



# ***Excalibur*: A Nonparametric, Hierarchical Wavelength Calibration Method for a Precision Spectrograph**

Lily L. Zhao<sup>1,2</sup> , David W. Hogg<sup>2,3,4,5</sup> , Megan Bedell<sup>2</sup> , and Debra A. Fischer<sup>1</sup>

<sup>1</sup>Yale University, 52 Hillhouse, New Haven, CT 06511, USA; [lily.zhao@yale.edu](mailto:lily.zhao@yale.edu)

<sup>2</sup>Flatiron Institute, Simons Foundation, 162 Fifth Avenue, New York, NY 10010, USA

<sup>3</sup>Center for Cosmology and Particle Physics, Department of Physics, New York University, 726 Broadway, New York, NY 10003, USA

<sup>4</sup>Center for Data Science, New York University, 60 Fifth Avenue, New York, NY 10011, USA

<sup>5</sup>Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany

Received 2020 August 26; revised 2020 November 14; accepted 2020 November 22; published 2021 January 20

## **Abstract**

*Excalibur* is a nonparametric, hierarchical framework for precision wavelength calibration of spectrographs. It is designed with the needs of extreme-precision radial-velocity (EPRV) instruments in mind, which require calibration or stabilization to better than  $10^{-4}$  pixels. Instruments vary along only a few dominant degrees of freedom, especially EPRV instruments that feature highly stabilized optical systems and detectors. *Excalibur* takes advantage of this property by using all calibration data to construct a low-dimensional representation of all accessible calibration states for an instrument. *Excalibur* also takes advantage of laser-frequency combs or etalons, which generate a dense set of stable calibration points. This density permits the use of a nonparametric wavelength solution that can adapt to any instrument or detector oddities better than parametric models, such as a polynomial. We demonstrate the success of this method with data from the Extreme Precision Spectrograph (EXPRES), which uses a laser-frequency comb. When wavelengths are assigned to laser comb lines using *excalibur*, the rms of the residuals is about one-fifth that of wavelengths assigned using polynomial fits to individual exposures. Radial-velocity measurements of HD 34411 show a reduction in rms scatter over a 10 month time baseline from  $1.17$  to  $1.05 \text{ m s}^{-1}$ .

*Unified Astronomy Thesaurus concepts:* [Exoplanet detection methods \(489\)](#); [Radial velocity \(1332\)](#); [Astronomical techniques \(1684\)](#); [Astronomical instrumentation \(799\)](#); [Spectrometers \(1554\)](#); [Computational methods \(1965\)](#)

*Supporting material:* machine-readable table

## **1. Introduction**

Precise radial-velocity (RV) programs have been fruitful in finding and characterizing extrasolar planets (e.g., Mayor et al. 2011; Bonfils et al. 2013; Plavchan et al. 2015; Butler et al. 2017). These programs typically make use of spectrographs with resolutions on the order of 50,000–100,000, which correspond to line widths on the order of  $3000 \text{ m s}^{-1}$ . The state-of-the-art RV precision had reached  $1 \text{ m s}^{-1}$  by 2016 (Fischer et al. 2016). The newest generation of instruments aims to reach  $0.1 \text{ m s}^{-1}$  precision, the required precision for detecting terrestrial worlds. This requires new spectrographs to be calibrated or stabilized to better than  $10^{-4}$  of a pixel (assuming that the spectrographs are well sampled). Two next-generation spectrographs, the Extreme Precision Spectrograph (EXPRES) and the Echelle Spectrograph for Rocky Exoplanets and Stable Spectroscopic Observations (ESPRESSO), have been commissioned for more than a year and are demonstrating  $<0.1 \text{ m s}^{-1}$  instrumental errors and  $\sim 0.2 \text{ m s}^{-1}$  errors on stars (Pepe et al. 2013; Jurgenson et al. 2016; Blackman et al. 2020; Brewer et al. 2020; Petersburg et al. 2020; Suárez Mascareño et al. 2020).

Traditionally, wavelength solutions are constructed by fitting a polynomial to lines from a calibration source in order to describe the relationship between wavelength and pixel for each echelle order (Butler et al. 1996; Lovis & Pepe 2007; Cersullo et al. 2019). In this framework, each calibration image is treated independently. The returned wavelength solutions work well at the level of  $1 \text{ m s}^{-1}$  precision.

The move toward  $0.1 \text{ m s}^{-1}$  RV precision necessitates higher-fidelity calibration data and wavelength models. These

models need to account for high-order spatial variations that can arise from small imperfections in the optics of an instrument and nonuniformity in detector pixel sizes/spacing. There has been a significant effort in using an entire set of calibration images to identify incongruous thorium–argon (ThAr) lines (Coffinet et al. 2019) or obtain high-resolution Fourier transform spectra of reference cells (Wang et al. 2020). It has also been found that using multiple polynomials in the dispersion direction, tuned to capture detector defects, better describes the wavelength solution than a single, continuous polynomial (Milaković et al. 2020).

Here, we propose to simplify and improve calibration programs for extreme-precision RV (EPRV) hardware systems with two practical yet innovative ideas. The first flows from the fact that calibration sources, which include arc lamps (in some wavelength ranges), etalons, and laser-frequency combs (LFCs), illuminate the spectrograph with very stable, very dense sets of lines—almost every location in the spectrograph image plane is surrounded by nearby, useful calibration lines. This enables use of a calibration methodology that is nonparametric, or not defined by a prescribed analytic function described by a finite number of parameters: if every point in the spectrograph detector is sufficiently surrounded by nearby calibration lines, the wavelength solution can, for example, be made simply as an interpolation of the calibration data. The density of lines removes the need to enforce any functional form for the wavelength solution (such as a continuous ninth-order polynomial). In some ways, this is a generalization of recent work that demonstrates the efficacy of constructing a wavelength solution as multiple, segmented polynomials (Milaković et al. 2020). A nonparametric approach will improve

calibration accuracy by not forcing the choice of a parametric form that may bias the calibration, especially when the chosen function is inappropriate (as, for example, polynomials are at detector edges).

The second simple idea follows from the observation that most physical systems have only a few dominant degrees of freedom, meaning most spectrographs vary along only a small number of axes in “calibration space,” or the (very high-dimensional) space of all possible wavelength solutions. This is particularly true of EPRV instruments, which are equipped with stringent environmental stabilizing. The thermomechanical stability of these instruments reduces the variations they experience to something that can be represented by a low-dimensional framework. That is, spectrographs, especially stabilized ones, should have few environmentally accessible degrees of freedom. This renders it inadvisable to fit each calibration exposure or calibrate each science exposure independently. Instead, all the calibration data (or all the data) should be used to determine the calibration space in which the instrument can and does vary. Subsequent calibration work then need only determine where in the small, accessible part of calibration space the spectrograph is located for each exposure.

In the context of probabilistic models, this structure is hierarchical: the calibration data are used not just to determine the wavelength solution at one moment, but also to determine the possible calibration space of wavelength solutions at all moments. In statistics, this concept is often described as denoising: we can improve calibration by recognizing that every calibration exposure contains information about every other calibration exposure. Thus, every exposure can be improved (i.e., denoised) with information from every other exposure.

The method we propose here—*excalibur*—embodies these ideas. It is a nonparametric, hierarchical, data-driven method to generate a wavelength model. By being nonparametric, it delivers enormous freedom to the wavelength solution to match or adapt to any instrument or detector oddities. By being hierarchical, it restricts that freedom tremendously, but it does so appropriately for the empirically determined variations in a spectrograph.

The method *excalibur* is designed for temperature-controlled, fiber-fed spectrographs with good calibration sources, such as LFCs or etalons. We have in mind EPRV instruments and EPRV science cases, primarily because the need for good wavelength calibration is so great in this field. Nevertheless, we expect *excalibur* to have applications for other kinds of spectrographs in other contexts. *Excalibur* should be applicable to spectrographs with low-dimensional variability, though the precision of the returned wavelengths will depend on the available calibration sources (more discussion in Section 6 below).

## 2. Method

The *excalibur* method is designed to take many calibration images, each containing a series of calibration lines with known wavelengths and well-fit detector positions, and denoise and interpolate this information into a full wavelength model applicable to all exposures taken with the instrument. It operates on two core ideas: the wavelength solution should be allowed flexibility, but it lives in a very low-dimensional calibration space where the degrees of freedom are set by the

limited kinematics of the spectrograph hardware. *Excalibur* therefore assumes that the space of possible calibration states for an instrument is low-dimensional, but assumes very little about the forms of those states.

*Excalibur* also assumes dense enough calibration line coverage with well-fit line centers to provide sufficient constraints on an interpolated wavelength solution across an echelle order. Upstream errors in line center positions may propagate through *excalibur* wavelength models. The required line density is dependent on the required precision of the returned wavelength model; larger spacing between lines offers less constraint and is likely to return worse wavelengths. We revisit and quantify these conditions in Section 6.

Wavelength calibration is usually posed in the following way: given an exposure  $n$  and echelle order  $m$ , there is a relationship between the 2D  $(x, y)$ -position on the detector and the wavelength  $\lambda$ :

$$\lambda(x, y, m, n) = f(x, y, m; \theta_n), \quad (1)$$

where  $\theta_n$  represents the parameters describing the wavelength solution for a given exposure.

Classically, pipelines employ polynomials to construct smooth wavelength solutions for each exposure. For example, the EXPRES pipeline sets the function  $f(x, y, m; \theta_n)$  from Equation (1) to a 2D, ninth-order polynomial, where  $\theta_n$  represents the polynomial coefficients  $c_{nij}$  unique to each exposure  $n$  (Petersburg et al. 2020):

$$\lambda(x, m, n) = \sum_{i=0}^9 \sum_{j=0}^9 c_{n,i,j} x^i m^j + \text{noise}. \quad (2)$$

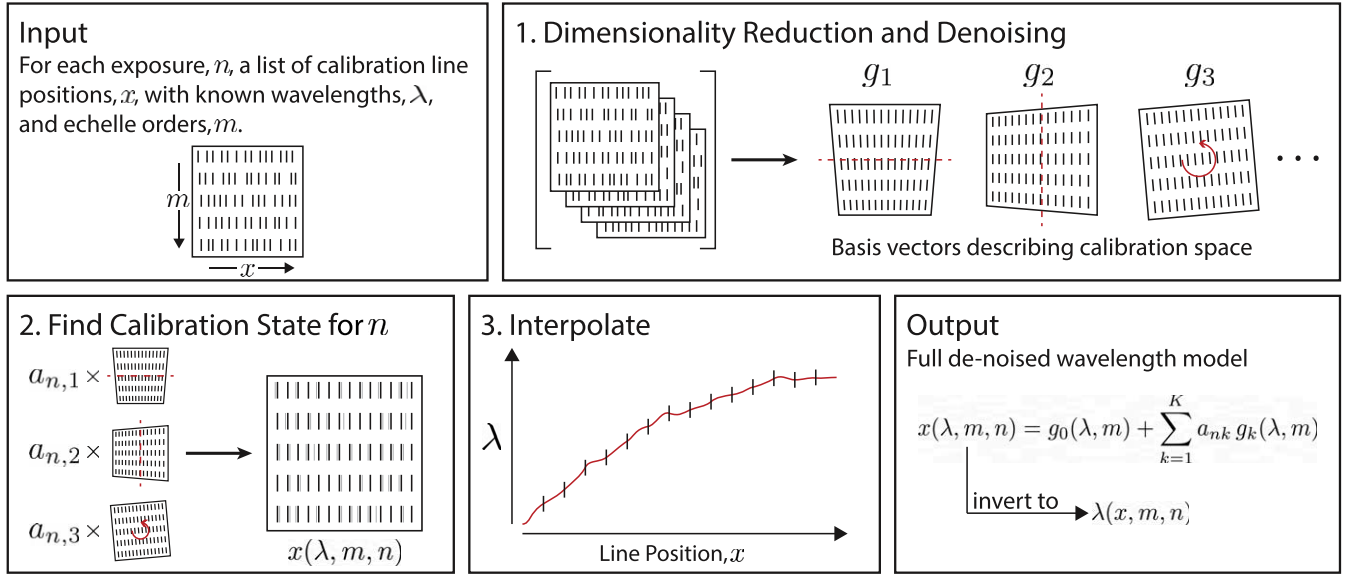
Here, the  $y$  dependence is dropped as, in our framing, this dependence is carried by the spectral order  $m$ . The line position on the detector is therefore uniquely identified by the echelle order  $m$  and pixel in the dispersion direction  $x$ . The coefficients  $c_{nij}$  are interpolated from the time of the calibration exposures to the time  $t_n$  of a science exposure  $n$  by a third-order polynomial with respect to time. This third-order polynomial is evaluated at the time of noncalibration, science exposures to reconstruct the coefficients for a 2D, ninth-order polynomial wavelength solution for that exposure. Each calibration image obtains its  $c_{nij}$  independently.

Given a stabilized instrument with low degrees of freedom, however, the calibration of any image can be reliably informed by the calibration of every other image. The calibration data themselves can be used to develop a low-dimensional basis for expressing the space of all possible calibrations for a spectrograph with few degrees of freedom.

If the space of all calibration possibilities is in fact  $K$ -dimensional (where  $K$  is a small integer, e.g., 2 or 8 or thereabouts) and if the calibration variations are so small that they can be linearized, then the function  $f(x, m; \theta_n)$  from Equation (1) should be low-dimensional. In *excalibur*, we transpose the calibration model—making the position  $x$  a function of  $\lambda$ —into the following form:

$$x(\lambda, m, n) = g_0(\lambda, m) + \sum_{k=1}^{K\Sigma} a_{n,k} g_k(\lambda, m), \quad (3)$$

where  $g_0(\lambda, m)$  is the fiducial, mean, or standard calibration of the spectrograph,  $a_{n,k}$  denotes the  $K$  scalar amplitudes for each exposure  $n$ , and  $g_k(\lambda, m)$  denotes the basis functions expressing the “directions” in calibration space in which the



**Figure 1.** A cartoon representation of the *excalibur* method, as described in Section 2. We exaggerate variations in measured line positions, changes in calibration space, and interpolation deviations for clarity. In step one, dimensionality reduction and denoising (Section 2.1), the complete set of line positions for all exposures is analyzed to return a set of  $K$  basis vectors  $g_n$ , which represent different ways the spectrograph calibration changes. These basis vectors span the  $K$ -dimensional calibration space of the spectrograph, which includes all possible wavelength solutions. In step two (Section 2.2), the amplitude of each basis vector,  $a_{n,k}$ , is interpolated to return the calibration state for a specific science exposure, returned as a set of denoised calibration lines. The assigned wavelengths of these denoised line positions are then interpolated onto other pixels in step three (Section 2.3) to construct a full wavelength model that returns the wavelength as a function of detector position  $x$  and echelle order  $m$ .

spectrograph can depart from the fiducial calibration. The resultant  $x(\lambda, m, n)$ , a list of calibration line positions for a given exposure, can be regarded as the calibration state of the spectrograph for that exposure. When this calibration structure is used to deliver a wavelength solution,  $x(\lambda, m, n)$  can be inverted into  $\lambda(x, m, n)$  to recover the wavelengths for each detector position  $x$  and echelle order  $m$  (Bauer et al. 2015).

The challenge is to learn these basis functions  $g_k(\lambda, m)$  from the data and get the  $K$  amplitudes  $a_{n,k}$  for every exposure  $n$ . There are many ways to discern the basis functions. In this paper, we present one implementation of *excalibur* using principal-component analysis (PCA; Pearson 1901; Jolliffe & Cadima 2016). PCA is justifiable in the limit where exposures have very high signal-to-noise ratios (S/N), as is usually the case with typical calibration images. There are many alternatives to PCA for this dimensionality reduction; we return to this point in Section 5 below.

### 2.1. Dimensionality Reduction: Denoising of Calibration Frames

*Excalibur* will use calibration images to determine (1) the space in which an instrument varies and (2) where in the accessible calibration space the spectrograph is located for each exposure. For each calibration exposure  $n$ , *excalibur* requires a full list of lines  $(\lambda, m)$  that are expected to appear in each calibration exposure. Each line is uniquely defined by a combination of echelle order  $m$  and “true” or theoretical wavelength  $\lambda$ . There are many strategies for identifying calibration line positions and matching them to their assigned wavelengths; this problem is left out of the scope of this work.

*Excalibur* assumes that line positions have been identified “correctly,” as the position of a calibration line is determined in the same way as the position of a stellar line when extracting

RVs. This also inherently assumes that the calibration lines are not subject to any effect that the science exposures are not—for example, differences in charge transfer inefficiency. We caution that systematic errors or large uncertainties in fitting line positions easily propagate to biases in the wavelength models returned by *excalibur*. For more discussion, see Section 7.

For each exposure  $n$ , every line  $(\lambda, m)$  has an associated fitted (measured) detector position  $x(\lambda, m, n)$ —for example, an  $x$ -pixel in a 2D extracted echelle order. Fitted line centers that are missing from an exposure (e.g., because the fit failed due to noise or the line is not in its usual echelle order) can be assigned a NaN (hardware not-a-number) for that exposure instead. Let there be  $P$  lines per exposure. *Excalibur* reads in an  $N \times P$  matrix of line positions for each of the  $P$  lines for each of the  $N$  exposures.

The mean of measured line positions over the set of calibration exposures represents the fiducial or standard calibration of the spectrograph,  $g_0(\lambda, m)$ . In this implementation of *excalibur*, PCA is performed on the difference between this fiducial calibration and each individual line position. The returned principal components serve as basis functions  $g_k(\lambda, m)$  expressing the possible deviations of the spectrograph from this fiducial calibration. The magnitude of each principal component for each exposure,  $a_{n,k}$ , represents the scalar amplitude of these deviations for each exposure. *Excalibur* then uses a small number  $K$  of principal components to reconstruct a denoised version of the line positions as formulated in Equation (3).

Missing line center measurements, which were previously marked by NaN, are replaced with denoised estimates. This is done iteratively until the estimates of missing line centers change by less than 0.01%. This process can be repeated on line centers deemed outliers by some metric, to account for

lines that may have been misidentified or misfit. The principal components from the final iteration are used to define the spectrograph's calibration space, while the associated amplitudes for each component pinpoint are used to identify where in that calibration space the spectrograph is located for each calibration exposure.

#### Algorithm 1. Dimensionality Reduction and Denoising

---

**Data:** Line positions  $x(\lambda, m, n)$  for each exposure  $n$ , with wavelengths  $\lambda$  and echelle orders  $m$

**Result:** Basis vectors of the low-dimensional calibration space,  $g_k(\lambda, m)$ , and location of exposures in calibration space expressed by amplitudes  $a_{n,k}$

**While** change in missing or outlier line centers  $> 0.01\%$  **do**  
 $g_0(\lambda, m) = x(\lambda, m, n)$ ;  
 using singular-value decomposition, find  $U, \Sigma, V$  s.  
 $t. U\Sigma V^* = (x(\lambda, m, n) - g_0(\lambda, m))$ ;  
 let  $a_{n,k} = U \cdot \Sigma$  and  $g_k(\lambda, m) = V$ ;  
 $x(\lambda, m, n) = g_0(\lambda, m) + \sum_{k=1}^K a_{n,k} g_k(\lambda, m)$  for  
 $x(\lambda, m, n) = \text{NaN}$ , where  $K$  is a small integer. **End**

---

### 2.2. Interpolating Calibration Position

In *excalibur*, the amplitude  $a_{n,k}$  of each principal component is interpolated to determine the calibration state of the spectrograph. For example, the amplitude can be interpolated with respect to time to recreate the calibration state of the spectrograph at different times. The choice of what to interpolate against depends on the dominant contribution to variation in the calibration of the instrument.

In the implementation of *excalibur* presented here, the amplitudes of the principal components are interpolated linearly with respect to time. This is discussed more in Section 5.1. Let a prime denote values related to a science exposure  $n'$  for which we want wavelengths. We use linearly interpolated magnitudes  $a_{n',k}$  at time  $t_{n'}$  to construct the calibration state of the spectrograph for that point in time. Using the interpolated amplitudes  $a_{n',k}$  and the basis vectors  $g_k(\lambda, m)$  returned by the denoising process, a new set of calibration lines  $x'(\lambda, m, n')$  can be constructed for any exposure as formulated in Equation (3).

### 2.3. Interpolating a Wavelength Solution

From the denoising step, *excalibur* can now construct a set of calibration lines  $x'(\lambda, m, n')$  for any exposure  $n'$ . To construct a wavelength solution, we invert  $x'(\lambda, m, n')$  to  $\lambda(x', m, n')$  by interpolating known wavelengths of the calibration lines over the detector position. For instance, interpolating the known wavelengths versus the line centers for an echelle order  $m$  onto every integer  $x$  will generate wavelengths for each pixel in the echelle order.

After experiments, we find that a cubic-spline interpolation that enforces monotonicity, such as a piecewise cubic Hermite interpolating polynomial (PCHIP) interpolator, works well for interpolating wavelengths onto pixels. A cubic spline allows for more flexibility than a parameterized function, while enforced monotonicity allows the wavelength solution  $\lambda(x', m, n')$  to be invertible and prevents the spurious deviations that may befall a cubic spline. Choices of interpolation scheme  $K$  and other tests are further discussed in Section 5.3.

#### Algorithm 2. Generating Wavelength Solution

---

**Data:** The fiducial calibration of the spectrograph,  $g_0(\lambda, m)$ ; the magnitudes of the principal components for each exposure,  $a_{n,k}$ ; the basis vectors spanning the calibration space of the spectrograph,  $g_k(\lambda, m)$

**Result:** Wavelengths for detector positions  $x'(m, n')$  of exposure  $n'$  with time  $t_{n'}$ , where the prime is used to denote values relevant to this new exposure

Find  $a_{n',k}$  by interpolating  $a_{n,k}$  with respect to  $t_{n'}$ ;  
 $x'(\lambda, m, n') = g_0(\lambda, m) + \sum_{k=1}^K a_{n',k} g_k(\lambda, m)$ , where  $K = 6$ .

**For** each unique  $m$  **do**  
 interpolate  $\lambda$  with respect to  $x'(\lambda, m, n')$  onto  
 pixels  $x'(m, n')$ .  
**End**

---

The implementation of *excalibur* described here is hosted on GitHub.<sup>6</sup>

### 3. Data

We test *excalibur* using data from EXPRES. EXPRES is an environmentally stabilized, fiber-fed optical spectrograph with a median resolving power  $R = \lambda/\delta\lambda = \sim 137,000$  over a wavelength range of 390–780 nm (Jurgenson et al. 2016; Blackman et al. 2020). EXPRES has two different wavelength calibration sources, a ThAr lamp and a Menlo Systems LFC. LFCs are unique in that the wavelengths of their emission lines are stable and exactly known at picometer accuracy (Wilken et al. 2012; Molaro et al. 2013; Probst et al. 2014).

Rather than using a simultaneous calibration fiber, two to three LFC exposures are obtained roughly every 30 minutes while the telescope is slewing to new targets. ThAr exposures are taken at the beginning and end of each night. All calibration data are taken through the science fiber, so that calibration light travels down the same optical pathway and is projected onto the same pixels as the science observations. Light passes through a pupil slicer and double scrambler before being injected into a rectangular fiber, which is fed through a mechanical agitator to ensure modal mixing (Petersburg et al. 2018).

The LFC lines cover echelle orders 84–124, which contain approximately 19,200 calibration lines. Though our results are primarily based on work with LFC data, there will be some discussion of applications to arc lamps below. The ThAr lines cover all 86 extracted orders of EXPRES (echelle orders 75–160), which include approximately 5300 lines. For both the LFC and ThAr data, lines that appear in less than 60% of the exposures are not included in the analysis. Similarly, exposures with more than 60% of expected lines missing are cut from the analysis. A list of echelle orders  $m$ , line wavelengths  $\lambda$ , and pixel positions  $x$  are calculated by the EXPRES pipeline (Petersburg et al. 2020) for every line of every exposure and read into *excalibur*.

Line positions from the EXPRES pipeline are generated as follows. A ThAr wavelength solution is generated from each ThAr exposure using the IDL code `thid.pro`, developed by Jeff Valenti. This code identifies ThAr lines by matching lines in an exposure against a line atlas. Line matching is carried out in an automated and unsupervised way with a Levenberg–Marquardt minimization routine. Once each line's position is identified and matched to a wavelength from the line atlas, a

<sup>6</sup> <https://www.github.com/lilylingzhao/excalibur>



sixth-order, 2D polynomial is fit over the pixel location  $x$ , echelle order  $m$ , and scaled wavelength  $m\lambda$  (wavelengths are scaled in order to distinguish lines that may appear in more than one order).

Flat-relative, optimally extracted LFC data is background-corrected using a univariate spline. Each peak in an echelle order is then fit with a Gaussian. The mean of this fitted Gaussian to a single peak is taken to be the center of the line. For each line, the ThAr wavelength solution is used to estimate the mode number of the line. The precise wavelength is then calculated using

$$f_n = n \times f_r + f_0 \quad (4)$$

where the repetition rate  $f_r$  is known from the design of the LFC and the offset frequency  $f_0$  has been determined by Menlo Systems, the manufacturer of the LFC.

In order to comfortably satisfy the assumption that there exists only low-order variation, which is needed for *excalibur*, we use exposures from after the LFC stabilized following servicing in summer 2019, when the photonic crystal fiber was replaced and the polarization was changed to shift the wavelength range of the LFC redward. In the results presented here, we use 1227 LFC exposures and 78 ThAr exposures taken between 2019 October 14 and December 18 on 29 unique nights.

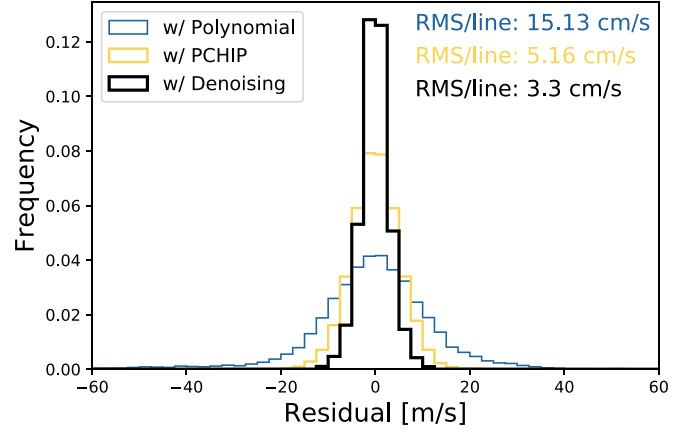
#### 4. Tests

We perform a series of tests to validate the performance of *excalibur* and benchmark *excalibur*-generated wavelengths against wavelengths generated by a classic, nonhierarchical, parametric method. To implement training/validation tests, we leave out a subset of calibration lines with known wavelengths as a “validation” sample, generate wavelengths for these lines using the remaining data, and compare the predicted wavelength to the assigned wavelength of each line. This inherently folds in errors in the measured line center of each calibration line, but this contribution to the residuals will be the same across all tests.

To assess the classic, polynomial-driven method of wavelength calibration, we take each LFC exposure and separate the lines into even- and odd-indexed lines. We then construct a wavelength solution using only the odd-indexed lines and use that wavelength solution to predict wavelengths for the even-indexed lines; i.e., a polynomial is fit to just the odd-indexed lines and then evaluated at the detector positions of the even-indexed lines (see Equation (2)). We then generate a wavelength solution using only the even-indexed lines and use it to predict wavelengths for the odd-indexed lines.

To test the interpolation step of *excalibur* (Section 2.3), we employ *excalibur* on all LFC exposures with odd-indexed lines removed. The resultant basis vectors  $g_k(x, y, m)$  and amplitudes  $a_{n,k}$  are therefore only informed by the even-indexed lines of each LFC exposure. We then predict wavelengths for the odd-indexed lines that have been excluded, and compare these predictions to their assigned wavelengths. This allows us to test how accurately an interpolated wavelength solution can predict wavelengths.

To test the denoising step of *excalibur* (Sections 2.1 and 2.2), we employ *excalibur* on a randomly selected 90% of all LFC exposures. This means the basis vectors  $g_k(x, y, m)$  and weights  $a_{n,k}$  are constructed using information from only 90% of all exposures. We use the results to predict wavelengths for all the lines in the remaining 10% of calibration exposures.



**Figure 2.** Difference in predicted and theoretical wavelengths for the wavelength calibration tests described in Section 4. The per-line rms as defined in Equation (5) is given in the top-right corner in each method’s corresponding color. Incorporating denoising returns the smallest spread in residuals.

This allows us to test how well we can pinpoint the calibration state of the spectrograph using *excalibur*.

The polynomial and interpolation tests remove the same 50% of lines from each exposure, while the denoising test completely removes a randomly selected 10% of calibration exposures and their associated line position measurements. Errors from interpolation will be localized, extending only to neighboring lines. We therefore aggressively remove every other line to ensure we capture these local effects. The PCA denoising, on the other hand, folds in information of all lines from all exposures. Here, it is sufficient to completely remove 10% of the exposures, a traditional training/validation fraction. Since the information being removed varies between each test depending on its focus, we present the results per line, treating each line as an independent test.

The residuals of a wavelength solution represent the difference between the wavelength solution evaluated at the line position of a calibration line and the assigned theoretical wavelength (i.e., that from Equation (4) for the LFC lines) on a line-by-line basis in every exposure. The reported rms of a wavelength solution is therefore the per-line rms, i.e.,

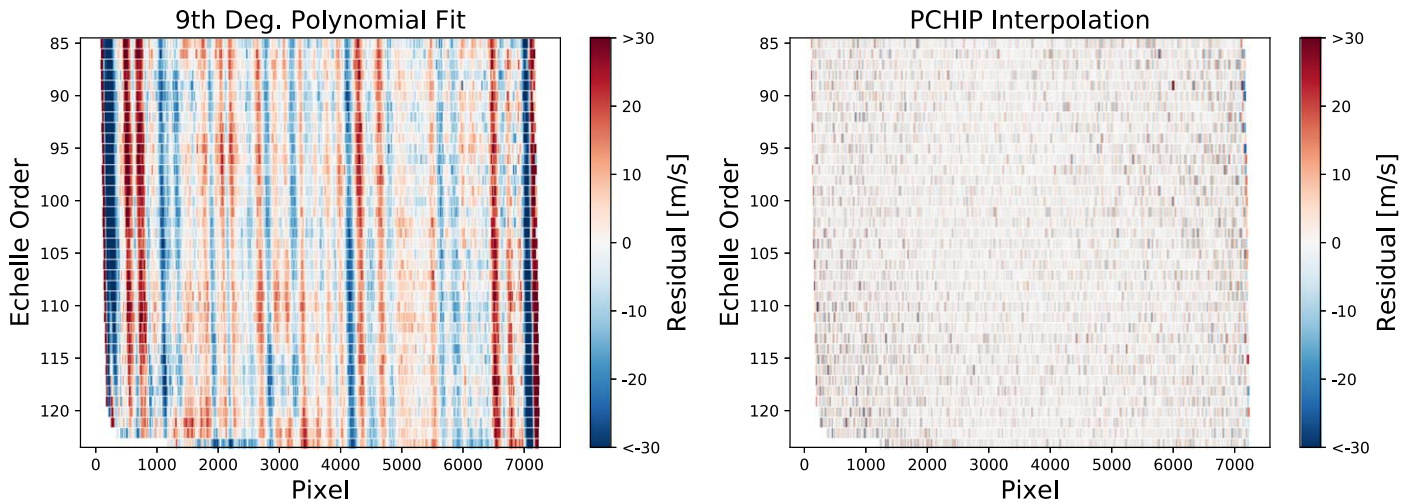
$$\text{RMS/line [m s}^{-1}] = \sqrt{\frac{\sum_{n=1}^N \sum_{p=1}^P \left[ \frac{(\lambda_{n,p,\text{pred.}} - \lambda_{p,\text{theory}}) \times c}{\lambda_{p,\text{theory}}} \right]^2}{N \times P}} \quad (5)$$

where  $\lambda_{p,\text{theory}}$  is the theoretical wavelength for line  $p$ ,  $\lambda_{n,p,\text{pred.}}$  is the wavelength predicted by the constructed wavelength solution for line  $p$  in exposure  $n$ , and residuals from all  $P$  lines from all  $N$  exposures are used, for a total of  $N \times P$  lines. The difference in wavelength is converted to units of meters per second, a more intuitive metric for EPRV work.

##### 4.1. Results

Histograms of the per-line residuals for each of the above-described polynomial, interpolation, and denoising tests are shown in Figure 2. Note that the spread in residuals is much smaller for both the denoising and interpolation tests relative to the results of the polynomial wavelength solution.

The per-line residuals from the denoising test also exhibit smaller spread than those from interpolation alone. This



**Figure 3.** Residuals of a single LFC exposure plotted with respect to detector position (as defined by echelle order and  $x$ -pixel) for parametric (left) and nonparametric (right) wavelength calibration methods. Each line is colored by the difference between the predicted wavelength and the theoretical wavelength for each line, given in units of meters per second. High-order structure, i.e., vertical stripes and patchiness, is apparent in the residuals to a polynomial wavelength solution, which assumes smoothness.

suggests that the spectrograph truly is accurately represented by a low-dimensional model. Recreating line positions using this model gives better line position estimates than treating each exposure independently. The low-dimensional model does not incorporate noise from individual line measurements. Returning more precise denoised line positions results in smaller per-line residuals.

*Excalibur*-generated wavelengths also exhibit less structure in the returned residuals. For a randomly selected example LFC exposure, Figure 3 plots each line with respect to its echelle order ( $y$ -axis) and  $x$ -pixel on the detector ( $x$ -axis) colored by the difference between the predicted and theoretical wavelengths for that line in units of meters per second.

The residuals of the classic, polynomial wavelength solution are shown in the top plot of Figure 3. There are a lot of vertical structures and some hints of periodic diagonal structures as well. The residuals of the interpolation test for the same exposure are shown in the bottom plot of Figure 3. There are no coherent structures here and smaller residuals.

This shows how the flexibility of an interpolated model can account for high-order instrument or detector defects, which emerge as structures in the residuals of the classic, smooth, polynomial-driven wavelength solution. This same flexibility may similarly allow interpolated wavelength solutions to account for position errors in pixel image blocks for different detectors depending on how the interpolation is framed (Fischer et al. 2016; Milaković et al. 2020).

Though the interpolated wavelength solution returns lower, less structured residuals than the polynomial wavelength solution when guided by LFC lines, the flexibility of an interpolated wavelength solution can result in much worse residuals when not properly constrained—for example, in regions between widely separated calibration lines. The left plot of Figure 4 shows the residuals when ThAr calibration lines, which are much fewer and less regularly spaced than LFC lines, are run through *excalibur* and used to predict wavelengths for the (completely independent) LFC exposures taken during the same range of time. Overplotted in yellow are the positions of the ThAr lines.

Note that running *excalibur* informed by only ThAr lines cannot be regarded as a direct comparison to the LFC runs, as

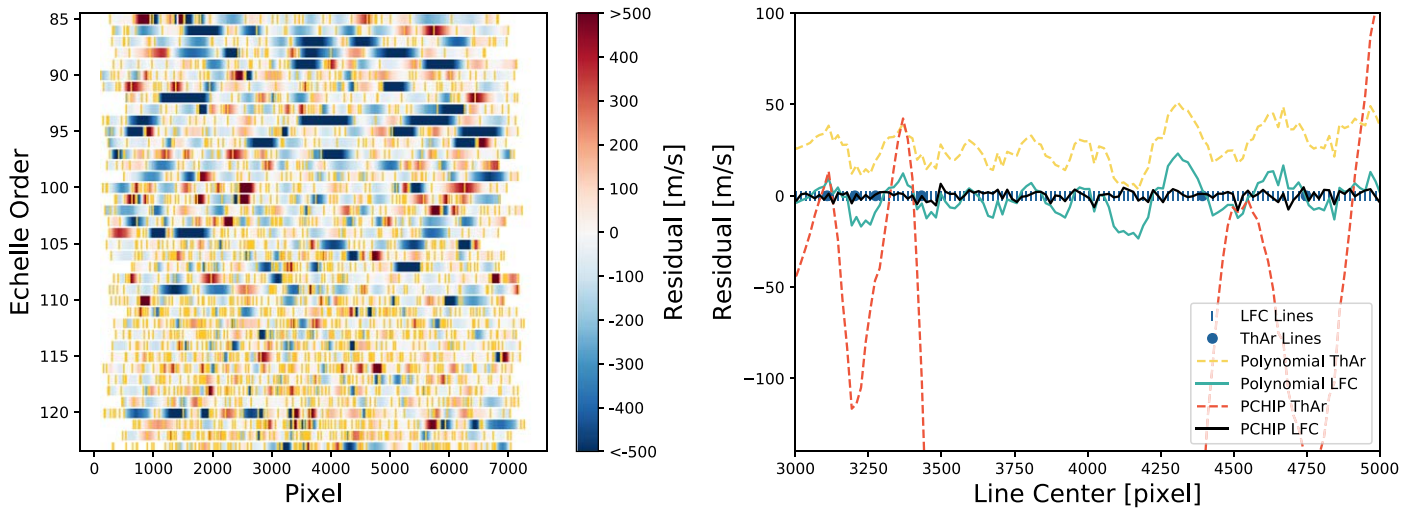
the increased uncertainty and variability in using ThAr line positions alone makes the resultant wavelength predictions an order of magnitude worse—hence the different scale of the colorbar in the left plot of Figure 4 as compared to Figure 3. All the same, the residuals are in general worse where lines are further apart (for example, in redder echelle orders) than where lines are denser.

Figure 4 (right) plots the residuals for a subset of order 94 for both a polynomial-based method and a PCHIP-based method guided by either ThAr lines or LFC lines. The PCHIP model with ThAr lines (orange, dashed curve) returns huge residuals between two widely separated ThAr lines that extend out of frame. The classic, polynomial fit exhibits similar residuals in both amplitude and shape whether the set of ThAr lines or the set of LFC lines is used. An interpolated wavelength solution using LFC lines (black, solid curve) exhibits the lowest residuals.

The move to an interpolated wavelength solution is driven by the assumption that a high density of calibration lines allows for more freedom in the resultant wavelength solution. This freedom allows the wavelength solution to more accurately ascribe wavelengths. This flexibility, however, is no longer justified in the regime where there are large separations between calibration lines, as this no longer provides sufficient constraint on the interpolated wavelength solution, as is the case in some regions of a classic ThAr lamp.

#### 4.2. Impact on RV Measurements

We test *excalibur*-generated wavelengths on RV measurements using EXPRES observations of HD 34411, which are presented in Table 1. HD 34411 is a G0V star. It is 4.8 Gyr old and relatively quiet ( $\log R'_{\text{HK}} = -5.09$ ; Brewer et al. 2016). Because HD 34411 has no known planets and should have a smaller contribution from stellar signals, it is a good star with which to test the effects of wavelength calibration on the rms of the returned RVs. We use 114 observations of HD 34411 taken between 2019 October 8 and 2020 March 5 with S/N of 250. RV measurements are derived using a chunk-by-chunk, forward-modeling algorithm run by the EXPRES team (Petersburg et al. 2020).



**Figure 4.** Residuals when using ThAr lines to predict wavelengths for LFC lines. Left: Residuals for a single LFC exposure plotted with respect to detector position and colored by residual, as in Figure 3. The positions of ThAr lines are overplotted in yellow. In general, residuals are greater between ThAr lines with greater separation. Right: Comparison of polynomial and interpolated wavelength solutions using either just ThAr lines or just LFC lines for a subset of echelle order 94. The shape of the residuals from a polynomial fit is similar whether using ThAr lines or LFC lines. A PCHIP-interpolated wavelength model guided by LFC lines returns the smallest residuals.

**Table 1**  
EXPRES RVs Obtained Using *Excalibur* Wavelengths

JD-2,440,000	RV ( $\text{m s}^{-1}$ )	Error ( $\text{m s}^{-1}$ )
18,764.4771	3.139	0.335
18,764.4791	1.035	0.332
18,764.4810	3.074	0.324
18,771.4179	-1.927	0.342
18,771.4196	2.688	0.357

(This table is available in its entirety in machine-readable form.)

Figure 5 compares the resultant RVs when using a classic, ninth-degree polynomial wavelength solution and an *excalibur*-generated wavelength model. Using *excalibur*-generated wavelengths reduces the rms of the entire data set from  $1.17 \text{ m s}^{-1}$  with the classic wavelength solution to  $1.05 \text{ m s}^{-1}$ . This is equivalent to removing an independent, additive noise component of  $0.52 \text{ m s}^{-1} (= \sqrt{1.17^2 - 1.05^2})$ .

We conduct a direct test of a classically generated wavelength solution with *excalibur* wavelengths on four other data sets. All targets show a reduced or comparable RV rms. However, the results from these data sets cannot be interpreted as directly as those with HD 34411, due to larger contributions from stellar variability, known planets, etc. As completely mitigating these different effects is out of the scope of this paper, we focus here on the results with HD 34411.

## 5. Choose Your Own Implementation

We have described and tested only one implementation of *excalibur*. Using PCA and an interpolated wavelength solution is a statistically straightforward step toward a complete hierarchical and nonparametric wavelength model. It is possible to upgrade both the denoising and wavelength solution steps to true models. It is also possible, of course, to implement either step individually. A hierarchical framework can be used to simply denoise the lines before they are fit to a parametric model, and a nonparametric model can be used on lines that have not been denoised.

For dimensionality reduction and denoising, PCA could be replaced by a probabilistic PCA model or other probabilistic linear reduction methods, such as heteroskedastic matrix factorization (Tsalantza & Hogg 2012). It is also possible to move to nonlinear reduction methods, like a Gaussian-process latent variable model, an autoencoder, or a normalizing flow (e.g., Kramer 1991; Woodbridge et al. 2020). Using a nonlinear denoising model could enable *excalibur* to capture large-scale changes as well as small variations in calibration state.

The wavelength solution could also move past interpolation. For example, a Gaussian process could be used that is constrained to ensure monotonicity. Replacing each step with a model will allow for full, hierarchical Bayesian inference. This means the uncertainty from wavelength calibration could be completely marginalized out. Doing so will have the largest impact if the wavelength calibration is a significant fraction of the error budget.

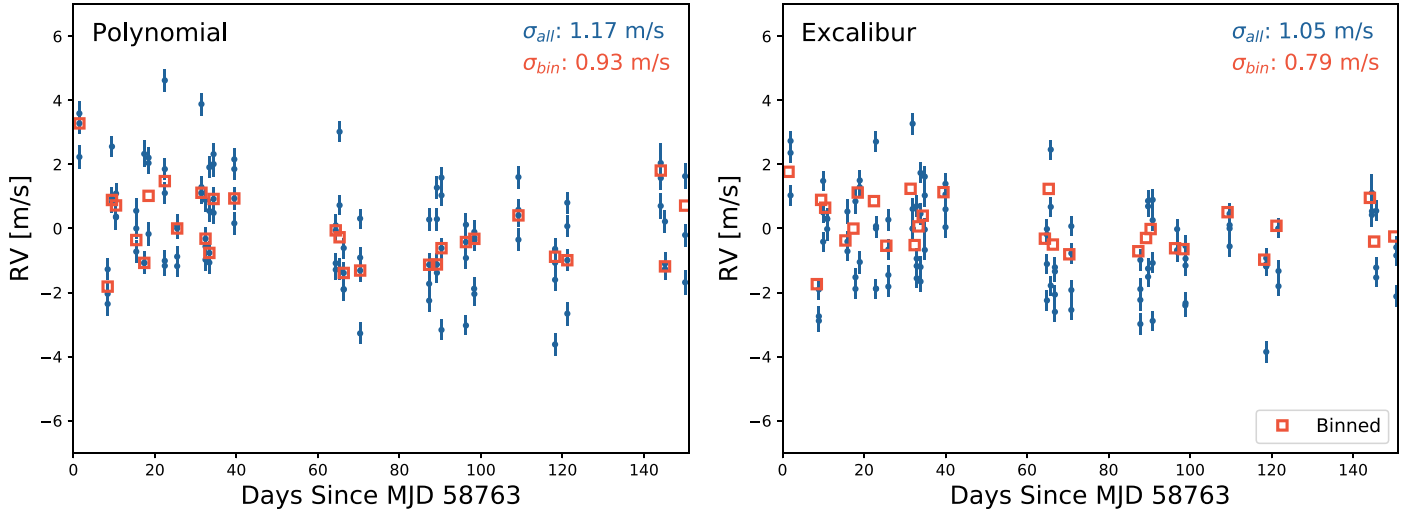
The implementation of *excalibur* presented here, using PCA for denoising and interpolating a wavelength solution, uses various global variables and methods that we believe are optimal or close to optimal for constructing a high-fidelity wavelength solution. The following subsections will describe each choice and the associated decision-making process.

### 5.1. Dimensionality of the Calibration Space, $K$

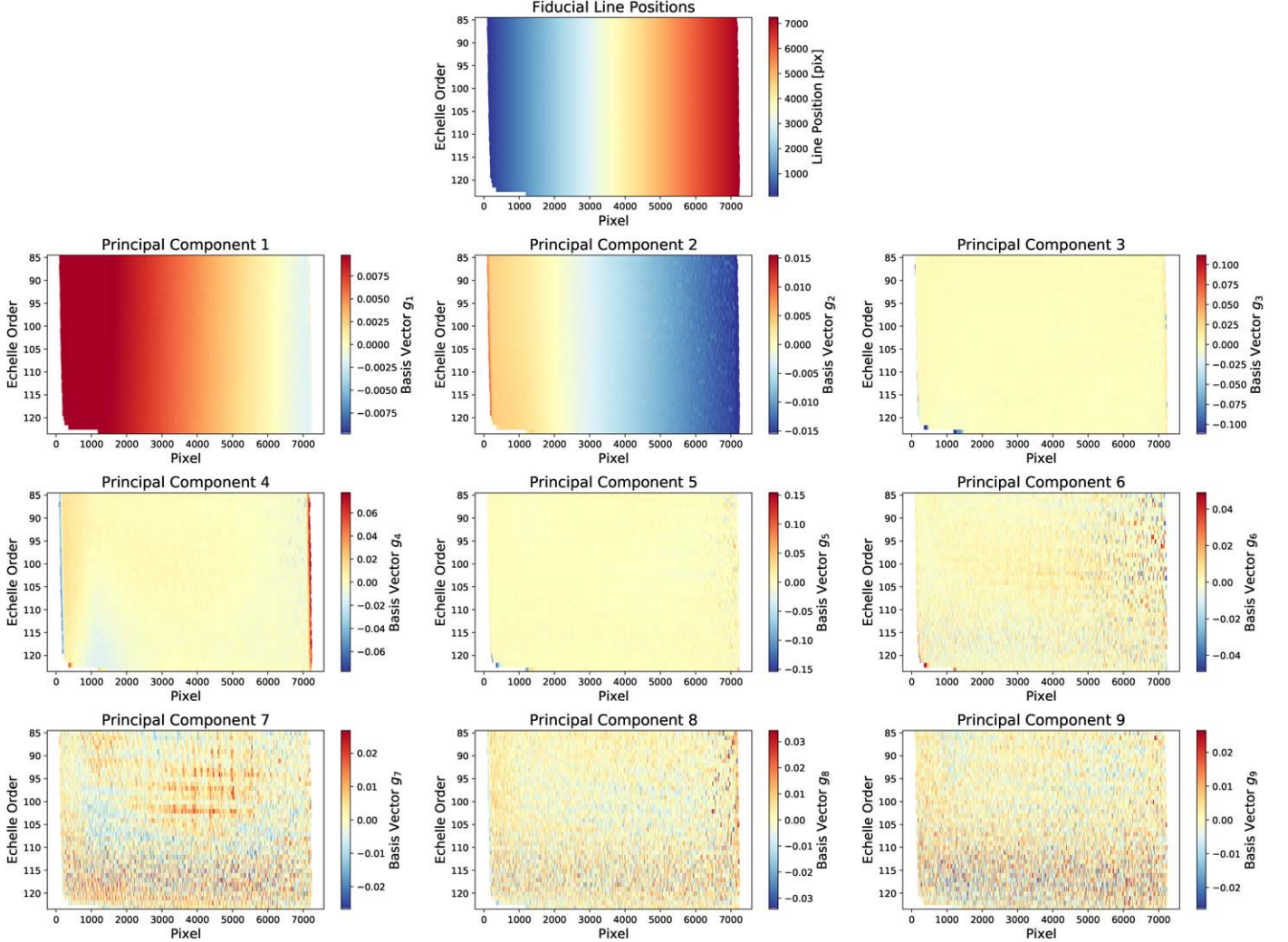
The dimensionality of the calibration space where the spectrograph is located is represented by  $K$ . In practice, it is the number of principal components, or basis vectors, used to reconstruct the denoised line centers.  $K$  needs to be large enough so that all variability in the spectrograph is captured. Too large, however, and the reconstruction will begin to incorporate noise, thereby defeating the purpose.

Figure 6 shows the fiducial calibration state and the first nine principal components constructed using LFC lines, which represent deviations from the fiducial calibration state. There is clear structure in the first and second principal components. Components three through six show smaller or more localized structures. Components three and four have aberrant behavior



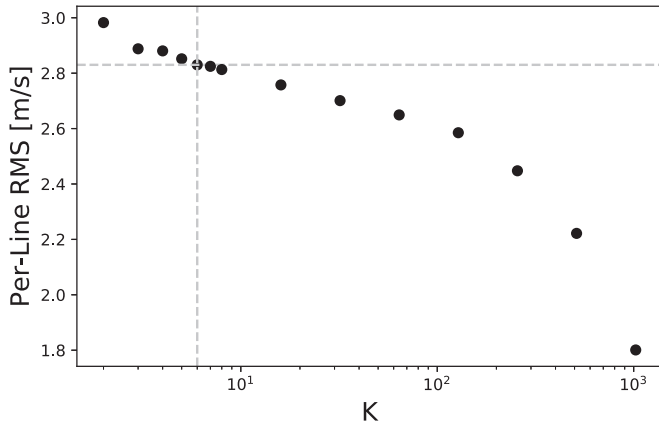


**Figure 5.** HD 34411 RVs measured with EXPRES. We zoom in on all but eight of the exposures, which were taken over 150 days later. All RVs are shown in blue; nightly binned RVs are overplotted as orange squares. The rms of the data set and the rms of the binned values are given in the top-left corner. Left: RVs derived using a polynomial-based wavelength solution. Right: RVs derived using wavelengths from the implementation of *excalibur* presented in this paper.



**Figure 6.** Top: The fiducial calibration of the spectrograph, i.e., the mean line positions for each line throughout the epoch of stability. The following  $3 \times 3$  grid of plots shows the first nine principal components constructed using LFC lines. These principal components represent the basis vectors along which the calibration of the spectrograph can deviate from the fiducial calibration. For each principal component, or basis vector, each calibration line is plotted according to its echelle order and  $x$ -pixel and colored by the value of the basis vector for that line. Principal components beyond the sixth one steadily become more dominated by noise.





**Figure 7.** Per-line rms of returned wavelength models for different values of  $K$ . The per-line rms, as defined in Equation (5), provides a measure of the accuracy of a wavelength model. There is a dotted vertical line at  $K = 6$ , and the dotted horizontal line is the rms for  $K = 6$ . The improvement around  $K = 6$  plateaus.

on the edges of the two bluest echelle orders, where a lower signal results in more variation in the line fits. Later principal components become dominated by noise and show less coherent structures.

In deciding a  $K$  value, we run denoising tests, as described in Section 4, for  $K$  values spanning 2 to 512. The resultant per-line rms for each test is plotted in Figure 7. One would expect the returned wavelengths to get increasingly more accurate with larger  $K$  until components that represent only noise are incorporated. Residuals might then get worse before ultimately starting to get better again with large  $K$ , which marks when the model starts overfitting the data. Though the returned rms never gets worse, we find that the improvement plateaus between  $K = 6$  and  $K = 128$ . Comparisons of wavelengths returned by a  $K = 6$  model versus a  $K = 32$  model show significant differences in less than 10 blue lines, which are known to have greater error and variance in their measured line positions. We therefore settle on a  $K$  value of 6.

### 5.2. Interpolation of Calibration State to Science Exposures

Figure 8 shows the amplitudes of the first and second principal components with respect to time on the left. Though there exists a complex overall shape to the amplitudes with respect to time, a clear linear trend exists within each night. This is shown by the right plots in Figure 8. As the beginning-of-night and end-of-night calibration sets always include LFC exposures, we use a simple linear interpolation to interpolate principal-component amplitudes with respect to time.

The choice of interpolation method can help identify how many wavelength calibration images are truly needed. It is unnecessary to take calibration images at times when the same information can be reconstructed at the desired precision by a well-chosen interpolation scheme. For example, with the EXPRES data shown here, it is clear that nightly calibration images are needed, but for a linear trend, only two calibration images throughout the night are strictly required.

We also test an implementation of *excalibur* where the  $K$  principal components within a night are fit to a cubic with respect to time rather than being linearly interpolated. This emulates the current, polynomial-based wavelength solution implemented in the EXPRES pipeline, where polynomial fits to calibration files are interpolated to science exposures by fitting

polynomial coefficients with respect to time to a cubic. We find that using a cubic in place of linear interpolation returns a comparable RV rms for most targets, though it appears to do better when a night has sparse calibration data. This suggests that the nightly behavior of EXPRES with respect to time is well described by a cubic function, but LFC exposures are typically taken with enough frequency that a linear interpolation provides a good approximation (see Figure 8).

The amplitudes  $a_{n,k}$  can also be interpolated with respect to any good housekeeping data, not just time. The best results will come from interpolating with respect to whatever is most strongly correlated with the calibration state of the spectrograph. For example, with EXPRES, which is passively temperature-controlled, the returned amplitudes  $a_{n,k}$  are extremely correlated with the optical bench temperature, as shown in the top-left plot of Figure 8, suggesting it would also be possible to interpolate the amplitudes with respect to temperature.

Another pertinent example would be a spectrograph that is mounted on a telescope and therefore moves with the telescope. In this case, it may be important to interpolate at least in part with respect to the position of the telescope, which enables the resultant calibration to incorporate the gravitational loading experienced by the spectrograph.

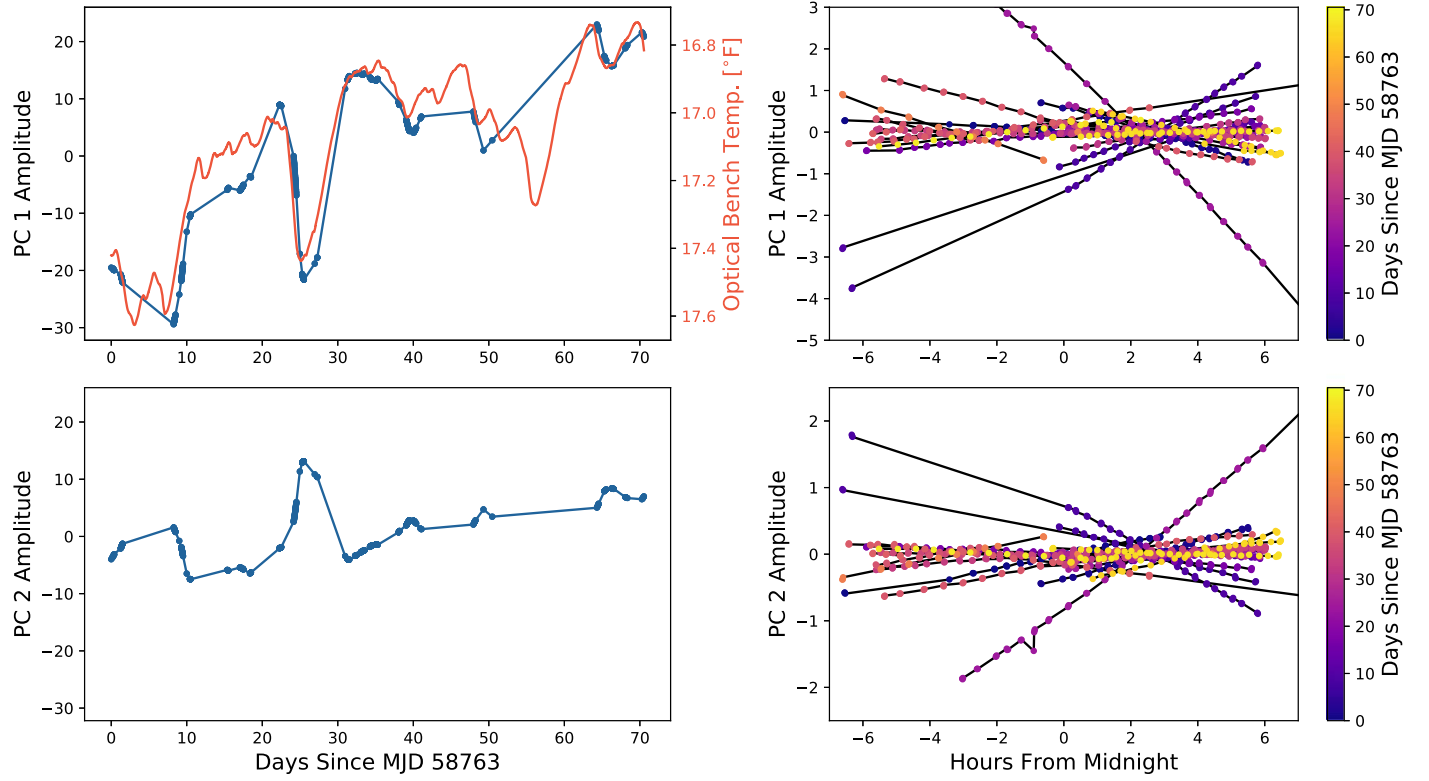
### 5.3. Interpolation of Wavelengths with Respect to Pixels

In the implementation described and tested by this paper, interpolation of wavelengths over pixels is done order by order using a PCHIP interpolator. This interpolation incorporates the flexibility needed to model the changing dispersion of the spectrograph across an echelle order along with any detector defects while also enforcing monotonicity, which we know must be true across any one echelle order.

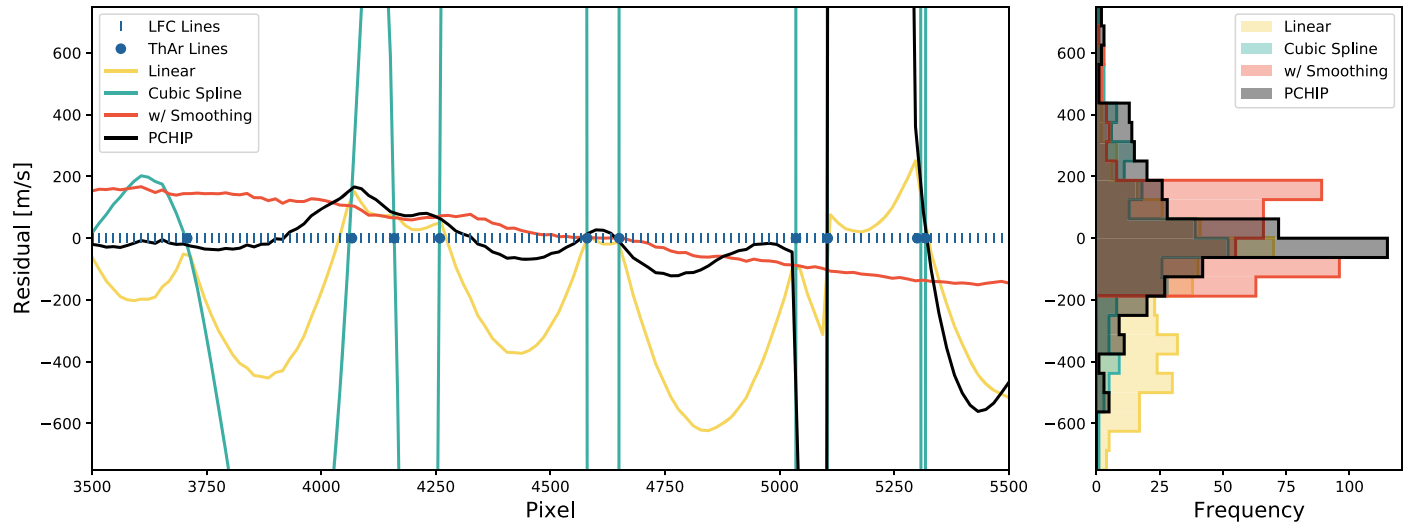
A simple linear interpolation would give erroneously low values everywhere. Due to the dispersion intrinsic to echelle spectrographs, the wavelength change between pixels grows greater with greater wavelengths, even within an order. This means that the function of wavelength versus pixel across an echelle order will always be monotonically increasing and concave down everywhere.

Though less of an issue with LFC lines, a more classic cubic-spline interpolation can run into issues with arc lines, which are irregularly spaced or even blended. Close lines appearing at virtually the same pixel location but with different wavelengths could coerce a cubic spline into a very high slope. This is demonstrated by the green line in Figure 9, which shows the results of interpolating between ThAr lines rather than between LFC lines. A line blend appears at approximately pixel 5300, causing the spline to twist to nearly vertical to account for both points. This leads to huge deviations from the correct wavelengths around this line blend as the extraneously high slope of the spline is accounted for.

These huge digressions can be avoided by allowing for some smoothing in the interpolation. In Figure 9, we show an example in orange using SciPy's implementation of a univariate spline. While the result appears to follow the calibration lines much better, the smoothing ultimately causes larger residuals that are spatially correlated (Figure 9, right). In all echelle orders, the edges are overestimated while the middle is underestimated, shown by the flattened shape of the histogram of residuals. The resultant wavelength solution underestimates the curvature of the pixel-wavelength relation,



**Figure 8.** Amplitudes of the first two principal components shown as a function of time (left) or hours from midnight (right). The top row of plots shows the amplitudes for the first principal component, while the bottom row shows the amplitudes for the second principal component. Lines show the result of a linear interpolation. In the top-right plot, the temperature of the optical bench is also plotted, in orange. In the right plots, the principal-component amplitudes for each night are artificially offset by the median amplitude per night. All days are therefore roughly on the same scale, but the y-axis is different from that on the left plots. In the right column, points are colored by the MJD of each exposure.



**Figure 9.** Residuals from different interpolation schemes over pixels in echelle order 100. ThAr lines, shown as blue circles, are used to construct a wavelength solution that is then evaluated at each LFC line, shown as blue vertical lines. The residuals of each wavelength solution for a subset of the order are shown on the left. Histograms of the residuals for each method for the complete order are shown on the right. Note: There is a blended ThAr line at approximately pixel 5300, the rightmost ThAr line plotted.

giving rise to issues similar to those with an inflexible, parametric wavelength solution. Introducing this smoothing parameter enforces a smoothness we have no reason to believe is true of the data, thereby reintroducing one of the problems encountered with parametric models.

We instead turn to the PCHIP algorithm, which damps down huge deviations in the traditional cubic spline by requiring the

resulting interpolated function to be monotonic. Monotonicity is a constraint we know must be true for any one echelle order. The PCHIP interpolator, like the classic cubic spline, shows issues of a ThAr line blend around pixel 5300, but the effect is much smaller and is exerted on fewer pixels. Figure 9, right, shows that using the PCHIP interpolator returns the lowest-spread residuals.

There likely exists an even more fitting model between an overly constrained polynomial fit and a completely free spline interpolation. For example, there has been success in interpolating wavelengths with respect to pixels using a segmented polynomial in the dispersion direction, especially when tuned to known detector defects (Milaković et al. 2020). Stiffer, more constrained splines or carefully chosen knot positions may afford the perfect marriage of freedom and constraint that will better describe wavelength changes with pixels.

## 6. Application to Other Spectrographs

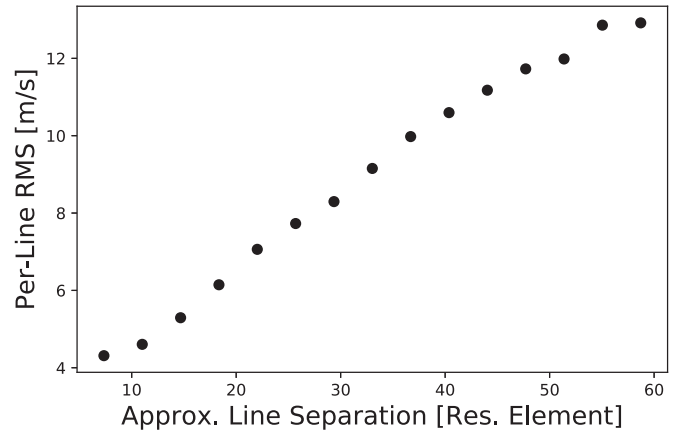
We focus on an EPRV use case here because there is a strong need for wavelength calibration in the EPRV community. EXPRES is representative of the newest generation of EPRV spectrographs, and an LFC provides stable, dense calibration lines with known wavelengths, ideal for *excalibur*. The applicability of *excalibur* to any one instrument is a detailed question of the kind of variance experienced by the spectrograph and calibration sources available, but we hope to provide some approximate benchmarks and diagnostics here.

Implementing *excalibur* will require an existing library of calibration exposures that span the range of calibration space accessible by the spectrograph, or at least the space accessed by the science exposures being calibrated. A new set of calibration exposures would be needed for any changes to the instrument that could significantly change the accessible calibration space. While there is no exact cutoff for how many exposures are needed, the number is certainly much larger than  $K$ , the dimensionality of the calibration space. More calibration exposures will help with denoising. The number of calibration exposures required throughout a night will depend on how much the instrument varies throughout the night as well as on the chosen interpolation scheme, as mentioned in Section 5.1.

As an empirical assessment of the required line density, we remove LFC lines from the EXPRES data and calculate the per-line rms of the returned wavelengths for the removed lines. These tests are similar to the interpolation test explained in Section 4, in that a fraction of lines are systematically removed from the analysis. An increasing fraction of lines are removed to simulate different line densities. For these line density tests, though, we also implement denoising, unlike the pure interpolation test of Section 4.

Figure 10 plots the scatter of the residuals of the wavelengths returned by *excalibur* as a function of the separation between lines in units of resolution elements. It gives an approximate estimation of the expected error between lines of different separation. Note, however, that the stability of the lines and their measured line centers quickly becomes the dominant error source in the calibration lamps over the separation between lines. Additionally, in this assessment, the lines remain uniformly spaced. The required line density depends on the resolution of the given spectrograph, the required precision of the returned wavelength solution, and the chosen interpolation scheme.

This implementation of *excalibur* to EXPRES data interpolates calibration information to the science data using surrounding calibration exposures. Simultaneous calibration data can be used to reinforce the determined calibration state of an instrument at the time of the science exposures. This simultaneous calibration data can come from calibration light



**Figure 10.** Per-line rms as a function of spacing between lines in units of resolution elements. The average line spacing is calculated using the average distance between LFC lines across the wavelength range.

shone through a reference fiber or any metadata (e.g., temperature, pressure, time) that correlates with the calibration. For example, we have seen that the calibration state of EXPRES is correlated with the temperature of the optical bench, even with the optical bench temperature varying by less than one degree. This correlation is not seen in the RVs, suggesting changes with temperature are calibrated out at the wavelength stage.

With a simultaneous reference fiber, the position of calibration lines taken through the reference fiber can simply be concatenated to the array of line positions taken through the science fiber. Both sets of lines will then be used when the low-dimensional basis is constructed. This allows the simultaneous calibration information to contribute to the construction of the complete calibration space of the spectrograph and pinpoint where the spectrograph is in that calibration space for any exposure.

The calibration for any science exposure with a simultaneous reference can be determined by finding the amplitude of each basis vector that most closely recreates the calibration line positions through the reference fiber. These amplitudes can then be used to recreate the calibration line positions through the science fiber as well. This replaces the need to interpolate the basis vector amplitudes from calibration exposures to science exposures, something that is done with respect to time in the example implementation described in this paper. The result is likely to be even more precise, as this method incorporates more data. This method, like all other analysis involving a simultaneous reference fiber, will work only as well as the reference fiber’s ability to trace changes in the main science fiber.

It is also possible to apply *excalibur* to etalon data with some modifications. The simplest implementation is if the free spectral range (FSR) and therefore the wavelength of each line of the etalon are well known. The etalon data can be interpolated onto a set of fiducial lines with set wavelengths. These fiducial lines would therefore be identifiable by echelle order and wavelength alone, with only their line positions varying with different exposures. This returns us to the same framework as developed for the case of an LFC. This marks the simplest implementation of *excalibur* on etalon data, as the



uncertainty of a line’s wavelength is upstream of the model rather than built in.

Incorporating the FSR as part of the *excalibur* model will require introducing a free parameter to capture changes in the FSR independent of the variation in an instrument’s calibration state. The calibration state can then be described with respect to the mode number, which will be used to uniquely identify a calibration line across exposures rather than the wavelength. The FSR is then used to determine how the mode number of each line maps to wavelength for a given exposure. The FSR must not vary so much that the change in this mode-number-to-wavelength mapping becomes nonlinear. This model would require a simultaneous reference or other housekeeping data that can be used to determine the FSR for every exposure.

In terms of dimensionality reduction, most physical systems should have only a few dominant axes along which they vary, meaning *excalibur* should be adaptable to a wide range of instrument designs. With PCA, this can be tested by plotting the amplitude of the returned principal components, which should fall quickly after a few components. It should be noted that this only provides a measure of the variance in the PCA space and is not an explicit test of end-to-end variation in the resulting model. This condition is therefore necessary but not sufficient if implementing *excalibur* with PCA.

It could still be possible to run *excalibur* on a spectrograph that has a high-dimensional calibration space, meaning a large number of basis vectors are required to capture all the variance in the spectrograph. In this regime, there is always the risk of overfitting. Regularizing the principal-component amplitudes—for example, by insisting the amplitudes for higher principal components be smaller—can help to return reasonable models (Foreman-Mackey et al. 2015). Within such a framework, *excalibur* may still deliver good results.

For the results presented here, the data is broken up into different periods of stability based on where the principal-component amplitudes show huge deviations. This is done visually, though there exist many change-point detection algorithms that could be used (Aminikhanghahi & Cook 2017). There is a trade-off in including more exposures between introducing greater variation, but also more data to offer constraints that may be optimized. Here, a period of stability is chosen manually in order to focus on separating out time-domain intervals in which the data is relatively homogeneous, e.g., when most exposures show the same calibration lines. Homogeneity is, of course, implicitly required when implementing PCA. Different denoising models will be able to account for different amounts of stability or a lack thereof.

Lastly, we caution that *excalibur* is extremely sensitive to upstream changes that may affect the line centers. For example, PCA is good for detecting variability but is agnostic to the source of the variability. This is why the principal components shown in Figure 6 exhibit errant values for bluer LFC lines, which have lower signals and therefore exhibit more variation in their fitted line centers. It is essential that the line positions being fed to *excalibur* capture only the changes in the spectrograph’s calibration state, not potential errors in the fitted line centers.

## 7. Discussion

We show that *excalibur* returns a lower per-line rms than classic, parametric methods by a factor of 5 (Section 4). The

residuals are also smoother, exhibiting less spatial correlation (Figure 3). Using *excalibur* wavelengths reduces the rms in the RVs of HD 34411 from  $1.17 \text{ m s}^{-1}$  to  $1.05 \text{ m s}^{-1}$  (Section 4.2).

In implementing *excalibur* on EXPRES data, we have successfully constructed a model of EXPRES’s accessible calibration space, confirming that EXPRES truly is an instrument with low degrees of freedom. *Excalibur* does not make any claims about what variability each basis vector represents. Those interested in interpreting the variability are encouraged to investigate how the amplitude of the different vectors varies with different housekeeping data to find its source.

Starting with a list of calibration lines with known wavelengths and well-fit line centers for each calibration exposure, *excalibur* will denoise and interpolate the given lines into a full wavelength solution. *Excalibur* leverages the stability of contemporary EPRV instruments and the high density of lines made available by new calibration sources, such as LFCs and etalons, to achieve more accurate wavelengths. *Excalibur* therefore assumes dense enough calibration lines to properly constrain a nonparametric wavelength model and assumes that the instrument has low degrees of freedom.

Denser calibration lines allow us to move to more flexible wavelength models, which can then account for nonsmooth features in the wavelength solution. Stabilized spectrograph hardware makes it more likely that the calibration space of the instrument is low-dimensional. All calibration images in a given period of stability can therefore be used to constrain the accessible calibration space of the spectrograph as well as where in that calibration space the spectrograph lies. We have described only one, fairly simplistic implementation of *excalibur* here. There are many other options for both the denoising and the interpolation steps, as mentioned in Section 5.

An advantage of this implementation of *excalibur*, where PCA is applied to line positions from all LFC exposures, is the ability to isolate exposures that exhibit errant variation, which is typically associated with flawed exposures. This allows us to quickly vet for problematic LFC exposures, which otherwise would have required visual inspection of all 1200+ LFC exposures. In a classic framework where each calibration exposure is treated independently, these aberrant exposures would likely have persisted undetected and are liable to sway the resultant wavelength solutions for at least an entire night.

On the other hand, PCA captures all variance, regardless of source. Though *excalibur* endeavors to capture only variation in the instrument, PCA is also sensitive to uncertainties and failures in the upstream line position fitting. For example, we have seen that lower-signal lines that are harder to accurately fit have greater variety in returned line positions, which is in turn captured by the PCA. In this sense, *excalibur* is actually a model of not just the instrument but all the upstream choices used to drive line positions. High-fidelity line positions are essential to ensuring the PCA is capturing variations in just the spectrograph’s calibration rather than including changes in how well a line can be fit or other effects.

Along these lines, we caution that with any wavelength solution, there is a perfect degeneracy between what is defined as the “position of the line” and the resultant wavelength solution. If, for example, a cross-correlation method is used to extract RVs from the data, a systematic difference may be

introduced depending on what exactly is defined to be the line position, whether it be the mode of the Gaussian fit, the first moment of a complicated point-spread function, or the output of a particular peak-finding algorithm. In principle, the best way to mitigate this uncertainty would be to use a calibration source that looks like a star.

Compared to traditional methods, which involve fitting high-order polynomials, *excalibur* has several useful statistical properties. *Excalibur* is technically robust to localized issues that arise from either the calibration source or the pipeline used to return line positions. With an interpolated wavelength model, one errant line position will only affect the resultant wavelength model out to close, neighboring lines. The effect of an outlier is diminished and kept localized. In contrast, an entire parametric fit will be affected by a single errant line, changing the wavelength solution for the entire exposure for a 2D fit. Through denoising and outlier rejection, *excalibur* adds additional robustness against erroneous line positions.

The locality of the interpolation carries other benefits as well. Manufacturing artifacts in the detector or other optical elements can lead to nonsmooth structures in the wavelength solution that cannot be captured by polynomials or other smooth functions (see Figure 3). An interpolated model introduces greater flexibility, accounting for such high-order effects. As discussed in Section 5.3, there are better and worse interpolators for the task, which may differ for different instruments and different calibration sources. Instead of using an interpolator at all, there might be better results from implementing something more sophisticated, such as a kernel method or a Gaussian process with a kernel adapted for the specifics of an instrument. There are in principle an enormous number of nonparametric methods to explore, which we leave outside the scope of this paper.

Similarly, PCA is just one of many possible dimensionality reduction methods. We choose to implement *excalibur* using PCA here for simplicity and computational tractability. PCA is a good option because the instrument changes here are small enough that a linear model is an appropriate representation of the changes. If *excalibur* is updated to a full probabilistic model, the PCA along with the interpolation model will have to be upgraded to something with better probabilistic properties. Other, nonlinear denoising methods may be more robust to large changes, allowing all calibration images ever taken with an instrument to be used to construct the accessible calibration state regardless of hardware adjustments. Further discussion of other implementations of *excalibur* can be found in Section 5.

*Excalibur* can be applied to any data that contains information about the calibration state of the spectrograph (see Section 6). For example, though LFC and ThAr exposures are used as examples in this paper, *excalibur* would work similarly for an etalon or any other arc lamp with a list of lines and assigned wavelengths. Simultaneous calibration information can easily be accounted for by simply including the line position information from the simultaneous reference when constructing a low-dimensional basis of the instrument's calibration space.

Once we have defined a calibration space that captures all possible degrees of freedom for a stabilized spectrograph, there will be many options for pinpointing where the spectrograph is located within that calibration space. Good housekeeping data, such as temperature or pressure, could be used in addition to or instead of time (as mentioned in Section 5.1). Telemetry that is



seen to be correlated with the calibration state of the spectrograph can even be added to the data used to construct the low-dimensional basis. Furthermore, all exposures taken with the spectrograph in principle contain information about the calibration state of the spectrograph. Theoretically, tellurics, lines in science exposures, or just the trace positions themselves could also be used to determine the instrument's calibration state, thereby providing free simultaneous calibration information.

*Excalibur* is designed and optimized for EPRV spectrographs. In the battle to construct a high-fidelity data pipeline for EPRV measurements, we have shown that *excalibur* represents a step toward mitigating the error from wavelength calibration, as demonstrated by tests using EXPRES data (Section 4). Though the focus is on EPRV instruments here, *excalibur* should be largely applicable to nearly any other astronomical spectrograph.

We gratefully acknowledge the help of the anonymous referee, whose insightful and well-thought-out comments significantly improved the paper. We thank the EXPRES team for their work on the instrument, software, and pipeline. We thank Dan Foreman-Mackey for an illuminating discussion. The data presented here were obtained through the Lowell Discovery Telescope (LDT) at Lowell Observatory. Lowell is a private, nonprofit institution dedicated to astrophysical research and public appreciation of astronomy and operates the LDT in partnership with Boston University, the University of Maryland, the University of Toledo, Northern Arizona University, and Yale University. We gratefully acknowledge ongoing support for telescope time from Yale University, the Heising-Simons Foundation, and an anonymous donor in the Yale community. We especially thank the National Science Foundation (NSF) for funding that allowed for precise wavelength calibration and software pipelines through NSF ATI-1509436 and NSF AST-1616086 and for the construction of EXPRES through MRI-1429365. L.L.Z. gratefully acknowledges support from the NSF Graduate Research Fellowship under grant No. DGE1122492.

*Software:* SciPy library (Virtanen et al. 2020), NumPy (Oliphant 2006; van der Walt et al. 2011), Astropy (Astropy Collaboration et al. 2013; Price-Whelan et al. 2018).

## ORCID iDs

Lily L. Zhao  <https://orcid.org/0000-0002-3852-3590>  
David W. Hogg  <https://orcid.org/0000-0003-2866-9403>  
Megan Bedell  <https://orcid.org/0000-0001-9907-7742>  
Debra A. Fischer  <https://orcid.org/0000-0003-2221-0861>

## References

- Aminikhanghahi, S., & Cook, D. J. 2017, *Knowl. Inf. Syst.*, 51, 339
- Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, *A&A*, 558, A33
- Bauer, F. F., Zechmeister, M., & Reinert, A. 2015, *A&A*, 581, A117
- Blackman, R. T., Fischer, D. A., Jurgenson, C. A., et al. 2020, *AJ*, 159, 238
- Bonfils, X., Delfosse, X., Udry, S., et al. 2013, *A&A*, 549, A109
- Brewer, J. M., Fischer, D. A., Blackman, R. T., et al. 2020, *AJ*, 160, 67
- Brewer, J. M., Fischer, D. A., Valenti, J. A., & Piskunov, N. 2016, *ApJS*, 225, 32
- Butler, R. P., Marcy, G. W., Williams, E., et al. 1996, *PASP*, 108, 500
- Butler, R. P., Vogt, S. S., Laughlin, G., et al. 2017, *AJ*, 153, 208
- Cersullo, F., Coffinet, A., Chazelas, B., Lovis, C., & Pepe, F. 2019, *A&A*, 624, A122
- Coffinet, A., Lovis, C., Dumusque, X., & Pepe, F. 2019, *A&A*, 629, A27

- Fischer, D. A., Anglada-Escude, G., Arriagada, P., et al. 2016, *PASP*, **128**, 066001
- Foreman-Mackey, D., Montet, B. T., Hogg, D. W., et al. 2015, *ApJ*, **806**, 215
- Jolliffe, I., & Cadima, J. 2016, *RSPTA*, **374**, 20150202
- Jurgenson, C., Fischer, D., McCracken, T., et al. 2016, *Proc. SPIE*, **9908**, 99086T
- Kramer, M. A. 1991, *AChE*, **37**, 233
- Lovis, C., & Pepe, F. 2007, *A&A*, **468**, 1115
- Mayor, M., Marmier, M., Lovis, C., et al. 2011, arXiv:1109.2497
- Milaković, D., Pasquini, L., Webb, J. K., & Lo Curto, G. 2020, *MNRAS*, **493**, 3997
- Molaro, P., Esposito, M., Monai, S., et al. 2013, *A&A*, **560**, A61
- Oliphant, T. 2006, NumPy: A Guide to NumPy (Spanish Fork, UT: Trelgol Publishing)
- Pearson, K. 1901, LIII. On Lines and Planes of Closest Fit to Systems of Points in Space, Zenodo, doi:10.1080/14786440109462720
- Pepe, F., Cristiani, S., Rebolo, R., et al. 2013, *Msngr*, **153**, 6
- Petersburg, R. R., McCracken, T. M., Eggerman, D., et al. 2018, *ApJ*, **853**, 181
- Petersburg, R. R., Ong, J. M. J., Zhao, L. L., et al. 2020, *AJ*, **159**, 187
- Plavchan, P., Latham, D., Gaudi, S., et al. 2015, arXiv:1503.01770
- Price-Whelan, A. M., Sipőcz, B. M., Günther, H. M., et al. 2018, *AJ*, **156**, 123
- Probst, R. A., Lo Curto, G., Avila, G., et al. 2014, *Proc. SPIE*, **9147**, 91471C
- Suárez Mascareño, A., Faria, J. P., Figueira, P., et al. 2020, *A&A*, **639**, A77
- Tsalmantza, P., & Hogg, D. W. 2012, *ApJ*, **753**, 122
- van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, *CSE*, **13**, 22
- Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, *Nature Methods*, **17**, 261
- Wang, S. X., Wright, J. T., MacQueen, P., et al. 2020, *PASP*, **132**, 014503
- Wilken, T., Curto, G. L., Probst, R. A., et al. 2012, *Natur*, **485**, 611
- Woodbridge, Y., Elidan, G., & Wiesel, A. 2020, arXiv:2004.10255