We thank the referee for these incredibly helpful comments and suggestions. We particularly appreciate the way in which the comments were phrased, making for a very constructive report! We have adopted changes for each comment. Responses to the comments (orange) are given below while changes to the paper draft are highlighted in red and blue text.

## A few general comments:

-Overall the writing is mostly clear, although there are some typos throughout and there are spots where there is some repetition.

We have checked more carefully for typos and unnecessary repetition throughout the paper.

-I encourage the authors to consider sharing this code with the community, if they are not already.

Of course! We have added a sentence at the end of Section 2.3, where we finish explaining out implementation, with a footnote to the GitHub repository that hosts the code we developed.

-Much of the discussion in this work relates to wavelength calibration states that are accessible by small environmental perturbations, but I wonder if a significant strength of the non-parametric approach would be its ability to describe both

i) the (more typical) wavelength calibration changes due to small environmental changes and

ii) the wavelength calibration changes due to other kinds of changes (e.g. power failures, opening the spectrograph, etc) that might be rare and therefore more difficult to even try to develop a descriptive model for.

The difference between these is likely somewhat related to the discussion about selection of "eras of stability," and may be beyond the scope of this work. But the authors might consider emphasizing if they have seen evidence that their method facilitates the diagnosis or bridging of these eras of stability.

This is such a cool idea that we did not consider adequately in writing the paper! It is probably not possible with a PCA-based implementation, since the PCA will be overcome by the variations introduced by those larger changes and rendered insensitive to the smaller, calibration change. However, there could exist other methods capable of capturing both types of changes.

To expand on this idea, we have added the following sentence to the discussion of choosing a denoising model in Section 5

“It is also possible to move to non-linear reduction methods, like a Guassian process latent-variable model, an auto-encoder, a normalizing flow (Kramer et al. 1991, Woodbridge et al. 2020}. Using a non-linear denoising model could enable excalibur to capture large-scale changes as well as small variations in calibration state.

the following sentence in Section 6 after introducing eras of stability,

“Different denoising models will be able to account for different amounts of stability or lack thereof.”

and the following sentence in the discussion.

“Other 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.”

## Specific comments:

*Section 1: "The state of the art RV precisions reached 1 m/s in 2016." This is very precise - is there a citation that is being referred to here?*

We have added a citation to Fischer et al. 2016.

Section 1: "This is particularly true of EPRV instruments, which are equipped with stringent environmental stabilizing, likely reducing the number and extent of the axes along which the instrument can vary." More fundamentally, isn't the dimensionality of "calibration space" more about the finite number and (white-pupil) arrangement of elements of the optical systems? And the "extent" of these axes more about the thermomechanical stability?

Yes, except we haven’t done the legwork to prove how the axes of the calibration space correlate with instrument hardware/optics. However, the results presented in this paper do show that the variations stabilized instruments (or at least EXPRES) experiences can be represented by a low-dimensional framework.

To clarify this point, we edited the sentence to read:

“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”

Section 2: PCA - perhaps a more appropriate citation (than Wikipedia) would be Pearson 1901, or a statistics text that describes the method.

Fair point. We have added a citation to Pearson 1901 and a recent 2016 review by Jolliffe et al.

Section 2.1: "There are many strategies for identifying calibration line positions and matching them to their assigned wavelengths... excalibur assumes that this has been done correctly." - The authors might consider if assuming the "correct" mapping may involve underlying assumptions that would be relevant to users of excalibur. I appreciate that the nuances of this question are out of scope for this work, but I wonder if a bit more detail here might be helpful, particularly in view of the later discussion which does treat the fiber-to-fiber differences as well as applications to stellar spectra. (Examples that come to mind: is there a clear choice of line position in the case where the spectrograph response is asymmetric, or where CTI might cause emission and absorption lines to bias differently?) It may be sufficient to simply point the reader to the later discussion about these issues in section 7.

We’ve added a more explicit discussion of “correctness” and a reference to section 7. The paragraph now reads:

“Excalibur assumes that line positions have been identified “correctly,” as in the position of a calibration line is determined 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 an effect that the science exposures are not, for example differences in charge transfer inefficiency, non-linearities in the PSF, etc. We caution that systemic errors or large uncertainties in fitting line positions easily propagate through to biases in the wavelength models returned by excalibur. For more discussion, see Section \ref{sec:discussion}.”

Algorithm 1 - The text description is clear, but the algorithm is the only place the matrices are used (U,Sigma,V). Without more information, I'm not sure I appreciate the details of the method being described. From notation, I suppose it is describing SVD? This should be clarified and/or streamlined.

We’ve added text to clarify we mean SVD in the algorithm box.

Section 2.2/2.3 and Algorithm 2 - What does the prime signify? Please check that it is used consistently.

We have added text to the algorithm to clarify the meaning of the prime. We checked sections 2.2/2.3 to ensure that the prime accompanies all values that change for an exposure that excalibur is constructing wavelengths for (for example, the wavelengths and orders of the lines will stay the same regardless of exposure, and so do not get primes).

Section 3: "In order to satisfy the assumption that there exists only low-order variation, which is needed for excalibur, we used exposures from after the photonic crystal fiber within the LFC was changed in Summer 2019, and the LFC stabilized." - Is this motivating the choice to avoid spanning this transition, or to only use the period afterwards? My rough understanding is that the PCF change could alter the overall envelope of the comb, but I don't see why the PCF would induce different higher-order variations in the line centroids.

Yes, actually most of the lines did stay in the same place! We stick to exposures after this transition out of both an abundance of caution and because when we replaced the PCF, we also changed the wavelength range of the LFC. We have added the following text to the paper that expands on this point:

“we used exposures from after the LFC stabilized following servicing in summer 2019, where the photonic crystal fiber was replaced and the polarization was changed to shift the wavelength range of the LFC redwards”

Section 4: If I am understanding correctly, the polynomial and PCHIP tests involve censoring 50% of the data (even/odd) while the denoising test involves censoring only 10% (randomly). I am therefore unsure whether the distributions shown in Fig 2 are comparing "apples to apples," although I suspect that defining the RMS per line does the trick. The authors should consider motivating and explaining this difference more clearly, as well as any potential implications in how one might understand the spreads compared in Fig 2.

We highlight this difference and use it to preface our decision to evaluate the results on a line-by-line basis by including the following paragraph:

“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 are capturing 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 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.”

Fig 2: It would be clearer if the legend referred to a "Denoising" test rather than "Excalibur" for the black curve, as the text describes.

The caption has been changed to read “Denoising” instead of “Excalibur”

Section 4: It might be helpful if the authors could clearly explain the origin of the difference in the PCHIP and "Excalibur/denoising" distributions.

We added the following paragraph discussing this difference:

“The per-line residuals from the denoising test also exhibit smaller spread than interpolation alone. This suggests that the spectrograph truly is accurately represented by a low-dimensional model, and 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.”

Fig 4: Right and left are switched. Also, the residuals being given in m/s (instead of or in addition to Angstroms) would enable direct comparison with earlier figures.

Right and left have been switched in the caption. The y-axis of the right plot has been rescaled to be in units of m/s.

Fig 5: Although it is good to see the time series bins down for the excalibur-calibrated RVs, it would be much more helpful to see a direct comparison to the polynomial-calibrated RVs. I strongly suggest adjusting the plot accordingly.

We have changed the plot to show polynomial-calibrated RVs in a left plot and the excalibur-calibrated RVs in a right plot.

Section 4.1: The magnitude of the improvement measured here really is substantial (20-50 cm/s in RSS)! The authors should consider if there is any additional information that might strengthen this evidence of the performance of excalibur, such as how many data sets were tested, what the distribution of improvements was, the properties of the stars, etc. An additional plot may be justifiable.

The other data sets proved less clean that the one we currently show, HD 34411. HD 34411 is a quiet star with no known planets, so a reduction in RMS is readily interpretable. Other systems did show a similar reduction, but have planets or are more active and so have a higher base RMS level. We tried to make this more transparent by removing the highlighted statement and replacing it with the following paragraph:

“We conducted a direct test of a classically-generated wavelength solution with excalibur-wavelengths on four other data sets. All targets showed a reduction in or comparable RV RMS. The results from these data sets cannot be interpreted as directly as with HD 34411, though, due to larger contributions from stellar variability, known planets, etc. As completely mitigating these different effects is out of scope for this paper, we focus here on the results with HD 34411.”

Section 5.2: "It is unnecessary to take calibration images at times where the same information can be constructed by a well-chosen interpolation scheme." - But is this not dependent on the SNR and desired precision?

Yes, we added a clause within the sentence to specify the dependence on precision. The sentence now reads:

“It is unnecessary to take calibration images at times where the same information can be *reconstructed at the desired precision* by a well-chosen interpolation scheme.”

Section 5.2: "We also tested an implementation of excalibur... this method returns comparable RV RMS for most targets, though appears to do better when a night has sparse calibration data." - Which two methods are being compared here? Is the cubic interpolation justified over the linear? This test, its motivation and conclusion, could be explained more clearly.

We have expanded the paragraph to delineate between the two methods more explicitly and added a sentence explaining the conclusion of the test. The paragraph now reads:

“We also tested an implementation of excalibur where the K principal components within a night were fit to a cubic *with respect to time rather than 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 exposure by fitting polynomial coefficients with respect to time to a cubic. We found that ~~this method~~*using a cubic in place of linear interpolation* returns comparable RV RMS for most targets, though 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)*.”

Figure 9: See earlier comment about Angstroms vs m/s. Also, the caption refers to a different y-axis scaling on the left and right sides, but they appear to be the same.

The y-axis for both plots has been changed to m/s. The sentence about the different y-axes was removed since it is wrong.

Section 5.3: "Due to the dispersion intrinsic to echelle spectrographs, the wavelength change between pixels grows greater with greater wavelengths." Although this is true across spectral orders, within an order I believe it is not. Please check this statement and that this paragraph is self-consistent.

This does appear to be the case within orders, though on a much smaller amplitude than across orders. We’ve seen this in all wavelength solutions. We have double checked that the paragraph is self-consistent.

Section 5.3: "A more classic cubic spline interpolation can run into issues with arc lines, which are irregularly spaced or even blended..." The authors should consider whether a comparison between "classic" cubic spline and PCHIP with the LFC lines (rather than ThAr, which this implementation of excalibur does not focus on) would more directly motivate their choice here. I believe this is mostly already shown in Figure 4.

The issue being described in that paragraph doesn’t really arise when using LFC lines (which is partly why they’re so great!). We have added text to the paragraph to clearly specify that this is only a concern with ThAr lines and that is what the figure is showing.

Fig 10: Perhaps the authors could explain the difference in magnitude of the per-line RMS shown in Figure 10 vs Figure 2? I am struggling to appreciate why they are so different.

We have clarified the line density tests used to make Figure 10 with explicit references to the test results shown in Figure 2. The descriptive paragraph now reads:

“As an empirical assessment of the needed line density, we removed LFC lines from the EXPRES data and calculated 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 are also implementing denoising unlike the pure interpolation test of Section 4*.”

Section 6: "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..." What if the fiber-to-fiber drift is independent of the "bulk" spectrograph drift?

The efficacy of this proposed method definitely does depend on the inherent assumption that the simultaneous fiber traces the same changes experienced by the science fiber. We’ve added the following sentence highlighting this:

“This method, as with all 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.”

Section 6: "If the FSR varies very little, its value can in principle be captured by a PCA or similar method..." - Is a "little" or a "lot" referring to the magnitude or complexity of the temporal behavior (or both)?

On further reflection, we have instead become convinced that incorporating the FSR as part of the PCA is impossible. We have deleted that paragraph and instead added a paragraph describing how we would incorporate the FSR as a free parameter and expand the excalibur model for etalons. The new paragraph reads:

“Incorporating the FSR as part of the excalibur model will require introducing a free parameter to capture changes in the FSR independent of variation in an instrument's calibration state. The calibration state can then be described with respect to mode number, which will be used to uniquely identify a calibration line across exposures rather than 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 non-linear. This model would require a simultaneous reference or other housekeeping data that can be used to determine the FSR for every exposure.”

Section 7: "Theoretically, tellurics, lines in science exposures, or just the trace positions themselves could also be used to determine an instrument's calibration state, negating the need for any additional wavelength-calibration source beyond defining the calibration space." - I appreciate the value of this additional (potentially under-utilized) calibration information. But to assert this could obviate the need for any calibration source, at least for PRV work, is perhaps going a bit too far.

We have controlled our enthusiasm by changing the ending of the sentence to read:

“Theoretically, tellurics, lines in science exposures, or just the trace positions themselves could also be used to determine an instrument's calibration state, *thereby providing free simultaneous calibration information*.”