Peer Review File

Manuscript Title: Light Bending and X-ray Echoes from Behind a Supermassive Black Hole

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

Wilkins et al
This is a very sophisticated analysis of a very high quality data set which, if correct, is a discovery of a long predicted phenomenon the detection of light bent from the 'backside' if the black hole (similar to the EHT result in some sense) . However the result is highly dependent on the interpretation of the time resolved energy resolved data and while the authors have presented a detailed analysis, the main supporting figures (2 and 7) do not, to this referee, support the result. In addition a fair amount of the description of the analysis is confusing (at least to this reader) and needs to be re-written.

Style- Its not completely clear to this referee if the physical modeling of this source requires the conclusions of this paper or is simply one of several possibilities. For example the 'need' to have the x-ray source move at high velocities during the flare or the interpretation of the time variability as the combination of GR effects and disk reprocessing. Given the complexity of the time variability, which the authors interpret (consistent with previous papers) as combing from a two component coronae, I find it hard to believe that the authors model is the only possible interpretation.

Content:

In the abstract it says "During the flare, the corona is accelerated upward,
away from the disk and its temperature drops"- I think it would be much better to say
" The data are consistent with a model in which during the flare, the corona is accelerated upward, away from the disk and its temperature drops"-
in parallel with the next set of sentences.

line 45- please refer to a figure2 which shows the flares
line 58- please refer to where in the supplementary material the value and its uncertainty are derived and the model used.
Line 61- which model is used to fit for the reflection fraction?

Much of the conclusions depend on 'uniqueness'- are there other models which can account for the lower than expected reflection fraction?- It is very common to have 'low' reflection fractions . I remember from ancient times that there were ideas that the disk either developed a 'hole' (truncated disk) or became so highly ionized that the fluorescent Fe K line was reduced (while the reflection continuum remained). See Xu et al 2020ApJ...893...42 or Zdziarski et al 2020arXiv200612829Z Are these models ruled out by the present data??

In the text the values for R(in) are introduced 'ad hoc' there should be references to the additional material which justify the values used.

I find the sentence "The ejected corona can escape the central region and play a role in the evolution of the host galaxy" confusing- While their fits find a velocity of 0.7c no errors are given. I assume that these values are for a certain spin since the escape speed depends on the BH spin. Also its not at all clear what the authors mean by "play a role in the evolution of the host galaxy"- I assume that they are referring to feedback, but do they mean that the x-ray emitting plasma somehow is escaping the central regions and 'causing' feedback or something else.

Looking at Gonzalez et al (ref 20) which is a very detailed description of the physics used in the present paper it is claimed that one can distinguish extended vs point like coronae- what is the situation with IZwI?? This is mentioned in the supplementary material with the sentence lines 121-126 "The simultaneous observation of an extended flat portion of the emissivity profile with a steep inner emissivity profile cannot be explained by a compact, point-like corona…." But no values are given

Is the rapid change in effective temperature of the disk a new result? In particular it is rather rare to have such low values of kT- what model is used to derive the temperature?
Please clarify the spectral changes (lines 80-84) are you are describing changes in the temperature while the slope remains constant before and during the flare but flattens afterwords?- In a standard Comptonization model the powerlaw index is directly connected to the optical depth and temperature and it is very constraining to have large changes in temperature and a constant slope. (line 78-84)

Figure 1 and 2 have different zero points for the time axis making comparison difficult.

Figure 2 and figure 8 are the crucial figures identifying the initial flare (red) and the response (blue) however I have great difficulty seeing the initial flare in any energy bin but the 6-8 kev bin. In particular (lines 290-293 in additional material) says "We apply this model of X-ray reverberation from a single flare simultaneously to the light curves in the 3-5keV and 6-8keV bands, in which the narrow, offset peaks were most significantly detected.- yet looking at figures 8 the first 'flare' indicated with the dashed line is only at 1-2σ significance. This is the major problem I have with the analysis. The authors give a sophisticated treatment of the detection of the flares and I have no reason to doubt the result, but to the eye, especially in the lower energy bins the flares certainly do not seem significant.

I am not sure what can be done to improve the quantitative analysis of the light curve identifying the flare times but this is rather worrying not to be able to 'see' the main result of the paper. As we all know creating a light curve from the power spectrum produces a wide variety of light curves and the authors have done this 106 times- which should certainly be enough- perhaps they could give a bit more detail on how they have done this (Timmer, J., & Koenig, M.??). A result this interesting needs to be very robust.

My eye sees a 'flare' in the 1-3 kev band at 2200 sec and in the 5-7 kev at 1700 sec which are not discussed. It also seems a bit odd to me that the 'response' (blue curve) is stronger (particularly in the 3-5 kev bin) than the input- is this due to gravitational lensing ? (I do not understand the modeling well enough to be definitive about this) .

The C statistic does not provide a goodness of fit- only chi-sq does. To quote from the Sherpa manual "The magnitude of the cash statistic depends upon the number of bins included in the fit and the values of the data themselves. Hence one cannot analytically assign a goodness-of-fit measure to a given value of the cash statistic. Such a measure can, in principle, be computed by performing Monte Carlo simulations."

11) What is the time binning shown in figure 8
12) Lines 292-296- Since the 5-7 keV bin is not independent of the 3-5 and 6-8 keV bin why have the authors presented it since they do not use it.

Given the energy resolution of the CCD they could have had 3-5.7 and 5.7-8 keV to increase the signal to noise.
13) The discussion in lines 300-305 is confusing – first they assume that the spectrum is time invariant and on line 306 say that it is not (consistent with the spectral analysis)
14) in Lines 308-316 they add the reverberation model and say that it reduced the C statistic by 24 they should state the statisticial signifgance of this given the number of parameters that have been added. However if I do my math correctly the fit is still not very good with C=311 (remembering that the absolute value of C does not indicate the goodness of fit- I would recommend that the authors either do a Baysian analysis to get the goodness of fit or a chi-sq analysis).
15) Line 320-333 indicates that they are introducing 2 more free parameters the height of the x-ray sources and the mass of the black hole- what is the improvement of the fit with these additional parameters to compare with C=311??

References:
Am I correct in saying that the theory behind this result have been submitted for publication as arxiv 2008.08083- this is not referenced
Need a reference for the 'complex' class of NLSYIs (presumably Gallo 2006) this nomenclature is not 'standard' (line 29)- the authors use the phrase "weaker in their X-ray emission compared to the optical and UV emission," perhaps they could quantify this using the α(ox) parameter?

Grammar
The sentence "This illumination of the inner disk by the coronal x-
ray continuum, the X-rays that are seen reflected from the disk [4], and the time delays in the reverberation of continuum variations off the disk [5, 6] provide a unique view of the extreme environment in the strong gravitational field just outside the event horizon." Is not clear
Do the authors mean
"The inner disk is illuminated by the coronal x-ray continuum and thus the X-rays that are seen reflected from the disk [4], and the time delays in the reverberation of continuum variations off the disk [5, 6] provide a unique view of the extreme
environment in the strong gravitational

Referee #2 (Remarks to the Author):

This paper reports new XMM-Newton and NuSTAR observations of the active galactic nucleus (AGN) I Zw 1. Confirming previous studies, these data find spectral evidence for the reflection of X-rays from the inner regions of an accretion disk around a spinning black hole. The authors report a flare during the course of the observation and report a decrease in the reflection fraction during the flare, a result that is interpreted as the relativistic ejection of the X-ray corona during the flare. The most novel (and headline) result is the reported detection of structure in the energy-resolved light curves that is interpreted as iron line reverberation from the accretion disk on the far side of the black hole. This would be the first detection of this particular prediction of the theory of reverberation in strong-field gravity.

The presentation makes a strong case for the existence of relativistic reflection signatures in this AGN. Regrettably, as I explain below, I do find the other claims to be far from compelling. Thus, I cannot recommend the publication of this paper in its current form.

Let me start with the headline result - the existence of a large flare after which there is the detection of this particularly interesting reverberation signature. I have stared at Figure 2 for a long time and have failed to convince myself that there are distinct structures in the light curves beyond what one might expect from the normal aperiodic variability. It is well known that the continuum variability in hard and soft bands is not 100% coherent, especially for high-frequencies (short timescales), and so the existence of subtle differences in the short term variability between the soft and hard bands is not by itself unusual. Further, there clearly are longer term (and hence not reverberation) induced changes in the blue peak of the iron line through the flare as shown in Figure 7 of the SI. Ultimately the entire claim rests upon the apparent shift in one peak, and it would be far from obvious that this peak was special if it were not for the superposed model on Figure 2. If one were to scan through the entire light curve, what is the chance that unrelated high-frequency variability would produce such a result? When assessing the statistical significance in the SI, the authors instead focus on a different and much more mundane question, i.e. the probability that the flares exist at all (relative to the null hypothesis of a constant flux). That is surely not the correct or relevant question to ask when assessing the significance of this possible reverberation signature.

Bold claims need robust evidence to back them up, and in my view the evidence presented in the current manuscript falls short. We must recognize that the presence of this reverberation signature is *not* expected to naturally arise in this system, somehow requiring the disk to see an intense ~100s flash when we observers actually see a much more extended 10,000s flare. This issue is briefly mentioned in the manuscript but dismissed with some creative story-telling.

In an effort to bolster the evidence, I can make two suggestions. Firstly, the authors should assess whether there are any other periods of time (either from this or previous datasets) for which the same time-delayed redward-moving peaks are seen. The significance of this and any other putative moving peaks should be assessed using simulated lightcurves that includes the frequency dependent changes in coherence (i.e. reduced coherence on short timescales). Secondly, the authors can attempt to make an impulse response function by plotting these data on the energy-time space (essentially doing with real data what Young & Reynolds [2000] did with simulated data). This may uncover other structure associated with the reverberation that supports the claim. If neither of this approaches yield additional evidence, I urge the authors to withdraw their claim and seek additional (new) data.

I also need to comment on the other principal result discussed, namely the suggestion that the corona has been set into mildly relativistic outflow as a means to explain the reduction in reflection fraction during the flare. Firstly, additional statistical evidence needs to be presented to demonstrate that the reflection fraction does actually drop - the 90% confidence regions for the pre-flare and flare reflection fraction do overlap and so there is a question mark concerning the statistical significance of this result. Secondly, the authors found that the photon-index of the continuum is unchanged before and during the flare. This is not what's expected if the corona is set into relativistic outflow - the photon starvation experienced by the moving corona would naturally produce a harder spectrum (Beloborodov 1999) by Delta-Gamma~0.2-0.3. Thus, an important prediction of the accelerated corona model is not seen.

Author Rebuttals to Initial Comments:

We thank the referees for their time and feedback, which has been extremely helpful in improving this manuscript. To address the concerns of the referees, we have completed a number of further analyses of both the present data and archival data from AGN, and we have reorganized the manuscript to improve the presentation and discussion of the results.

Responses to the specific queries of the referees are below and new sections of the manuscript have been highlighted.

We would like to thank the Editor and the referees in advance for their consideration of the revised manuscript.

Referee 2:

Let me start with the headline result - the existence of a large flare after which there is the detection of this particularly interesting reverberation signature. I have stared at Figure 2 for a long time and have failed to convince myself that there are distinct structures in the light curves beyond what one might expect from the normal aperiodic variability. It is well known that the continuum variability in hard and soft bands is not 100%coherent, especially for high-frequencies (short timescales), and so the existence of subtle differences in the short term variability between the soft and hard bands is not by itself unusual. Further, there clearly are longerterm (and hence not reverberation) induced changes in the blue peak of the iron line through the flare as shown in Figure 7 of the SI. Ultimately the entire claim rests upon the apparent shift in one peak, and it would be far from obvious that this peak was special if it were not for the superposed model on Figure

2. If one were to scan through the entire light curve, what is the chance that unrelated high-frequency variability would produce such a result? When assessing the statistical significance in the SI, the authors instead focus on a different and much more mundane question, i.e. the probability that the flares exist at all (relative to the null hypothesis of a constant flux). That is surely not the correct or relevant question to ask when assessing the significance of this possible reverberation signature.

Bold claims need robust evidence to back them up, and in my view the evidence presented in the current manuscript falls short. We must recognize that the presence of this reverberation signature is *not* expectedto naturally arise in this system, somehow requiring the disk to see an intense ~100s flash when we observersactually see a much more extended 10,000s flare. This issue is briefly mentioned in the manuscript but dismissed with some creative story-telling.

In an effort to bolster the evidence, I can make two suggestions. Firstly, the authors should assess whether there are any other periods of time (either from this or previous datasets) for which the same time-delayed redward-moving peaks are seen. The significance of this and any other putative moving peaks should be assessed using simulated lightcurves that includes the frequency dependent changes in coherence (i.e. reduced coherence on short timescales). Secondly, the authors can attempt to make an impulse response function by plotting these data on the energy-time space (essentially doing with real data what Young & Reynolds [2000] did with simulated data). This may uncover other structure associated with the reverberationthat supports the claim. If neither of this approaches yield additional evidence, I urge the authors to withdraw their claim and seek additional (new) data.

In order to address these concerns, we have completed a number of further stages of the analysis.

Firstly, we have improved the presentation of the section in which the significance of the flares is discussed. In particular, we have recreated Figure 10, into a format that better shows the significance at which the peaks aredetected. The significance is defined in terms of the probability of the peaks being caused by Poisson noise fluctuations from the underlying smoothed time series. We find that the detection of the peaks is significant with respect to this hypothesis.

It is true that the coherence between light curves drops at high frequencies, however simulating random light curves for the count rates observed in I Zw 1, we find that the measured drop in coherence is consistent with that expected from Poisson noise (there is no evidence for an additional drop in these data). We show that the peaks are detected significantly with respect to the null hypothesis of Poisson noise and, hence, they are not consistent with the null hypothesis that they are seen due to the same drop in coherence that is seen across thefull light curves.

We have extended the analysis to the second flare, and we find, in Fig 10 (the peak detection plot) that there isa set of similar offset peaks that occurs the same time into the second flare (though these peaks are weaker anddetected less significantly). We then proceed to stack (sum) the light curves across the two flares, aligning them in time based upon the time of these detected peaks. Upon doing this, we find that the rising profiles of the flares line up. This means that these peaks of interest occur the same time into both flares. We therefore now conduct the final modelling of the peaks and reverberation response on the summed light curve to improve theconstraints.

To determine whether a reverberation response is expected in the data, we have separately analyzed the average reverberation time lag, following the method published by Uttley et al. 2014. The 2020 observations were not long enough to provide a high signal-to-noise measurement alone, but we combined them with the archival observations to obtain an average measurement of the height of the corona above the disk. We find that the measured lag time in the iron K lag is consistent with a corona at a height of 4 gravitational radii, which is consistent with the height found independently by fitting the timing of the peaks during the flare with the reverberation model. Thus, for a flare from the corona at the average height that is detected, we expect to see narrow peaks when the X-ray emission re-emerges from behind the black hole at the measured times.

To further investigate the reverberation interpretation for the peaks, as suggested by Referee 2, we a first attempts to directly recover the response function from the (noise) data, by calculating the residuals of the spectrum in each time bin to the best-fitting power law. The procedure is detailed in the letter and the Supplementary Material. In the residuals (Fig. 5a), we see the expected features from the response function matching the timing of the detected peaks. We see the 'loop' and evidence for the 'waterfall' tail of the model response function and the loop spans the expected duration in time and energy. Finally, we use this to plot theresidual spectrum at different times (Fig. 5b). We see the expected shifting in the centroid of the iron line as different parts of the accretion disk respond to the flare.

Combining all of these factors, we find significant evidence in favor of the reverberation interpretation of thesepeaks and their interpretation as the detection of the re-emergence of reflected X-rays from behind the black hole. We have added a discussion of the uniqueness of these findings and find that no other common modes ofvariability seen in AGN can readily explain the observations.

We have clarified the point about the duration of the peaks with respect to the length of the total flare, which was not well articulated in the original version of the manuscript. We find that the appearance of the peaks is consistent with the reverberation response to the start of the flare, however we note as a caveat to this interpretation that the model requires that the peaks are only seen in response to the first 2000s of the flare, otherwise the narrow peaks would become smeared out at late times. We note that this is plausible in light of the measurement of the drop in reflection fraction during the flare. This means that as the flares progress, the disk is less strongly illuminated than when they start, thus the strongest reverberation response is only expectedfrom the start of the flares, consistent with what is seen.

> I also need to comment on the other principal result discussed, namely the suggestion that the corona has been set into mildly relativistic outflow as a means to explain the reduction in reflection fraction during the flare. Firstly, additional statistical evidence needs to be presented to demonstrate that the reflection fractiondoes actually drop - the 90% confidence regions for the pre-flare and flare reflection fraction do overlap and so there is a question mark concerning the statistical significance of this result. Secondly, the authors found that the photon-index of the continuum is unchanged before and during the flare. This is not what's expectedif the corona is set into relativistic outflow - the photon starvation experienced by the moving corona would naturally produce a harder spectrum (Beloborodov 1999) by Delta-Gamma~0.2-0.3. Thus, an important prediction of the accelerated corona model is not seen.

We have clarified this point in the revised version of the manuscript. It is not that evidence was found that the photon index remains constant, rather we find a lack of significant evidence for a variation in photon index. Given the larger error bar during the flare, the data are consistent with the null hypothesis of constant photon index, though the data also show a hint that the photon index hardens.

We have revised the estimates of the photon index and reflection fraction during the flare, reducing the uncertainty on these quantities. We find that due to the low level of absorption at soft X-ray energies during theflare, we are able to use information from the soft X-ray spectrum to tighten these constraints – namely the level of soft X-ray flux seen places upper limits on the reflected flux and the softness of the continuum spectrum. This is now discussed in the section of the Supplementary Material about the variability in the X-ray spectrum.

It should be noted that the model of Beloborodov et al. 1999 incorporates a number of simplifying assumptions,namely that the temperature of the corona remains constant. There are, however, a number of uncertainties inhow the corona is energized and how the X-ay continuum emission is produced. Beloborodov et al. 1999 assume the continuum arises through Comptonization within an idealized "blob" corona. In light of the evidence we find that the temperature of the corona drops during the flare, we expect there to be a competing effect that softensthe photon index, which counteracts the expectation that the motion of the corona causes it to harden. We haveincluded a discussion of these factors in the Supplementary Material.

Referee 1:

Much of the conclusions depend on ‘uniqueness’- are there other models which can account for the lower than expected reflection fraction?- It is very common to have ‘low’ reflection fractions . I remember from ancient times that there were ideas that the disk either developed a ‘hole’ (truncated disk) or became so highly ionized that the fluorescent Fe K line was reduced (while the reflection continuum remained). See Xu et al 2020ApJ...893...42 or Zdziarski et al 2020arXiv200612829Z Are these models ruled out by the present data??

We have added an analysis of inner radius of the accretion disk to our analysis of the variability in the X-ray spectrum. The inner radius of the disk is constrained by the extent of the redshifted wing of the broad iron line.We find that during the flares, when the reflection fraction is lowest, the inner radius is most tightly constrained to the lowest values, thus the drop in reflection fraction is inconsistent with these findings.

It is true that over-ionization of the inner disk can reduce the strength of the fluorescent iron K line and cause the disk to appear truncated. This is also inconsistent with the above finding. The reflection fraction, however, is defined in terms of the total reflected flux, so is also robust to over-ionization of the inner disk by the detection of the Compton hump.

We have included a discussion of these factors as well as some estimates of how truncated a disk would need to be in order to produce the low reflection fractions that are seen.

In the text the values for R(in) are introduced ‘ad hoc’ there should be references to the additional material which justify the values used.

The inner radius of the accretion disk was fit as a free parameter to the observed X-ray spectrum. Assuming theaccretion disk extends down to the innermost stable orbit, this inner radius is parametrized in terms of the black hole spin parameter.

I find the sentence “The ejected corona can escape the central region and play a role in the evolution of the host galaxy” confusing- While their fits find a velocity of 0.7c no errors are given. I assume that these values are for a certain spin since the escape speed depends on the BH spin. Also its not at all clear what the authors mean by “play a role in the evolution of the host galaxy”- I assume that they are referring to feedback, but do they mean that the x-ray emitting plasma somehow is escaping the central regions and ‘causing’ feedback orsomething else

We have removed this statement as it is periphery to the main focus of the paper. We have further added uncertainties on the velocity estimates derived from the reflection fractions

Looking at Gonzalez et al (ref 20) which is a very detailed description of the physics used in the present paperit is claimed that one can distinguish extended vs point like coronae- what is the situation with IZwI?? This is mentioned in the supplementary material with the sentence lines 121-126 “The simultaneous observation ofan extended flat portion of the emissivity profile with a steep inner emissivity profile cannot be explained by a compact, point-like corona….” But no values are given

These findings were first presented in Wilkins & Fabian 2012. The flat portion in the middle of the emissivity profile is produced either by the extent of the corona over the disk or by the height of the corona above the disk. If the corona is high above the disk though, while the flat portion extends further, the inner part of the profile becomes less steep, thus to have both an extended flat region of the profile and a steep inner profile, it is necessary for the corona to be extended.

We note, however, that from the findings of Wilkins et al. 2016, the extended part of the corona varies slowly and when considering the rapid variability on reverberation timescales, we are only seeing the response to the core of the corona, thus for the reverberation model used to model the peaks, we can simplify the corona as just its core (equivalent to a point source).

These points have been clarified in the text and the values for the extended corona are derived from fitting themodel of a reflection spectrum with a broken power law emissivity profile to the spectrum. The extent of the corona (and uncertainty) is given by the “outer break radius” parameter in Table 1.

Is the rapid change in effective temperature of the disk a new result? In particular it is rather rare to have suchlow values of kT- what model is used to derive the temperature? Please clarify the spectral changes (lines 80-84) are you are describing changes in the temperature while the slope remains constant before and during the flare but flattens afterwords?- In a standard Comptonization model the powerlaw index is directly connected to the optical depth and temperature and it is very constraining to have large changes in temperature and a constant slope. (line 78-84)

The temperature of the corona is measured using the RELXILL model of the broadband continuum spectrum andthe reflection from the disk. The temperature of the corona controls the high-energy cut-off in the continuum. The continuum transitions from a power law to an approximately exponential cut-off at an energy that corresponds to the temperature. Furthermore, as discussed by Garcia et al. 2015, changing the temperature ofthe corona changes the energy balance in the reflecting accretion disk, thus it changes the relative strength of the iron K line and Compton hump as measured by NuSTAR. We simultaneously fit the spectrum from before the flares, during and after. We know that the iron abundance cannot change between these time periods. Wetherefore find that even after marginalizing over uncertainties in the ionization parameter, modelling the variation seen in the ratio of the iron K line to the Compton hump in the reflection spectrum requires a change in coronal temperature. We have added an explanation of this.

Figure 1 and 2 have different zero points for the time axis making comparison difficult

All of the light curves have now been recreated with a common zero-point corresponding t the start of the flares,except in Fig 1 (the full observation). The captions to later figures now explicitly state the start time relative to Fig 1.

My eye sees a ‘flare’ in the 1-3 kev band at 2200 sec and in the 5-7 kev at 1700 sec which are not discussed. Italso seems a bit odd to me that the ‘response’ (blue curve) is stronger (particularly in the 3-5 kev bin) than theinput- is this due to gravitational lensing ? (I do not understand the modeling well enough to be definitive about this) .

We have added a clearer discussion of the other fares or peaks that appear. The most notable of these are the ones that appear at the maximum point of the flares, and the one that appears at the beginning of the second flare. These appear simultaneously across all of the energy bands (except the one at the beginning of the secondflare is only seen in the lower energy bands, it is too soft to be significantly seen int eh 6-8 keV band).

The fact that these peaks appear simultaneously across the energy bands means that they can readily be described by a broad emission component, not a line arising from different parts of the disk. They are readily understood as rapid changes to the continuum emission. They appear in Fig. 10 because the smoothing appliedto the light curve to characterize the peaks does not fully account for the sharpness of the top of the flares, thusthere is excess continuum variability seen here where the continuum changes vary rapidly right at the peak of the flares.

The C statistic does not provide a goodness of fit- only chi-sq does. To quote from the Sherpa manual “The magnitude of the cash statistic depends upon the number of bins included in the fit and the values of the datathemselves. Hence one cannot analytically assign a goodness-of-fit measure to a given value of the cash statistic. Such a measure can, in principle, be computed by performing Monte Carlo simulations.”

A note has been added to the Spectral Analysis section to clarify that we use the modified C-statistic defined inXSPEC to approximate chi-squared when there are at least 5 counts per bin (applicable here). We now only state the C-statistic in supporting the choice of spectral model. We no longer quote the change in C-statistic when modelling the timing of the peaks, since, in retrospect, this is not a meaningful comparison.

What is the time binning shown in figure 8

The time binning has now been added to the light curve captions.

Lines 292-296- Since the 5-7 keV bin is not independent of the 3-5 and 6-8 keV bin why have the authors presented it since they do not use it. Given the energy resolution of the CCD they could have had 3-5.7 and 5.7-8 keV to increase the signal to noise.

The limiting factor in defining the light curve energy bands is not the energy resolution of the CCD, rather it is the number of counts per time bin. We also deliberately keep 3-5 and 6-8 keV separate to explicitly separate outthe red- and blue-shifted sides of the iron K line, while excluding the core of the line from the outer disk from these bands. We feel, however, that it is instructive to show the 5-7 keV and to illustrate the behavior of the response function and how the peak in this band lies at a time interim to that in the 3-5 and 6-8 keV bands.

> The discussion in lines 300-305 is confusing – first they assume that the spectrum is time invariant and on line306 say that it is not (consistent with the spectral analysis)

We have now clarified this point. We begin working from an assumption that the spectrum is invariant on shorttimescales (between light curve time bins), though we know it changes on longer timescales (from before to after the flare). The residuals we find to this model are the evidence for the spectral variability, namely the reverberation response in time after the flare.

> Am I correct in saying that the theory behind this result have been submitted for publication as arxiv 2008.08083 - this is not referenced.
>
> Need a reference for the 'complex' class of NLSYIs (presumably Gallo 2006) this nomenclature is not 'standard'(line 29)- the authors use the phrase "weaker in their X-ray emission compared to the optical and UV emission," perhaps they could quantify this using the α(ox) parameter?

The theory behind the results, and the corresponding method for computing the model reverberation responsefunctions, were published in Wilkins et al. 2016. arXiv 2008.08083 discusses a theoretical extension to this modelto account for relativistic effects around the black hole that return the reflected emission to the disk, but that extension is not directly relevant to this work.

A reference to Gallo 2006 has been added for the complex class of NLS1s (reference [9]).

Reviewer Reports on the First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

Review of Wilkins et al "Light Bending and X-ray Echoes from 2 Behind a Supermassive Black Hole"

The authors have done a tremendous amount of work and constructed a very sophisticated model and data analysis method, far beyond what is usual in this field. As such this paper sets a new standard. They have answered most of the points that I have raised in the original review except one- the uniqueness of their model. The result that they claim does not 'leap out' of the data (the light curves at a simple level look very similar to lots of other AGN x-ray light curves) and requires the level of analysis that they have performed to reveal and thus whether one accepts their result or not depends on understanding what they have done. The authors have in a very lengthy supplemental material explained in detail what they have done, but I have no way of verifying this. Thus whether this paper belongs in Nature or not is hard to judge. In many ways this paper resembles a lot of the recent cosmological or exoplanet papers in which a very complex analysis method and a very detailed model are presented, but even experts have to study this to understand the reliability. Unfortunately I think that the field is headed more and more in this direction (especially if AI is involved).
I suspect that it requires special conditions in the source (e.g. spin, viewing angle, accretion rate or whatever) to allow this phenomenon to be seen (IZw I has rather unusual properties in the x-ray and is a member of a small class of objects) and I strongly recommend that this be added to the paper. If this is not true then similar phenomena should be seen in other sources and the authors should add a sentence or 2 giving the implications for other objects.

As both referees pointed out in their reports the result does not 'strike one' in looking at the energy dependent light curve, timing analysis and spectral analysis separately e.g. there is no 'smoking gun' of the predicted relativistic effect, but the authors make a strong case that when one combines all 3 techniques other explanations of the data do not work nearly as well.

Comments on the paper
The paper needs to be slightly re-organized: for example there are several places where the size is discussed and several places where the authors jump between a compact 'simple' coronae and 'complex' corona which has both an extended and compact component and its hard to keep track

nature portfolio

of which model is being used. I believe the paper would benefit if the authors focused on the key result ‘ that the proper interpretation of the spectral and timing information requires that parts of the signal be due to “reverberation off the back side of the disk, bent
around the black hole and magnified by the strong gravitational field, a key prediction of General Relativity.” The parts of the paper on the structure of the corona and the need for acceleration of the corona, while interesting, distract from the key result in the presentation.

Detailed Comments

1) In the first section I would refer to figure 2 after the words “accretion disk is clearly detected” – however as the other referee and I pointed out this signal is not clear in the data and needs to be ‘teased out’ with the sophisticated analysis. Thus I would say that the signal is detected after a deep analysis of the data.

2) On pg. 1 line 30-35 I would make it clear that the geometry of the x-ray emitting region is complex and not a lamppost- but that the paper will ‘occasionally’ use the lamppost approximation for (explain the reasons). The first use is on lines 59-62
where a simplified model is used,

3) There is confusing text with respect to absorption in this object. In line 36 it says that there are absorption features which indicate a variable outflow-e.g. the Silva et al paper which argues for a variable warm absorber, but later in the text it says that you fit the spectrum with a simple absorbed power law (with hydrogen
column density corresponding to the Galactic column density along the line of sight to I Zw 1, N(H) =4.6 × 10!' cm"- however in the supplementary section to justify the exclusion of E< 3 keV data the text says “At X-ray energies below 3keV intrinsic absorption has previously been found to significantly impact the X-ray spectrum of I Zw 1, we therefore concentrate on the 3-50keV band in order
to obtain the cleanest measurements of the reflected X-rays from the inner disk.” Also the light curve in the 1-3 keV band is shown and there is reference to a soft excess which implies fitting below 3 keV.

Please clarify what was done and how this apparent set of inconsistencies can be reconciled.

4) The paper does not state the confidence level of the error bars (1 sigma, 90% etc.) in any of the tables or text and thus it is difficult for the reader to judge the significance of the results.

5) In fig 1 it would be good to indicate the flares that the paper is interpreting as being the result of “reverberation off the back side of the disk”

6) There are many places in the text and supplementary section where a height of
~4rg is derived. It’s not clear why there are so many and how they are connected. I would suggest some connection in the text between these estimates of the coronas height.

7) In figure 5 what is the difference between the gray and the blue shading in the right panel? - and what does the shading represent (1 sigma uncertainty or something else). What is the color scale in the left panel_e.g. red is an intensity of x or N sigma above the mean value or ??)

8) In fig 2 what does the vertical dashed line represent?

9) As expressed above page 2 of the supplementary material (which is not numbered or paged so its hard to refer to parts of it) is quite confusing about which energy bands are actually used- 1-50 or 3-50 keV- please clarify .Figure 5 only shows 3-50 keV but figure 7 shows .5-10 keV and figure 9 3-50 keV. This discussion continues in the section called “Variability of the X-ray spectrum: where it says “The soft X-ray spectrum is not significantly affected by absorption,

such that we are able to describe the spectrum below 1 keV with a commonly employed phenomenological model of soft excess emission described by a black body, and absorption from neutral gas. This provides additional constraints on the maximum reflected flux that could contribute to the soft X-ray band, improving the precision to which the reflection fraction is measured." This definitively indicates that despite what it says earlier about analysis of the 3-50 keV band that the authors have analyzed the .5-50 keV spectrum.

10) Need a reference for the factor of a few that spectral dilution changes the apparent timescales.

11) Why do the authors pick the narrow frequency range (5-7x10-4 Hz) to show the lag in ?? what is special about this frequency range?

12) What do the authors mean by "the emissivity falls as r-3, as expected for the illumination of a disk by a central source in flat spacetime, with no gravity."

13) Please give a reference for "If the reflection fraction drops due to the acceleration of the corona away from the disk, it is expected that the photon index decreases,"

14) Please define Compton amplification : is this the mean energy gain per scattering of a photon, the increase in the total power of the scattered spectrum or something else??

15) In table 1 there are no values in the emissivity section for the 'flares" – please explain

16) The figure 10 caption should explain how the time in this figure maps to figure 1- the right panel time axis is not given. The is a typo and part of the caption should read "between the bands appear at 6.8 ks and at 18.8 ks."

Referee #2 (Remarks to the Author):

I thank the authors for their serious consideration of my previous comments and the additional work/figures that result. I am still left very uneasy by the fact that the core result replies upon pattern matching of a very small number of features in the time-domain --- in essence, I remain unconvinced that this variability pattern could not arise randomly from red/pink noise variability that becomes incoherent at high frequencies.

However, at this point, I will not object to publication in Nature given one additional request - I ask that the authors clearly add (in the main paper) a formal statement of the significance of this the echo (and describe in the SI the procedure used to come to that number).

Author Rebuttals to First Revision:

We thank both referees for their feedback, which has been extremely valuable in both improving the analysis and the presentation of these results.

Referee #1

The paper needs to be slightly re-organized: for example there are several places where the size is discussed and several places where the authors jump between a compact 'simple' coronae and

‘complex’ corona which has both an extended and compact component and its hard to keep track of which model is being used. I believe the paper would benefit if the authors focused on the key result ‘ that the proper interpretation of the spectral and timing information requires that parts of the signal be due to “reverberation off the back side of the disk, bent
around the black hole and magnified by the strong gravitational field, a key prediction of General Relativity.” The parts of the paper on the structure of the corona and the need for acceleration of the corona, while interesting, distract from the key result in the presentation.

Following the recommendation of the referees and the Editor, we have removed the sections of the paper discussing the structure of the corona, its acceleration and evidence for a drop in temperature. The paper now focuses on the key result of the X-ray reverberation and the X-ray echoes from behind the black hole.

In addition, we have addressed the detailed comments made by the referees, although some of these comments apply to the removed sections of the paper about the structure and acceleration of the corona, which are no longer included.

Detailed Comments

1) In the first section I would refer to figure 2 after the words “accretion disk is clearly detected” – however as the other referee and I pointed out this signal is not clear in the data and needs to be ‘teased out’ with the sophisticated analysis. Thus I would say that the signal is detected after a deep analysis of the data.

We have made clear in the introductory paragraph that the signal arises from a deep analysis of the observations. In this context, “the reflection from the accretion disk is clearly detected” was intended to mean the spectral signatures of reverberation in the spectrum obtained from NuSTAR (the broad iron line and the Compton hump). This has now been clarified.

2) On pg. 1 line 30-35 I would make it clear that the geometry of the x-ray emitting region is complex and not a lamppost- but that the paper will ‘occasionally’ use the lamppost approximation for (explain the reasons). The first use is on lines 59-62
where a simplified model is used,

We have clarified that the geometry of the corona is complex and that we use the ‘lamppost’ approximation in some aspects of the analysis, and have added a sentence to the section of the main paper “X-ray Echoses from Behind the Black Hole”:

“While the structure of the corona is complex, the rapid variability is dominated by the compact core, thus we adopt a simplified model in which the rapidly variable X-ray emission originates from a compact, point-like source [17].”

3) There is confusing text with respect to absorption in this object. In line 36 it says that there are absorption features which indicate a variable outflow-e.g. the Silva et al paper which argues for a variable warm absorber, but later in the text it says that you fit the spectrum with a simple absorbed power law (with hydrogen
column density corresponding to the Galactic column density along the line of sight to I Zw 1, N(H) =4.6 × 10!' cm"- however in the supplementary section to justify the exclusion of E< 3 keV data the text says “At X-ray energies below 3keV intrinsic absorption has previously been found to significantly impact the X-ray spectrum of I Zw 1, we therefore concentrate on the 3-50keV band in order
to obtain the cleanest measurements of the reflected X-rays from the inner disk.” Also the light curve in the 1-3 keV band is shown and there is reference to a soft excess which implies fitting below 3 keV.

Please clarify what was done and how this apparent set of inconsistencies can be reconciled.

We specifically focus on the 3-50 keV spectrum when fitting the relxill model to the spectrum of the X-rays reflected from the accretion disk, since we see no evidence for additional spectral features from other emission or absorption components in this energy range. Specifically, it is below 1 keV that complexities in the spectrum arise. The outflows reported by Silva et al. 2018 imprint absorption features on the spectrum below 1 keV, but the best-fitting photo-ionization models of these outflows do not imprint significant features on the spectrum above 3 keV (and we see no evidence for any such features). In addition, below 1 keV, the X-ray spectra of AGN are dominated by the so-called ‘soft excess’ which forms part of the reflected X-ray emission from the disk (though there may be other contributions to it). This feature is not fully described by the relxill reflection model since it requires detailed modelling of the density and ionizations structure of the disk that is beyond the scope of this paper and would distract from the key result. The soft excess is variable in nature and continues to affect the detailed shape of the spectrum up to 2 keV, even though it is no longer the dominant emission component, thus it does not affect the interpretation of the 1-3 keV continuum band. Uncertainties in all of the parameters derived from the model are marginalized over all possible values that describe the 3-50 keV spectrum without incorporating the lower energy part of the spectrum in the fitting.

We include the X-ray spectrum below 3 keV in Fig. 4(a), showing the variability of the reflection spectrum. The variability in the soft X-ray emission can be seen here, but a detailed description of it was not included in the relxill model that was applied to the time-averaged spectrum. The affects Galactic absorption can be readily accounted for from detailed databases available of Galactic column densities such that this figure represents the emission intrinsic to the source.

4) The paper does not state the confidence level of the error bars (1 sigma, 90% etc.) in any of the tables or text and thus it is difficult for the reader to judge the significance of the results.

We have specified the confidence level of the error bars in the text and in the captions to the figures and tables.

5) In fig 1 it would be good to indicate the flares that the paper is interpreting as being the result of "reverberation off the back side of the disk"

We have created a new Fig. 1 that incorporates the old Fig. 1 and Fig. 2 and focusing on the flares of interest. In Fig. 1(a), the flares are highlighted and Fig. 1(b) zooms in on this time interval.

6) There are many places in the text and supplementary section where a height of ~4rg is derived. It's not clear why there are so many and how they are connected. I would suggest some connection in the text between these estimates of the coronas height.

We measure the height of the corona using different methods, first from fitting a model to the average time delay between the continuum and the iron K line, and then determining the corona height that is required to explain the timing of the narrow peaks that we interpret as the re-emerging X-rays, or echoes from the back side of the disk. The intention of deriving these two estimates is to demonstrate that they are consistent with one another and that the narrow peaks are as is expected for a flare in a corona at the previously-measure height. This has now been clarified in the text.

7) In figure 5 what is the difference between the gray and the blue shading in the right panel? - and what does the shading represent (1 sigma uncertainty or something else). What is the color scale in the left panel_e.g. red is an intensity of x or N sigma above the mean value or ??)

In Fig. 5(b), the blue shading shows the spectrum in each time bin. The grey shading shows the spectrum in the previous time bin (i.e. the previous panel), so the variation with time can be seen.

The shading corresponds to the residual or difference between the observed spectrum and the best fitting power law (which is essentially 'drawn' through the reflection features, such as the iron K line, thus residuals are negative and positive). The residual is in units of ct/s/keV. We have added a color bar and specified the units in the text.

8) In fig 2 what does the vertical dashed line represent?

In the simplified reverberation model that we use to explain the peaks, the flare is modelled as a short flash of emission from a point-like corona. The red dashed line indicates the timing of this short flare in the model and represents the fact that the disk responds most strongly to the continuum emission in just the first part of the flare. This has been clarified in the caption and is described in the main text.

9) As expressed above page 2 of the supplementary material (which is not numbered or paged so its hard to refer to parts of it) is quite confusing about which energy bands are actually used- 1-50 or 3-50 keV- please clarify .Figure 5 only shows 3-50 keV but figure 7 shows .5-10 keV and figure 9 3-50 keV. This discussion continues in the section called "Variability of the X-ray spectrum: where it says "The soft X-ray spectrum is not significantly affected by absorption,
such that we are able to describe the spectrum below 1 keV with a commonly employed phenomenological model of soft excess emission described by a black body, and absorption from neutral gas. This provides additional constraints on the maximum reflected flux that could contribute to the soft X-ray band, improving the precision to which the reflection fraction is measured." This definitively indicates that despite what it says earlier about analysis of the 3-50 keV band that the authors have analyzed the .5-50 keV spectrum.

This comment relates to the analysis that was used to infer the acceleration and temperature change in the corona which has now been removed from the paper.

10) Need a reference for the factor of a few that spectral dilution changes the apparent timescales.

This reference has been added.

11) Why do the authors pick the narrow frequency range (5-7x10-4 Hz) to show the lag in ?? what is special about this frequency range?

When measuring the X-ray reverberation lag, it is necessary to determine the frequency range in which the lag is measured at the highest signal to noise ratio, as discussed in Uttley et al. 2014 and also Kara et al. 2016. This represents the range of timescales on which reverberation is the dominant process that drives the variability we observe. We have clarified this in the text and noted that this is the frequency range over which X-ray reverberation in this object had previously been detected at the highest signal to noise ratio.

12) What do the authors mean by "the emissivity falls as r-3, as expected for the illumination of a disk by a central source in flat spacetime, with no gravity."

This was related to the structure of the corona, which has now been removed from the paper.

13) Please give a reference for "If the reflection fraction drops due to the acceleration of the corona away from the disk, it is expected that the photon index decreases,"

This discussion of the acceleration of the corona has now been removed from the paper.

14) Please define Compton amplification : is this the mean energy gain per scattering of a photon, the increase in the total power of the scattered spectrum or something else??

This was related to the measurement of the change in coronal temperature, which has now been removed from the paper.

15) In table 1 there are no values in the emissivity section for the 'flares" – please explain

This was related to the measurement of the change in coronal structure, which has now been removed from the paper.

16) The figure 10 caption should explain how the time in this figure maps to figure 1- the right panel time axis is not given. The is a typo and part of the caption should read "between the bands appear at 6.8 ks and at 18.8 ks."

We have put all the light curves onto a common time axis such that their relationship to one another can easily be seen. The time period of the flares is now defined to begin at t=0 (hence, in Fig. 1, thereis a negative time range before the flares begin).

We thank the referee for noticing the typo, which has now been corrected.

<u>Referee #2</u>

I thank the authors for their serious consideration of my previous comments and the additional work/figures that result. I am still left very uneasy by the fact that the core result replies upon pattern matching of a very small number of features in the time-domain --- in essence, I remain unconvinced that this variability pattern could not arise randomly from red/pink noise variability that becomes incoherent at high frequencies.

However, at this point, I will not object to publication in Nature given one additional request - I ask that the authors clearly add (in the main paper) a formal statement of the significance of this the echo (and describe in the SI the procedure used to come to that number).

A formal statement of the significance of the echo feature has been added to the main text.

We have added a new section to the Methods detailing how we derive this significance level. We do this by estimating the probability that the feature we observe arises due to random variations, either due to (1) Poisson noise or (2) random red noise variability that is uncorrelated between the light curves. By measuring the coherence of the light curves in different energy bands, we are able to place an upper limit on the fraction of the red noise variability that can be uncorrelated to be consistent with the observations, now discussed in the Methods.