrees of freedom, given the null hypothesis. Smaller values of *F* imply a higher probability that the groups share the same mean. For the real data $F = 2.8$, for which only $p = 1.6\%$ of the F-distribution has larger values, and so we reject the null hypothesis and conclude that not all methods have the same mean. If we remove KR/Pmap and SP(K2) from the calculation because they were the only methods that had failed t-tests, then $F = 0.9$. In this case, 45% of the distribution has larger values, and we do not reject the null hypothesis. We conclude therefore that KR/Pmap and SP(K2) eclipse depths are biased relative to the other techniques. For the simulated data $F = 0.8$ (all techniques), for which 57% of the distribution has larger values, and so we do not reject the hypothesis of equal means.

We emphasize that null hypothesis significance tests like *t* and *F* tests are limited in scope and predictive power. In particular, they only allow us to *reject* the hypotheses of equal means, but not to accept them. Their probability distributions

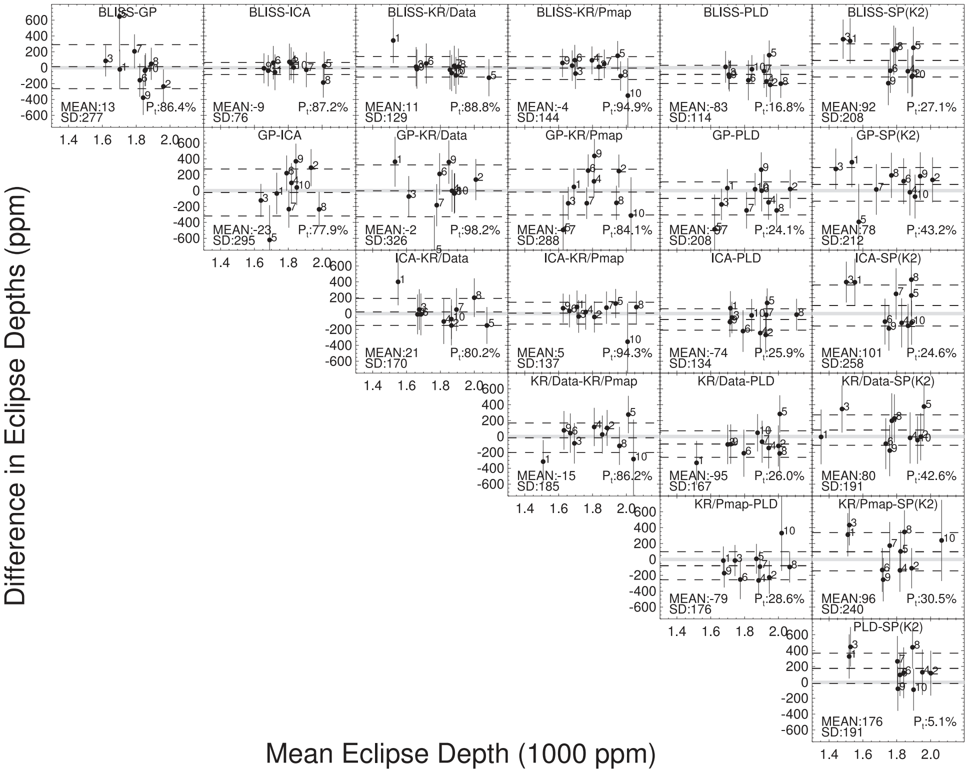

Figure 18. Mean/difference plots comparing decorrelation techniques to each other for simulated data. See caption to Figure 17 for more details.

give the probability, assuming the means are equal, that the corresponding statistic has the measured value, *not* the probability that the means are or are not equal given the measured statistic. Nevertheless, they still have value, at least as a first approach to an inter-method comparison. Bayesian estimation with Monte Carlo simulations would provide a more robust and comprehensive framework from which to analyze differences and similarities between results (e.g., Killeen 2005; Kruschke 2013), but is beyond the scope of this work.

4. DISCUSSION

4.1. Repeatability and Accuracy of IRAC Eclipse Depth Measurements

We have analyzed 10 real and 10 simulated[14] eclipses of hot Jupiter XO-3b using seven correlated noise-removal methods. The simulations were in some ways an attempt to replicate the real data, but were given larger pointing fluctuations and drifts, thereby increasing correlated noise and decreasing the positional redundancy that many noise-removal techniques rely on.

For the real data, the statistical uncertainties determined on individual eclipse depths accurately describe the scatter in eclipse depths over the 10 visits. In only one case, BLISS mapping, did the uncertainty need to be increased by 50%. For the simulations, all techniques except BLISS mapping required an increase of 20%–60%, implying that the methods may need slight adjustment to allow individual uncertainties to track the increased pointing fluctuations.

We defined three terms relating to measurement stability: repeatability (R), the expected difference between repeated measurements; reliability (r), the ratio of intrinsic (photon-limited) to measured *variability*; and accuracy (a), the ratio of intrinsic to measured *error*. Repeatability and reliability are inversely related, and reliability is a normalized estimate of precision. Accuracy combines both trueness and precision, and can theoretically never have a value less than the reliability.

For real XO-3b data, eclipse depths are repeatable within $R \lesssim 220$ ppm. In other words, any two single eclipses are expected to be within 220 ppm of each other 68% of the time. The most repeatable techniques have $R \leqslant 116$ ppm, which is about 1.5 times the photon limit ($\sqrt{2}\,\sigma_{\rm phot} \approx 75$). For the synthetic data, the repeatability is somewhat larger, $R \lesssim 300$ ppm.

When comparing the scatter in eclipse depths with the intrinsic uncertainty due to photon noise, all techniques come within a factor of 3 of the photon limit for the real data (reliability $r > 0.33$). ICA and the kernel regression techniques (KR/Data and KR/Pmap) exhibit a scatter consistent with more than two-thirds photon noise ($r \gtrsim 0.65$). For the simulations, the eclipse depth scatter is within a factor of 2–4 of the photon limit. Only the BLISS technique had the same value of reliability for the simulations as the real data, whereas the kernel regression techniques showed reductions of 40%. Even though BLISS may not be as precise as other techniques in the best circumstances, its precision is the most robust to an increasing positional spread.

The simulations afforded a unique view into the analysis of eclipses, allowing us to evaluate the accuracy and bias of each method based on knowledge of the true depth. The root mean-squared eclipse depth error ranged from 2.5 to 5 times the photon noise limit, yielding accuracy values ranging from 0.2

[14] This portion leaves out the SP(K2) technique, which was not developed for *Spitzer*.

to 0.4. Most techniques obtained an average eclipse depth within 60 ppm of the true depth (1875 ppm).

We stress that repeatability, reliability, and accuracy are statistics that refer to the quality of *single* measurements. To say that a technique has a reliability of r means that an individual measurement of the eclipse depth has a 68% chance of being consistent with other measurements to within $1/r$ times the photon limit (assuming Gaussian statistics). An accuracy of a means that an individual measurement has a 68% chance of being within $1/a$ times the photon limit of the true value. Techniques that give lower values of these quantities may nevertheless be extremely accurate when the results are averaged over multiple epochs. For example, PLD ($r = 0.38$ for simulated data) has larger overall scatter in individual measurements than BLISS ($r = 0.44$), but because PLD has a much lower bias than BLISS (5 versus –59, when averaged over 10 visits), both techniques have similar values of RMSE and are thus considered equally accurate ($a = 0.39$ for PLD and 0.40 for BLISS).

4.2. Is there a "Best" IRAC Correlated Noise-removal Technique?

After examining the results of processing the 2 × 10 data sets with seven different techniques for data reduction and eclipse depth measurement, we can make some tentative statements about the relative merits of the methods.

1. When the pointing fluctuations are at a normal level, ICA and the kernel regression techniques (KR/Data, KR/Pmap) return repeatability that is within a factor of ~1.5 of the photon limit ($r_{\mathrm{real}} \geqslant 0.65$), followed by PLD with $r_{\mathrm{real}} \sim 0.5$, BLISS and SP(K2) with $r_{\mathrm{real}} \sim 0.4$, and GP with $r_{\mathrm{real}} \sim 0.3$. (Here we have used inverse reliability as a normalized proxy for repeatability—see Section 3.4.2.)
2. BLISS is the most precise of all methods when the pointing fluctuations are larger ($r_{\mathrm{sim}} \sim 0.4$).
3. The precision of BLISS is the most robust to changes in the pointing fluctuations and drift ($r_{\mathrm{real}} = r_{\mathrm{sim}}$).
4. BLISS, PLD, and ICA are the most accurate and the most reliable (both a and $r \sim 0.4$), at least when pointing fluctuations are larger (simulated data).
5. PLD (with a quadratic phase curve model) yields the least biased results of all methods (however, all other methods used flat or linear phase curves—see below).
6. KR/Pmap and SP(K2), both of which did not include phase curve variations in their eclipse fits, return eclipse depths that are strongly biased, for both real data (they are not consistent with having the same mean as the other methods) and simulated data (their measured average biases are more than twice those of most of the other methods).

We emphasize that we have not separately controlled for centroiding, photometry, correlated noise removal, or eclipse depth fitting. In comparing techniques above, we are really comparing the entire data reduction pipelines that go along with each method. In particular, the out-of-eclipse phase curve model can significantly bias the measured eclipse depth. The simulated eclipses have nonlinear time-dependent phase variations that are concave downward (see Figure 1). Therefore, one expects eclipse fits using a linear (BLISS, GP, ICA, KR/Data) or flat (KR/Pmap, SP(K2)) phase curve to yield a center-of-occultation flux that is lower than the truth, as indeed seems to be the case.

We can calculate the *true* eclipse depth bias due to the phase curve model from the (noiseless) input light curve, L(t) (Appendix A.2), by fitting various phase models to the flux outside occultation and measuring the depth. For the XO-3b simulation shown in Figure 1, the fit eclipse depth is biased by −51 ppm for a flat phase model, −27 ppm for a linear model, and −2 ppm for a quadratic model. Not including SP(K2), these values account for approximately 50% of the measured average biases (Table 4, column 12). Given that the uncertainties in the mean depths (Table 4, column 7) have similar magnitudes to the biases, a larger ensemble of measurements would be necessary to make any definite claims regarding bias. Nevertheless, much of the *true* bias for the methods that used a flat or linear phase model would have been reduced dramatically by a quadratic phase curve. In the BLISS processing, quadratic and sinusoidal models were tried but were not favored by the Bayesian Information Criterion (BIC; Equation (43)). The only method whose reported depths are based on a quadratic phase curve, PLD, yields a relatively low positive bias (+5 ppm), which is consistent, within its uncertainty, with the expected true bias of −2 ppm.

This leads to the question: given that phase variations are expected to be nonlinear (if they exist at all, they are usually periodic), how should we interpret the BIC when it favors linearity? The BIC often helps minimize free parameters and ensure that models are generalizable among similar data sets; but it also is known to underfit (Dziak et al. 2012), not allowing for sufficient variability and sometimes leading to biased results. Another quantitive model selection technique, the Akaike Information Criterion (AIC) tends to *overfit* data (allow for too many free parameters) and therefore be too tied to the specifics of a given data set. One approach, suggested by Dziak et al. (2012) would be to select the best models according to both the BIC and AIC, and bracket a *range* of model sizes, instead of specifying definitively one model as the "best." In the end, model selection still requires human judgement to balance quantitative criteria such as the AIC and the BIC with reasonable expectations based on theory.

4.3. Are IRAC Eclipse Depth Uncertainties Underestimated?

A recent study by Hansen et al. (2014) derived systematic uncertainties for IRAC eclipse depths. They compared 10 two epoch pairs of *Spitzer* eclipse depth measurements for six different planetary systems, each epoch measured by different teams, including measurements from three IRAC wavelength bands and one MIPS band, as well as IRAC data taken using both dithers and staring mode. They estimated the systematic variance in each depth from the squared difference in eclipse depth values between epochs, minus the sum of reported variances (squared uncertainties) for each epoch. This is equivalent to our estimate of f_{dis} (Section 3.2), but for a sample size of $N = 2$ instead of 10. In 5 out of 10 comparisons the difference between epochs was larger than the reported uncertainty by more than a factor of 2. Combining results across data analysis methods, planetary systems, IRAC wavelength bands, and from both staring and dithering mode, they concluded that in general single eclipse measurements made with *Spitzer*/IRAC either have an uncertainty floor of 500 ppm, or that their uncertainties should be multiplied by a factor of $f_{\mathrm{dis}} = 3$. They used their inflated uncertainties to assert that features seen in broadband spectra are more likely due to instrumental systematics than molecular bands.

Following this, some authors have echoed the conclusions of Hansen et al. (2014). For example, Schwartz & Cowan (2015) obtained theoretical estimates on the properties of 50 exoplanet atmospheres after first assuming that many of the reported *Spitzer* eclipse depth uncertainties were underestimated by a factor of 3. Most recently, a general review on the observation of exoplanet atmospheres (Crossfield 2015) also accepted the Hansen et al. (2014) assertion regarding overestimated *Spitzer* precision, stating that, "it is debatable whether broadband photometry usefully determines atmospheric abundances in *any* transiting exoplanets (emphasis added)." If this statement were true, many recent analyses using modern reduction techniques and realistic (but not inflated) uncertainties would be invalidated. For some examples, see the Wong et al. (2016) claims regarding high-altitude silicate clouds in WASP-19b and enhanced C/O ratio in HAT-P-7b; or the Sing et al. (2016) categorization of the atmospheres of 10 hot Jupiters from clear to cloudy using *HST* and *Spitzer* data.

Our conclusions contradict those of Hansen et al. (2014). To avoid the influence of confounding variables that affect measurement stability, the present paper focuses on a single planetary system, using data from a single IRAC band and single observing mode (staring mode), and involves a parallel analysis isolating different correlated noise-removal techniques (and their associated data reduction pipelines). In contrast to the $f_{\rm dis} = 3$ estimate of Hansen et al., we have found for both real and simulated XO-3b data that the statistical uncertainties do not need to be increased by more than 50% to accomodate the scatter in data (for all decorrelation methods except SP(K2), which was created for K2 and not optimized for *Spitzer*), and in many cases no inflation was necessary. This holds even for simulated data, which had increased correlated noise and decreased spatial redundancy. Our estimates of $f_{\rm dis}$ include confidence intervals based on 10 epoch samples (column 6 of Tables 3 and 4), which vary by $\sim\pm 1/3 f_{\rm dis}$. As emphasized by Lyons (1992), the uncertainty on $f_{\rm dis}$ for $N = 2$ (the sample size used by Hansen et al. 2014) is much larger, up to a few times the actual value of $f_{\rm dis}$.

The chief source of the discrepancies between separate eclipse depth measurements examined by Hansen et al. (2014) is the evolution in both observing and data reduction strategies that has occurred to accomodate exoplanet observation. One key example of non-repeatability of IRAC eclipse depths cited by Hansen et al. is the 4.5 μm measurement for HD 209458b. An early study of this hot Jupiter used broadband *Spitzer* secondary eclipse spectra from 3.6 to 24 μm to infer the existence of an atmospheric inversion layer in the planet (Knutson et al. 2008). These 2005 measurements were among the earliest eclipse observations made with IRAC, and were obtained using the (then) standard practice of alternating exposures between each IRAC channel, which required a repointing every 4 × 64 subarray images. When *Spitzer* is commanded to continuously observe an inertially fixed target ("staring" mode), a source's position will fluctuate over a region of about 0.08 px diameter in one hour, while also incurring a slow linear drift of about 0.01 px per hour. Experience shows that this usually yields sufficient redundancy in source position to decorrelate intra-pixel gain in a set of photometric measurements. On the other hand, *Spitzer's* blind repointing accuracy is much worse: about 0.3 px rms. It is not surprising, then, that the 2005 measurements of HD 209458b, which were repointed every 256 frames, yielded large discontinuities in the target position, making it extremely difficult to decorrelate the data at 3.6 and 4.5 μm and extract accurate eclipse depths (especially using a low-order polynomial fit to the intra-pixel gain, as was the common practice). Subsequent measurements of the full phase curve of HD 209458b (by a team that included two of the three authors on the earlier study) were taken in continuous staring mode with no repointing, and the data were decorrelated using kernel regression as a function of x, y, and noise pixels (Zellem et al. 2014). The new methodology resulted in a 35% lower 4.5 μm eclipse depth that did not require an atmospheric temperature inversion.

Hansen et al. (2014, Table 2) use the difference between the 4.5 μm eclipse depth derived by Knutson et al. (2008) and that derived by Zellem et al. (2014) as a baseline estimate of the *systematic uncertainty* in *Spitzer*/IRAC measurements at 4.5 μm. This is incorrect, since it treats both approaches to measurement and reduction as equally valid, and equally indicative of the possible range in measurable eclipse depths. The 2005 IRAC measurements of HD 209458b were taken in such a way as to make the intra-pixel systematics in the InSb arrays virtually uncorrectable. In more recent years, observational practice has evolved toward a more optimal staring mode configuration, especially with the 2009 advent of PCRS Peak-Up to ensure that targets are repeatably positioned (to within 0.1 px) in a region with minimal intra-pixel gain variations (Ingalls et al. 2012). Eclipse data taken in this manner eliminate the discontinuous position jumps present in the 2005 data.

Also, the techniques for removing correlated noise have improved dramatically from the early days of low-order polynomial fitting. Even the sub-optimal 2005 measurements of HD 209458b were shown to be consistent with later measurements after reanalysis using BLISS (Diamond-Lowe et al. 2014) and GP (Evans et al. 2015). One of the criticisms made by Hansen et al. was that reported uncertainties for published eclipse depths were unrealistic and did not sufficiently take systematics into account. We agree that early methods did not adequately estimate the errors, but this is not a problem in most of the newer approaches, as seen in the current paper.

In his review of the study of exoplanet atmospheres, Burrows (2014) pointed out that observers and theorists have tended to overinterpret the earliest measurements. The article is a sobering reminder that results from a young field may be overturned by improved approaches to observation, reduction, and theory. The decrease in *Spitzer*/IRAC correlated noise due to staring mode and PCRS Peak-Up, as well as the improved understanding of systematics and development of better decorrelation techniques, have led to a situation in which the variations in eclipse depths described in Hansen et al. (2014) are now outliers when compared to variations observed today. Hansen et al. (2014) is a watershed work that attempted to quantify the uncertainties in *Spitzer* single exoplanet eclipse depths hinted at by Burrows (2014), via comparisons between paired studies. However, like the earliest theoretical conclusions that were biased by outlier eclipse depth *measurements*, Hansen et al. may have been similarly biased and overinterpreted the earliest *variations* in eclipse depths.

4.4. Application to Future Space Missions

Future space missions such as *JWST* (Clampin 2008) and TESS (Ricker et al. 2015), and proposed missions such as ARIEL (Tinetti 2015) and FINESSE (Deroo et al. 2012), will have similar needs to verify the repeatability and accuracy of their eclipse and phase curve measurements. These observatories will benefit from having been designed with precision measurements of transiting exoplanets in mind, and so the instrumental systematics will not be as significant as for *Spitzer*/IRAC, where correlated noise can be as much as 2 orders of magnitude larger than eclipse depths. However, systematics will still be present in future missions: *JWST* will have similar jitter to pixel scale ratios as found in *Spitzer*/IRAC (Beichman et al. 2014), which will lead to photometric variability due to intra-pixel gain fluctuations. Furthermore, observers will demand increasingly more precise measurements as more detailed questions are asked regarding e.g., atmospheric variability. Next generation space observatories will undoubtedly be pushed to the limits of their systematic error budgets and, like *Spitzer*, require a thorough assessment of their stability and accuracy.

5. CONCLUSIONS

We have performed a *Spitzer*/IRAC repeatability analysis of 10 real and 10 simulated eclipses of XO-3b using 7 correlated noise-removal techniques. Most methods are capable of estimating accurate uncertainties on individual eclipse depths. The eclipse depth repeatability (expected difference between pairs of measurements) under normal pointing variations averages ~150 ppm, only twice the photon limit, but can worsen as the spread in target positions increases. The BLISS technique, however, is most robust to such changes. The BLISS, PLD, and ICA techniques are the most accurate and repeatable when the pointing fluctuations are larger. Future analysis might benefit from separating the phase curve model from the decorrelation technique, as it can bias eclipse depths.

A few recent publications have claimed that *Spitzer* eclipse depth uncertainties should be increased by a factor of 3. Such claims rest upon a comparison of literature estimates of varying provenance and quality, using only two epochs per target, and are not substantiated by our more controlled analysis with a larger, more uniform sample.

Although we have controlled reasonably well for most important observing variables, our conclusions are strictly valid only for the IRAC 4.5 μm array, and in the particular signal-to-noise regime of XO-3b (photon noise limit on an eclipse depth of ~50 ppm). As multi epoch *Spitzer*/IRAC measurements accumulate for a variety of exoplanet targets, the data will better support more broad-based repeatability analysis, which will constrain further the limits of variability for reduction techniques, and ultimately for the instrument itself.

Some of the lessons learned with IRAC can be usefully applied to future space missions. The high degree of repeatability demonstrated in this paper was facilitated by a careful characterization and optimization of pointing during exoplanet observations (Grillmair et al. 2012; Ingalls et al. 2012). This understanding of the systematics was greatly facilitated by a set of dedicated calibration observations. The IRAC team has also found that hosting exoplanet data workshops and engaging the active research community has led to the optimization of observing strategies and improved the quality of data greatly. This paper shows that state of the art reduction techniques do an excellent and consistent job of mitigating systematic noise. Focused data challenges could prove equally effective for future exoplanet space missions.

We thank the anonymous referee for comments that greatly improved the quality of the paper. We thank Nick Cowan for a thorough reading of an early version of the manuscript and for discussions that helped clarify the conclusions. This work is based on observations made with the *Spitzer* Space Telescope, which is operated by the Jet Propulsion Laboratory, California Institute of Technology under a contract with NASA. G.M. received support from ERC project number 617119 (ExoLights). K.S. recognizes support from the Sagan Fellowship Program, supported by NASA and administered by the NASA Exoplanet Science Institute (NExScI). This research has made use of the Exoplanet Orbit Database and the Exoplanet Data Explorer at exoplanets.org.

Facility: Spitzer(IRAC).

Software: IRACSIM (http://dx.doi.org/10.5281/zenodo.46270), IDL.

APPENDIX A
IRACSIM: AN IRAC DATA SIMULATOR FOR POINT SOURCE IMAGES

To produce the simulated XO-3b observations used for the Data Challenge, we used `IRACSIM`,[15] a package built in the IDL programming language. The program uses a model of the *Spitzer*/IRAC system to create synthetic IRAC point source measurements, outputting FITS image (or image cube) files similar to those produced by the IRAC BCD pipeline. The simulator model is built on three major components of *Spitzer*/IRAC behavior: (1) pointing, (2) imaging, and (3) Fowler sampling. We give an overview of this model here.

A.1. Pointing

The IRAC pointing model specifies the position of a point source as a function of time, $(x[t], y[t])$. The model has four main components, based on the known structure of *Spitzer* pointing variations (Grillmair et al. 2012): a high-frequency fluctuation or "jitter" with amplitude ~0.05 px; a sawtooth-shaped "wobble" due to a battery heater cycling on and off (period ~40 minutes, amplitude ~0.05 px); an approximately 30 minute initial drift of up to 0.1 px; and a long term drift of ~0.3 px per day. (See also Hora et al. 2014, Figure 8 for high-fidelity measurements of jitter, wobble, and drift.) The pointing as a function of time is given by

$$x(t) = x_j(t) + x_{\rm w}(t) + x_{\rm sd}(t) + x_{\rm ld}(t); \qquad (22)$$

$$y(t) = y_j(t) + y_{\rm w}(t) + y_{\rm sd}(t) + y_{\rm ld}(t). \qquad (23)$$

The jitter component is the sum of a sine wave plus a randomly generated $1/f$ noise:

$$x_j(t) = A_j \sin[2\pi(t - t_0)/P_j + \phi_j] \cos(\theta_j) + {\rm FBM}(A_{\rm fbm}, \beta, t) \qquad (24)$$

$$y_j(t) = A_j \sin[2\pi(t - t_0)/P_j + \phi_j] \sin(\theta_j) + {\rm FBM}(A_{\rm fbm}, \beta, t). \qquad (25)$$

[15] http://dx.doi.org/10.5281/zenodo.46270

Table 5
Pointing Model Parameter Ranges

	Parameter	Range[a]	Type
Jitter	A_j	0.04 px	C
	P_j	60 s	C
	ϕ_j	0–π rad	U
	θ_j	−45°	C
	A_fbm	0.4 px	C
	β	1	C
Wobble	A_w	0.018–0.034 px	U
	P_w	1200–2800 s	U
	ϕ_w	−1–1 rad	U
	S_w	0.1–	G
	$\Delta\ A_{w,\mathrm{max}}$	0.01 px	C
	$\Delta\ P_{w,\mathrm{max}}$	10 s	C
	θ_w	−80 to −45°	G
Short-Term Drift	A_sd	0–1 px	U
	P_sd	395.6 s	C
	ϕ_sd	$7\pi/4$ rad	C
	τ_sd	−1800–1800s	U
	θ_sd	100°	C
Long Term Drift	A_ld	0–0$''$.0208 hr^{-1}	U
	θ_ld	−95 to −55°	U

Note.
[a] Ranges give either hard limits of a uniform deviate ("U" in column 3), $\pm 1\sigma$ of a Gaussian deviate ("G" in column 3), or a constant value ("C" in column 3).

Here, A_j is the jitter amplitude; t_0 is the time of the last spacecraft pointing reset, usually via PCRS Peak-Up; P_j is the jitter period; ϕ_j is the phase shift of the jitter; and θ_j is the "axis" of the jitter, the angle on the pixel grid (with respect to the x axis) over which the sinusoidal component of jitter oscillates. The term $\mathrm{FBM}(A_\mathrm{fbm}, \beta, t)$ is a random variable representing a fractional brownian motion noise with power spectral index proportional to $f^{1/\beta}$ and having peak amplitude A_fbm, constructed according to the prescription of (Stutzki et al. 1998, Section 4).

The wobble component is modeled as a "skewed sinusoid":

$$w(t) = A_\mathrm{w}(t)\sin[2\pi(t - t_0)/P_\mathrm{w}(t) + \phi_\mathrm{w} + \phi_\mathrm{sk}(t)], \quad (26)$$

where $A_\mathrm{w}(t)$ is the amplitude of the wobble, $P_\mathrm{w}(t)$ is the period, ϕ_w is a constant phase shift, relative to t_0, and $\phi_\mathrm{sk}(t)$ is an additional phase shift that varies with time, giving w its skewed shape. Let $q(t) \equiv (t - t_0)/P_\mathrm{w}(t) + \phi_\mathrm{w}/(2\pi) (\mathrm{mod}\ 1)$. The skew phase function is

$$\phi_\mathrm{sk}(t) = \begin{cases} \pi\left(\frac{1}{2S_\mathrm{w}} - 2\right)q & : 0 \leqslant q < S_\mathrm{w} \\ \pi\left(\frac{q - S_\mathrm{w}}{1 - 2S_\mathrm{w}} - 2q + \frac{1}{2}\right) & : S_\mathrm{w} \leqslant q < 1 - S_\mathrm{w} \\ \pi\left(\frac{1}{2S_\mathrm{w}} - 2\right)(q - 1) & : 1 - S_\mathrm{w} \leqslant q < 1. \end{cases} \quad (27)$$

Here, S_w is the phase of the peak amplitude (as a fraction of the period), which defines the amount of skewness. In a normal sine wave, S_w equals $1/4$, i.e., the curve peaks when the argument of the sine equals $\pi/2$. If $0 < S_\mathrm{w} < 1/4$, then the curve has a faster than sinusoidal rise and is skewed to the left. If $1/4 < S_\mathrm{w} < 1/2$, then the curve has a slower than sinusoidal rise and is skewed to the right. Either set of S_w choices results in a smoothly varying sawtooth-like curve. Additional flexibility is enabled by a time variable wobble amplitude $A_\mathrm{w}(t)$ and period $P_\mathrm{w}(t)$, with the values varying continuously in a random walk having maximum excursions assignable via the parameters $\Delta\ A_{w,\mathrm{max}}$ and $\Delta\ P_{w,\mathrm{max}}$. One final parameter that specifies the x and y projections of the wobble is the axis, θ_w:

$$x_\mathrm{w}(t) = w(t)\cos(\theta_\mathrm{w}) \quad (28)$$

$$y_\mathrm{w}(t) = w(t)\sin(\theta_\mathrm{w}). \quad (29)$$

The short-term drift appears to have periodic and asymptotic behavior, and so we model it with a rapidly decaying sinusoid:

$$s(t) = \frac{A_\mathrm{sd}}{\sin(\phi_\mathrm{sd})}\sin[2\pi(t - t_0)/P_\mathrm{sd} + \phi_\mathrm{sd}] \times \exp[-(t - t_0)/\tau_\mathrm{sd}], \quad (30)$$

where A_sd is the "asymptotic decay," the difference between the initial ($t = t_0$) and final ($t \rightarrow \infty$) values of the function; ϕ_sd is the phase of the sinusoid; P_sd is its period; and τ_sd is the decay time. The short-term drift is projected along the axis, θ_sd, onto the pixel grid:

$$x_\mathrm{sd}(t) = s(t)\cos(\theta_\mathrm{sd}) \quad (31)$$

$$y_\mathrm{sd}(t) = s(t)\sin(\theta_\mathrm{sd}). \quad (32)$$

Finally, the long term drift is a simple linear function of time:

$$x_\mathrm{ld}(t) = A_\mathrm{ld}(t - t_0)\cos(\theta_\mathrm{ld}) \quad (33)$$

$$y_\mathrm{ld}(t) = A_\mathrm{ld}(t - t_0)\sin(\theta_\mathrm{ld}), \quad (34)$$

where A_ld is the drift rate and θ_ld is the axis of projection.

Table 5 lists the range of inputs to the pointing model used in simulating the XO-3b data. We chose not to duplicate exactly the pointing fluctuations as observed in the real data set, but attempted to simulate a range of possible *Spitzer* observing conditions, and thus a range of possible decorrelation situations. To do this, most of the parameters for a given epoch were generated randomly within the predefined ranges given.

A.2. Imaging

After using the IRAC pointing model to predict the position of a point source as a function of time, the IRAC Point Response Function (PRF) allows one to compute the image of the source at each of those positions.[16] The PRF is essentially a convolution of the optical point-spread function (PSF) and the intra-pixel response function, sampled on each of the IRAC detector arrays. There are 25 PRF image files per IRAC array, each computed for a different region of the array. The files in turn contain 5 × 5 interleaved sets of point source realizations offset $1/5$ pixel from each other. For a point source at a given ($x[t]$, $y[t]$) decimal pixel location, the image of the source at pixel $i_\mathrm{p}j_\mathrm{p}$ is made by interpolating between the 5 × 5 PRF realizations

$$I(i_\mathrm{p}, j_\mathrm{p}, t) = \mathrm{PRF}^{\mathrm{interp}}_{i_\mathrm{p}j_\mathrm{p}}(x[t], y[t]). \quad (35)$$

Since the core PRF files currently available were built from cryogenic data, we have converted them to post-cryogenic IRAC by assuming that the structure of a point source image is the same as in the cryogenic mission (the optical PSF is

[16] The core PRFs from the cryogenic mission are packaged along with IRACSIM. They can be downloaded separately at http://irsa.ipac.caltech.edu/data/SPITZER/docs/irac/calibrationfiles/psfprf/. See also the IRAC Instrument Handbook, Appendix C.1.

unchanged), but that the intra-pixel response has changed. To account for this, we scaled the 25 cryogenic PRF realizations *with respect to each other* such that aperture photometry varied according to the measured post-cryogenic photometric gain map at the same intra-pixel offsets. In addition, all PRF centers were shifted such that the center of light centroid (Equations (1) and (2)) yields the correct result at zero pixel phase.

The absolute scaling of $I\,(i_{\rm p}, j_{\rm p}, t)$ is arbitrary. We rescale it to electron flux using (1) an input desired aperture flux for the point source, $f_{\rm ap}\,(r_{\rm ap})$ (Jy), (2) an aperture radius $r_{\rm ap}$ (px) for which the flux will be obtained, (3) a normalized light curve specifying the relative flux variations, $L(t)$ and (4) a scaling relationship giving the number of photoelectrons per second in the peak image pixel, divided by the flux in a three pixel aperture, $E_{\rm peak}\,(3) \equiv \dot{e}_{\rm peak}/f_{\rm ap}\,(3)$. If $\mathrm{PRF}^{\rm peak}$ is the peak value in the set of PRF images and $a(r)$ is the aperture correction in an aperture of radius r, then the rate of photoelectron production in each pixel is

$$\dot{e}\,(i_{\rm p}, j_{\rm p}, t) = L\,(t) I\,(i_{\rm p}, j_{\rm p}, t) \frac{E_{\rm peak}\,(3)}{\mathrm{PRF}^{\rm peak}} \frac{a\,(r_{\rm ap})}{a\,(3)} f_{\rm ap}\,(r_{\rm ap}). \qquad (36)$$

A.3. Fowler Sampling

Given $\dot{e}\,(i_{\rm p}, j_{\rm p}, t)$, a function that can be evaluated at arbitrary time, we produce a simulated IRAC image by mimicking the integration and sampling properties of the IRAC electronics.

IRAC acquires data using the Fowler-sampling technique, defined by the sample time, Δt, the Fowler number, FN, and the Wait Ticks, WT (IRAC Instrument Handbook, Section 2.4). The sample time Δt is fixed at 0.2 s for full array readout and 0.01 s for subarray readout. At the beginning of an IRAC measurement, each detector (pixel) is reset. Charge is then accumulated due to photoelectron production and noise. The accumulated charge in a pixel is read out every Δt seconds, for FN “pedestal” reads, $P_i(t)$, WT “wait” samples are skipped, and FN “signal” reads, $S_i(t)$ are measured. Figure 19(a) is a schematic depiction of Fowler sampling, and its relationship with the pointing model.

Each data file contains either one 256 × 256 pixel image for full array readout mode, or 64 32 × 32 pixel images for subarray readout. Define δt as the rate at which the pointing model is sampled. The total charge accumulated in one pointing sample at time t is $e\,(i_{\rm p}, j_{\rm p}, t) = \dot{e}\,(i_{\rm p}, j_{\rm p}, t)\delta t$. To capture possible rapid fluctuations in pointing that might affect the stellar image over the integration, we let the Fowler sample time Δt be somewhat larger than δt. We typically use $\delta t = \Delta t/10$, so that there are $n = 10$ PRF realizations to be integrated per Fowler sample.

We compute $e\,(i_{\rm p}, j_{\rm p}, t)$ every δt seconds over the course of the entire integration, which lasts $t_{\rm int} = [2(\mathrm{FN}) + \mathrm{WT}]\Delta t = [2(\mathrm{FN}) + \mathrm{WT}]n\delta t$ seconds. The number of total model samples in the integration is therefore $N\mathrm{samp} = [2(\mathrm{FN}) + \mathrm{WT}]n$.

The accumulated charge is stored for *every* Fowler-sampling interval (starting at P_1 and ending at $S_{\rm FN}$), including the wait ticks for proper noise accumulation. For the kth Fowler-sampling interval, the total accumulated charge is

$$e_k^{\rm mean}\,(i_{\rm p}, j_{\rm p}) = e_{k-1}(i_{\rm p}, j_{\rm p}) + \sum_{l=1}^{n} e\,(i_{\rm p}, j_{\rm p}, t_{kl}), \qquad (37)$$

where $e_0\,(i_{\rm p}, j_{\rm p}) \equiv 0$ and t_{kl} is the time of the lth pointing model subsample of the kth Fowler-sampling interval. The superscript “mean” indicates that this is an estimate of the mean charge. The actual electron counts will vary due to counting noise, and this is modeled via a Poisson random deviate $\boldsymbol{P}$:

$$e_k\,(i_{\rm p}, j_{\rm p}) = \boldsymbol{P}\,[e_k^{\rm mean}\,(i_{\rm p}, j_{\rm p})]. \qquad (38)$$

Here $\boldsymbol{P}[\mu]$ indicates a Poisson random variable with mean μ.

We separate out Fowler pedestal and signal samples by realizing that the ith pedestal read is overall sample i, whereas the ith signal read is overall sample (FN + WT + i). We also note that each time we read the detectors, we must add readout noise, which we model as a Gaussian random variable $\boldsymbol{G}\,(\mu = 0, \sigma = \sigma_{\rm RN})$. The readout noise SD, $\sigma_{\rm RN}$, is listed in Table 2.3 of the IRAC Instrument Handbook.[17] Therefore,

$$P_i\,(i_{\rm p}, j_{\rm p}) = e_i\,(i_{\rm p}, j_{\rm p}) + \boldsymbol{G}\,(0, \sigma_{\rm RN}); \qquad (39)$$

$$S_i\,(i_{\rm p}, j_{\rm p}) = e_{(\mathrm{FN}+\mathrm{WT}+i)}(i_{\rm p}, j_{\rm p}) + \boldsymbol{G}(0, \sigma_{\rm RN}). \qquad (40)$$

The result of Fowler sampling is an image which measures the mean electron counts accumulated over the “exposure time,” $t_{\rm exp} \equiv (\mathrm{FN} + \mathrm{WT})\Delta t$, or the time between the ith pedestal and the ith signal:

$$e\,(i_{\rm p}, j_{\rm p}) = \sum_{i=1}^{\rm FN} \frac{S_i\,(i_{\rm p}, j_{\rm p}) - P_i\,(i_{\rm p}, j_{\rm p})}{\rm FN}. \qquad (41)$$

The analog-to-digital converter of the IRAC electronics measures photoelectron accumulation in terms of digital data number (DN) via the proportionality $\mathrm{DN} = e/\mathrm{GAIN}$, where $\mathrm{GAIN} \approx 3.7$. Then, the SSC data pipeline produces BCD images in units of MJy sr^{-1}:

$$\mathrm{BCD}(i_{\rm p}, j_{\rm p}) = \frac{(\mathrm{FLUXCONV}) \times e\,(i_{\rm p}, j_{\rm p})}{\mathrm{GAIN} \times t_{\rm exp}}. \qquad (42)$$

Here, FLUXCONV is the flux conversion factor between MJy sr^{-1} and DN s^{-1} derived by the SSC.[18] The IRACSIM package produces images (and image cubes, for subarray measurements) in BCD units.

A.4. Input and Output

In addition to the pointing model parameters, IRACSIM accepts the following inputs: (1) the position(s) of one or more point sources (in either celestial or pixel coordinates; if positions are given in celestial coordinates then a reference coordinate and its pixel position must also be given); (2) date and time of observation; (3) the source flux density in an aperture, $f_{\rm ap}$ (and the aperture optionally); (4) a source light curve $L(t)$; and (5) the full set of observational parameters allowable in the *Spitzer* Planning Observations Tool (Spot). (For example: the instrument channel number, frame time, number of repeats, full or subarray readout).

The output of the program is a facsimile of the output of a real *Spitzer*/IRAC observation: a set of BCD image files and uncertainty files, with realistic FITS headers containing standard time and astrometry information that is correct for the simulated observation. We also add history items, comments, and new keywords that are specific to the simulation. For example, the mean pixel location of the target throughout the integration is printed in the header.

[17] http://irsa.ipac.caltech.edu/data/SPITZER/docs/irac/iracinstrumenthandbook/7/

[18] See http://irsa.ipac.caltech.edu/data/SPITZER/docs/irac/warmimgcharacteristics/ for the values of FLUXCONV for the InSb arrays.

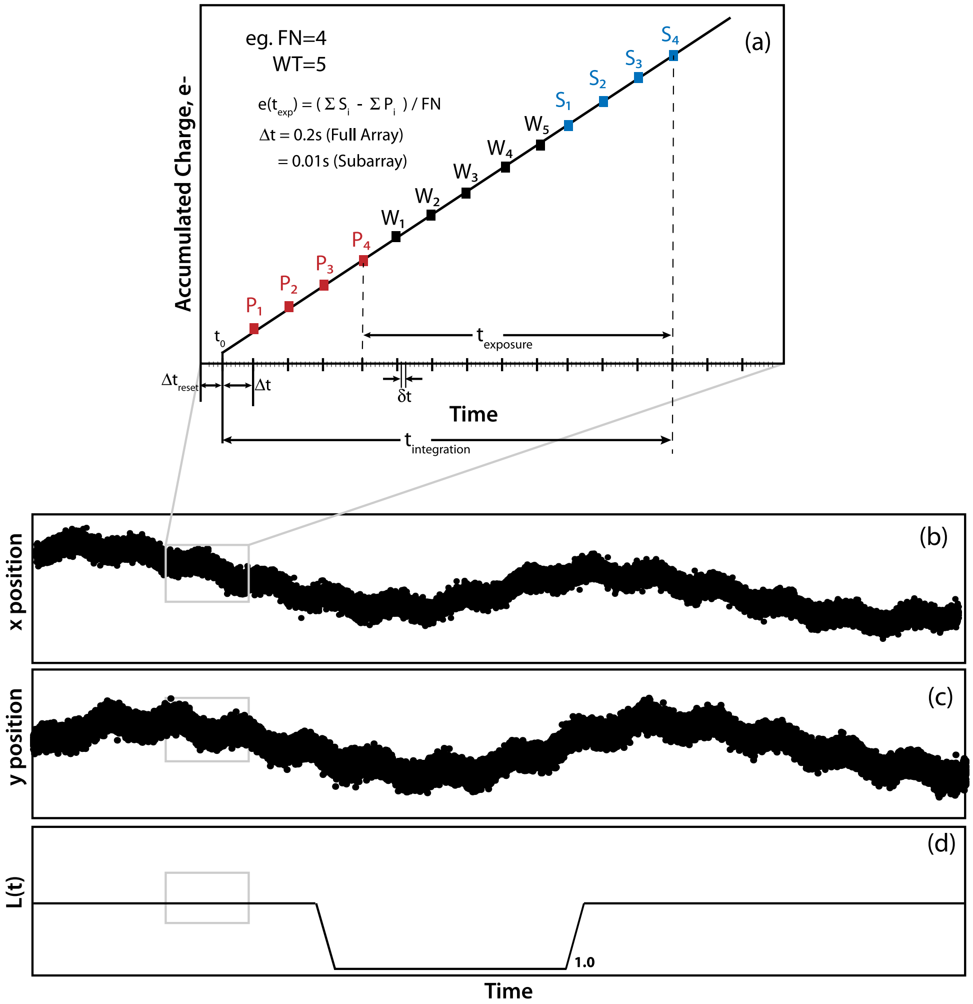

Figure 19. Schematic diagram showing the pointing and sampling aspects of the IRAC simulator. Panel (a) shows the charge on a pixel as a function of time during an IRAC measurement. We indicate the location of the Fowler sampling "pedestal"P_i and "sample" S_i measurements, for $FN = 4$ and $WT = 5$. (This sampling yields a non-standard frame time of 2.6 s, shown here only for illustrative purposes.) The IRAC sample time is shown as Δt, and the time resolution of the simulation δt is indicated. Panels (b) and (c) indicate the time evolution of the x and y pixel position of a point source on an IRAC array. Panel (d) displays the raw light curve for an unresolved eclipsing planetary system $L(t)$. Gray boxes on panels (b)–(d) showwhere the pixel sampling in panel (a) takes place.

A.5. Exoplanet Wrapper

An additional wrapper subroutine has been written to accomodate simulation of exoplanet measurements with IRAC-SIM. The wrapper features realtime access to the Exoplanets.org database of planetary system parameters (Han et al. 2014). Its main job is to create model exoplanet phase curves as input light curves $L(t)$ to the IRAC simulator. It uses the thermal phase variations model of Cowan & Agol (2011), and the transit (and eclipse) shape model of Mandel & Agol (2002), allowing for the effects of nonlinear limb darkening in the transit. The specifics of *Spitzer* recommended exoplanet observational practice are built-in: long AORs are broken into 12 hr pieces, with a 30 min settling AOR at the beginning, and the enhanced accuracy of target centering with PCRS Peak-Up is simulated.

APPENDIX B
DESCRIPTION OF CORRELATED NOISE-REMOVAL TECHNIQUES

We review below the seven techniques for removing correlated noise used to reduce the XO-3b data sets described in this paper, adding specific notes on implementation.

B.1. BLISS Mapping

BLISS mapping (Stevenson et al. 2012) uses bilinear interpolation over a photometric data set, to predict the intra-pixel response at a given (x, y) location. The procedure establishes a subpixel rectangular grid of node points, referred to as "knots," spanning the data set. Each knot is assigned the mean flux value from among all points in the data set for which that knot point is the nearest. The intra-pixel gain at a given data point is then computed from the knot fluxes via bilinear interpolation to the point (x, y).

For the implementation described here, performed by H. Diamond-Lowe and K. Stevenson, photometric measurements were obtained using the POET pipeline described in Stevenson et al. (2012), which produced artifact-corrected BCD images interpolated to a $1/5$ pixel grid. Centroid positions were measured by fitting a 2D Gaussian profile with fixed width (see the supplemental information for Stevenson et al. 2010) on the resampled images, and fluxes were measured using aperture photometry. Intra-pixel effects were removed using BLISS mapping, and various models were attempted to fit the decorrelated light curve, including a flat or possible linear detector "ramp" (time-dependent flux baseline) and flat, linear, quadratic, or sinusoidal phase variations. The eclipse depth, duration, and time of ingress and egress were fit separately for each epoch, as well as commonly among all visits. Acceptance of model parameters was decided by minimizing the BIC:

$$\mathrm{BIC} = -2\ln\hat{L} + k\ln(n), \tag{43}$$

where $\hat{L} \equiv p(x|\hat{\theta}, M)$ is the maximized value of the likelihood function of the data x given the maximizing parameters $\hat{\theta}$ and the model M, k is the number of free parameters, and n is the number of data points in x. A Differential Evolution Markov Chain (DE–MC) routine (ter Braak & Vrugt 2008) was used to explore the phase space of parameters and estimate their uncertainties (for details, see Stevenson et al. 2012).

B.2. Gaussian Process Regression (GP)

Gaussian process regression is a procedure for using the correlation properties of a data set to predict the value at an arbitrary point. It is alternately known as Kriging and Wiener-Kolmogorov prediction, and was first described in the astrophysical literature by Rybicki & Press (1992) as a means of interpolating irregularly spaced data. The technique was used to model instrumental systematics in exoplanet observations by Gibson et al. (2012) and first applied to *Spitzer*/IRAC data by Evans et al. (2015).

For the Data Challenge, the GP analysis by T. Evans started with a maximum likelihood fit to the eclipse depth, mid-eclipse time, and the variance in the white noise, plus a set of parameters for kernel functions describing how the *covariance* between two photometric measurements varies with their distance in pixel (x, y) and time. The covariance kernel functions are used analogously to the kernel regression function described below, except the standard kernel regression is applied directly to the photometry. Uncertainties for the eclipse parameters were obtained using Markov Chain Monte Carlo (MCMC) with Metropolis–Hastings sampling in the region of maximum likelihood. In the final MCMC step, the covariance kernel parameters and white noise variance were held fixed to allow rapid evaluation of the likelihood. One drawback to the GP method is that the evaluation of the $N \times N$ empirical covariance matrix among N data points is often prohibitive with large data sets. To avoid this difficulty, fluxes and centroids were binned as a function of time in groups of ~30 s, resulting in $N \sim 1000$ data points in each eclipse light curve (see Evans et al. 2015, for more details).

B.3. Independent Component Analysis

ICA is a non-parametric technique for separating blended signals, with little specific a priori knowledge of their structure. This is the classic "cocktail party problem" of signal processing, which attempts to mimic the human brain's innate capacity for hearing multiple speakers in a crowded room (Hyvärinen & Oja 2000). In contrast to *principle* component analysis, ICA does not assume that the statistically independent signals follow Gaussian distributions, and in fact attempts to maximize the *non*-Gaussianity after separation. The methods of ICA were first developed for exoplanet light curve analysis by Waldmann (2012) and used on *Spitzer* data by Morello et al. (2014).

The ICA data reduction of the XO-3b real and simulated eclipse data sets, by G. Morello, used a new "wavelet-pixel" variant on the approach introduced in Morello et al. (2014) and Morello (2015) for transits. In this variant, the source separation operates on wavelet-transformed individual pixel light curves, after which the resulting components are transformed back to the time domain. The wavelet transform was useful for enhancing the signal-to noise ratio in the lower frequency instrument systematics components prior to ICA. By operating on the individual pixel light curves, ICA circumvents a built-in degeneracy that occurs for most decorrelation techniques, which decorrelate aperture fluxes using (x, y) centroids.[19]

The sum of an eclipse light curve model (including a linear phase variation) and scaled versions of the non-eclipse independent components was then fit to the raw light curve. An Adaptive Metropolis MCMC algorithm with delayed rejection produced chains of 300,000 values to serve as samples of the posterior distributions of the fit parameters. These distributions yielded estimates of the parameters and their uncertainties. The final error bars were then increased to include the ICA component separation error. A full description of the implementation of ICA on the XO-3b real data set is given by Morello et al. (2016).

B.4. Kernel Regression (KR/Data, KR/PMap)

Kernel regression is the first non-parametric technique used to measure and correct the intra-pixel sensitivity of the *Spitzer*/IRAC InSb detectors. In mathematics and engineering, the general use of kernel-based methods was originally applied to the estimation of density functions (e.g., histograms). Eventually they were proposed as potential tools for regression (i.e., the fitting or prediction of function values; Nadaraya 1964; Watson 1964). The kernel regression estimator is a weighted average of the measured data, with a kernel function specifying how the weight decreases with distance from the target point $\boldsymbol{x} = (\mathrm{x}, \mathrm{y},\ldots)$ to be estimated. Limiting the contributing data points to the k nearest neighbors to the target is an additional expedient for faster computation (Stone 1977). The first application of kernel regression to estimate the intra-pixel response in *Spitzer* photometry was done by Ballard et al. (2010). The use of the Noise Pixel parameter, $\tilde{\beta}$, as a third component of the distance metric of the weighting kernel (in addition to x and y pixel centroid) was first described by Lewis et al. (2013).

The most commonly used version of kernel regression, KR/Data, uses the data to be corrected as its own "training set," i.e., the data (except the single datum being corrected) are used in the kernel average to obtain the correction. This requires that the observations contain sufficient redundancy in positioning to allow estimation of its own correlated noise via the inverse distance-weighted average, even in the presence of temporal variations in the astrophysical source. The published reduction of the real XO-3b eclipse data set, described in Wong et al. (2014), used KR/Data. A complementary analysis of the synthetic XO-3b data set was performed by I. Wong for the Data Challenge. For both analyses only the x and y centroids (as measured using the center of light technique; see Section 3.1) were used in the kernel's distance metric, but $\tilde{\beta}$ was employed for most eclipses as a scale factor in determining the optimal aperture size for the photometry. Wong et al. chose $k = 50$ nearest neighbors for the weighted

[19] Since flux and centroid are both weighted sums of pixel intensities (center-of-light centroids are linear sums, whereas Gaussian fits are effectively nonlinear sums), flux and centroid are always correlated *by definition*. This intrinsic correlation effectively adds "noise" to the flux versus centroid signal.

sums. They fit the data in two ways: each epoch separately, and all epochs combined. The separate fits were only concerned with the eclipse depth, time of mid-eclipse, and linear slope of phase curve, whereas the global fits also included the planet-to-star radius ratio, the orbital inclination, and the semimajor axis to stellar radius. Both fits were performed using a Levenberg–Marquardt (L–M) nonlinear least squares algorithm. They then used both a prayer-bead method and an MCMC routine to estimate the distributions of each parameter and their uncertainties, and reported the largest uncertainty of the two methods.

A variation on the kernel regression technique, KR/Pmap, uses the photometry of a separate calibration star as the training set for the regression (Krick et al., 2016). For each science data point to be corrected, the k nearest neighbors in the pixel mapping (pmap) data set are found, based on the Euclidean distance in x and y centroid and $\tilde{\beta}$. Similarly to the KR/Data implementation, $k = 50$ was chosen. The kernel-weighted pmap data are then summed and normalized by the calibration star flux averaged over the pixel. The potential benefit of KR/Pmap over KR/Data is that the correction is not built from the science measurements themselves, and therefore time-varying astrophysical signal does not contribute to the kernel averages. On the other hand, detector variability (e.g., latent charge) may differ between the calibration star measurements and those of the data to be corrected, and bias the regression.

The KR/Pmap analysis of the Data Challenge measurements was performed by J. Krick and J. Ingalls. Different calibration data sets were used to correct the real and synthetic XO-3b measurements. For the real data, the SSC has accumulated approximately 400,000 pixel mapping measurements of BD+67 1044, a star that is not known to vary, which are positioned near the "sweet spot" (peak of response) of the Ch 2 (4.5 μm) subarray pixel (15, 15). Since the IRAC simulator uses an idealized PRF that cannot replicate the detailed structure of the actual pixel response, the real BD +67 1044 data set is not appropriate for reduction of simulated data. Instead, a synthetic pixel mapping set was created. The measurements were designed to mimic the actual pmap measurements of BD +67 1044, with similar source flux, integration parameters, mapping centers, and number of data points. The pointing model parameters were taken from the same ranges as for the XO-3b simulations (Table 5), to approximate realistic motions during integration and sampling.

Eclipse parameters were derived from the KR/Pmap-decorrelated data by fitting a Mandel & Agol (2002) light curve shape with no phase trend using an L–M nonlinear least squares algorithm. The uncertainties returned were solely the formal uncertainties in the L–M fit, and should be considered underestimates. As a check on the results, Transit Analysis Package (TAP; Gazak et al. 2012) was also used to fit the eclipses, after setting the limb darkening coefficients to zero. While the uncertainties tended to be more realistic under TAP's MCMC analysis, the eclipse depths themselves were systematically low compared to both the L–M fit and the mean of the other techniques. We therefore decided to use the L–M results and assess the uncertainties using the "overdispersion factor" described in Section 3.2.

B.5. Pixel Level Decorrelation

The PLD technique (Deming et al. 2015) is a parametric method that expresses the correlated noise in terms of a Taylor expansion sum of individual pixel values, instead of a function of centroid position. The Taylor expansion partial derivatives become linear coefficients multiplying the (normalized) pixel values, a function that can be fit and removed. As with ICA, using the individual pixel values avoids the flux/centroid degeneracy inherent in most decorrelation methods.

The PLD reduction of the Data Challenge observations, by D. Deming, used 2D Gaussian centroiding (Agol et al. 2010) only to determine where to place the circular aperture, but not as a decorrelation variable. An eclipse function was fit to the photometry simultaneously with the pixel coefficients and a quadratic phase curve (see Deming et al. 2015, Equation (4)). Due to time limitations, a full MCMC analysis of uncertainties was not possible, and so error bars were estimated using the slope of the SD versus bin size relationship for the residuals (as described in Deming et al. 2015) to extrapolate to bins the width of the eclipse duration.

B.6. Segmented Polynomial, K2 Pipeline [SP(K2)]

The segmented polynomial algorithm was originally developed for use with K2 data (Buzasi et al. 2015), where detrending is normally required due to the presence of spacecraft pointing resets, and other less significant sources of correlated noise. The approach is reminiscent of polynomial surface fitting as used on *Spitzer* data, but with some differences. Detrending is carried out in two stages. In the first stage, a third-order polynomial is fit to the flux versus (x, y) centroid for the entire time series and removed. This process is repeated, with successive third-order polynomial fits being applied to each set of residuals, until there is $<1\%$ further reduction in the high-frequency noise SD. In the second stage, the resulting time series is divided into segments, each of which is iteratively decorrelated using polynomial fitting. This segmented detrending is repeated for 10 different segment lengths between 0.04 and 0.125 of the total time series length, and the final time series is the result of applying a median filter to the 10 results.

SP(K2) detrending was applied to the *Spitzer* Data Challenge measurements by D. Buzasi. Due to time limitations a simple box function was used to fit the eclipse profile, and the uncertainties reported are formal fit errors.

No attempt was made to tune SP(K2) for *Spitzer* data. Future analysis might benefit from adjustment of the segmentation and fitting strategy. For K2, the segmentation is partially necessary to accommodate unpredictable discontinuous jumps in source position when the pointing is reset, but this is not as much of an issue for *Spitzer* staring mode observations (except for observations longer than 12 hr, for which there will be predictable pointing resets). Furthermore, IRAC's intra-pixel gain variations are much larger than those of K2—the gain of IRAC varies by ~8% in Ch 1 (3.6 μm) and ~4% in Ch 2 (4.5 μm) across a pixel, whereas the K2 effect is only about 2%. *Spitzer* data might be more amenable to spatial, rather than temporal, segmentation of the data in stage 2.

REFERENCES

Agol, E., Cowan, N. B., Knutson, H. A., et al. 2010, ApJ, 721, 1861
Altman, D. G., & Bland, J. M. 1983, The Statistician, 32, 307
Ballard, S., Charbonneau, D., Deming, D., et al. 2010, PASP, 122, 1341
Bartlett, J. W., & Frost, C. 2008, Ultrasound Obstet Gynecol, 31, 466
Beaulieu, J. P., Tinetti, G., Kipping, D. M., et al. 2011, ApJ, 731, 16
Beichman, C., Benneke, B., Knutson, H., et al. 2014, PASP, 126, 1134
Burrows, A. S. 2014, PNAS, 111, 12601
Buzasi, D. L., Carboneau, L., Hessler, C., Lezcano, A., & Preston, H. 2015, arXiv:1511.09069v1

Charbonneau, D., Knutson, H. A., Barman, T., et al. 2008, ApJ, 686, 1341
Clampin, M. 2008, AdSpR, 41, 1983
Cowan, N. B., & Agol, E. 2011, ApJ, 726, 82
Crossfield, I. J. M. 2015, PASP, 127, 941
Deming, D., Knutson, H., Kammer, J., et al. 2015, ApJ, 805, 132
Demory, B.-O., Gillon, M., Madhusudhan, N., & Queloz, D. 2016, MNRAS, 455, 2018
Deroo, P., Swain, M. R., & Green, R. O. 2012, Proc. SPIE, 8442, 844241
Diamond-Lowe, H., Stevenson, K. B., Bean, J. L., Line, M. R., & Fortney, J. J. 2014, ApJ, 796, 66
Dziak, J. J., Coffman, D. L., Lanza, S., & Li, R. 2012, Sensitivity and Specificity of Information Criteria, Tech. Rep. 12-119 (State College, PA: Pennsylvania State Univ.), https://methodology.psu.edu/media/techreports/12-119.pdf
Evans, T. M., Aigrain, S., Gibson, N., et al. 2015, MNRAS, 451, 680
Fazio, G. G., Hora, J. L., Allen, L. E., et al. 2004, ApJS, 154, 10
Fraine, J. D., Deming, D., Gillon, M., et al. 2013, ApJ, 765, 127
Garnett, J. D., & Forrest, W. J. 1993, Proc. SPIE, 1946, 395
Gazak, J. Z., Johnson, J. A., Tonry, J., et al. 2012, AdAst, 2012, 30
Gibson, N. P., Aigrain, S., Roberts, S., et al. 2012, MNRAS, 419, 2683
Grillmair, C. J., Carey, S. J., Stauffer, J. R., et al. 2012, Proc. SPIE Operations: Strategies, Processes, and Systems IV, 8448, 84481I
Han, E., Wang, S. X., Wright, J. T., et al. 2014, PASP, 126, 827
Hansen, C. J., Schwartz, J. C., & Cowan, N. B. 2014, MNRAS, 444, 3632
Hora, J. L., Witzel, G., Ashby, M. L. N., et al. 2014, ApJ, 793, 120
Hyvärinen, A., & Oja, E. 2000, NN, 13, 411
Ingalls, J. G., Carey, S. J., Lowrance, P. J., Grillmair, C. J., & Stauffer, J. R. 2014, Proc. SPIE, 9143, 91431M
Ingalls, J. G., Krick, J. E., Carey, S. J., et al. 2012, Proc. SPIE, 8442, 84421Y
Killeen, P. R. 2005, Psychol. Sci., 16, 345
Knutson, H. A., Charbonneau, D., Allen, L. E., Burrows, A., & Megeath, S. T. 2008, ApJ, 673, 526
Knutson, H. A., Lewis, N., Fortney, J. J., et al. 2012, ApJ, 754, 22
Knutson, H. A., Madhusudhan, N., Cowan, N. B., et al. 2011, ApJ, 735, 27
Krick, J. E., Ingalls, J., Carey, S., et al. 2016, ApJ, in press (arXiv:1603.03383)
Kruschke, J. K. 2013, J. Exp. Psych.: General, 142, 573
Lewis, N. K., Knutson, H. A., Showman, A. P., et al. 2013, ApJ, 766, 95
Lust, N. B., Britt, D., Harrington, J., et al. 2014, PASP, 126, 1092
Lyons, L. 1992, JPhA, 25, 1967
Mandel, K., & Agol, E. 2002, ApJL, 580, L171
Morello, G. 2015, ApJ, 808, 56
Morello, G., Waldmann, I. P., & Tinetti, G. 2016, arXiv:1601.03959v1
Morello, G., Waldmann, I. P., Tinetti, G., et al. 2014, ApJ, 786, 22
Morello, G., Waldmann, I. P., Tinetti, G., et al. 2015, ApJ, 802, 117
Nadaraya, E. A. 1964, Theory Prob. & Appl., 9, 141
Reach, W. T., Megeath, S. T., Cohen, M., et al. 2005, PASP, 117, 978
Ricker, G. R., Winn, J. N., Vanderspek, R., et al. 2015, JATIS, 1, 014003
Rybicki, G. B., & Press, W. H. 1992, ApJ, 398, 169
Schwartz, J. C., & Cowan, N. B. 2015, MNRAS, 449, 4192
Sing, D. K., Fortney, J. J., Nikolov, N., et al. 2016, Natur, 529, 59
Stevenson, K. B., Harrington, J., Fortney, J. J., et al. 2012, ApJ, 754, 136
Stevenson, K. B., Harrington, J., Nymeyer, S., et al. 2010, Natur, 464, 1161
Stone, C. J. 1977, AnSta, 5, 595
Stutzki, J., Bensch, F., Heithausen, A., Ossenkopf, V., & Zielinsky, M. 1998, A&A, 336, 697
ter Braak, C. J. F., & Vrugt, J. A. 2008, Stat. Comput., 18, 435
Tinetti, G. 2015, in DPS, Univ. College London, #416.20
Waldmann, I. P. 2012, ApJ, 747, 12
Watson, G. S. 1964, Sankhyā Ser. A, 26, 359
Werner, M. W., Roellig, T. L., Low, F. J., et al. 2004, ApJS, 154, 1
Wong, I., Knutson, H. A., Cowan, N. B., et al. 2014, ApJ, 794, 134
Wong, I., Knutson, H. A., Kataria, T., et al. 2016, ApJ, 823, 122
Zellem, R. T., Lewis, N. K., Knutson, H. A., et al. 2014, ApJ, 790, 53