The submitted paper develops a novel theory for the temporal evolution of the critical Shields stress as a function of flow magnitude. The theory is calibrated using field data to show that it improves estimates of critical Shields stresses in a gravel bed river. This is the first paper to develop an equation that predicts both flow strengthening and weakening effects on critical Shields stresses. It is also the first paper, to the best of my knowledge, that develops such a theory using field data; all other equations for flow effects on temporal changes in critical Shields stresses are based on laboratory data. I believe this study constitutes a major step forward in our potential to predict critical Shields stress changes over time. I have a few major comments that I think should be easy to address, which are mostly about clarifying the assumptions/calibration in the equations or placing this work in the context of equation application. Other than these comments and some minor line by line comments, I think the paper is ready for final publication. - Elowyn Yager

We thank the reviewer for her thoughtful and thorough review of our submission. Below, we respond to the reviewer's comments and describe the edits we plan to make to the text addressing their helpful review.

Major comments:

Calibration of tau*c_min and tau*c_max: This may just be a matter of preference, but I think that Figure S1 (how tau*c_min/max is calibrated) might be better suited in the main text rather than in the supporting information because this demonstrates how the empirical parts of your main equations were developed? I think it might help your reader better understand, for example, why tau*c_max and tau*c_min vary with slope?

We agree with the reviewer's suggestion. We have refined Figure S1 and plan to include it in the updated version of our MS as a new Figure 2. We have also added the observed Erlenbach values of t*c as the reviewer later suggests. We also added supporting language to the text to clarify how this parameterization was developed.

Here is the revised figure and new caption:

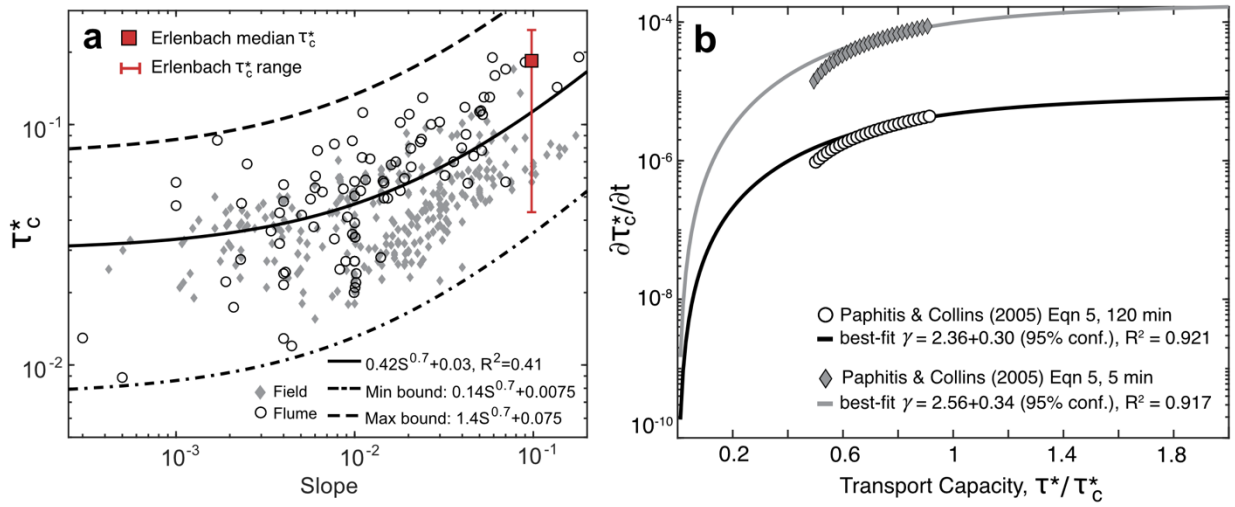

Figure 2. a) Compiled τ_c^* data as a function of slope from both field studies (grey diamonds) and flume experiments (open circles). Range of τ_c* from the Erlenbach field site is plotted in red. Compiled data are limited to slopes, S < 0.2 and median grain sizes, D50 ≥ 2 mm. Most data were compiled by Prancevic and Lamb (2015), building on Buffington and Montgomery (1997), with additional data from Olinde (2015) and Lenzi et al. (2006). b) . Calibration of strengthening term exponent, γ, based on Paphitis and Collins (2005) for both the shortest conditioning time (5 minutes) and longest conditioning time (120 minutes) spanned by the Paphitis and Collins (2005) data. Regressions to their best-fit empirical equation give gamma exponents within uncertainty of each other.

We will also revise the text, starting at Line 95 to read:

> *Figure 2a shows that the empirical τ_cmin^* and τ_cmax^* relations in Equation 2 capture the slope-dependence of gravel thresholds of motion compiled in both flume and field settings, while asymptoting to reasonable bounds at low channel slopes (e.g. Johnson, 2016; Lamb et al., 2008; Prancevic & Lamb, 2015). Field and flume data were weighted equally in the best-fit regression, removing possible bias from there being ~3.5 times more flume data points. Minimum and maximum bounds were determined visually to accommodate almost all data points, assuming a consistent best-fit exponent (0.7). Our model is similar in form to that of Johnson (2016)."*

Also since all assumed parameter values impact the equation accuracy and potential future application, I think it is useful for your reader to see in the main text how this calibration was performed? I think you also might need to discuss (in the main text) that you are calibrating tau*c_min/max to spatial (and not temporal) variations in observed tau*c. This method of calibration indirectly assumes that all measured spatial variability in tau*c (for a given slope) is because of the same factors that would cause temporal changes in tau*c in a given stream. However, variations of tau*c between streams have been attributed to factors that may be independent of those that cause tau*c temporal variations and I think this likely needs to be acknowledged? For example, this scatter in tau*c between different streams with a given slope has been attributed to differences in relative roughness effects (e.g., Lamb et al., 2008,

2017) or differences in the underlying grain size distribution (e.g., Kirchner et al., 1990; Buffington et al., 2002; Shvidchenko et al., 2001). To help support your assumption of a kind of space for time substitution, could you plot the measured values of tau*c for the Erlenbach on Figure S1? Or could you simply place arrows or lines on this plot that denote the range of temporal variation in tau*c in the Erlenbach? If the range of temporal variations in tau*c in the Erlenbach is similar to the spatial variations in tau*c between many different sites, this could help further justify your assumption that all (or most) spatial variations in tau*c are being caused by the same factors that control temporal variations in tau*c at a given site (i.e., that you can use the range of spatial variations as a proxy for the range of temporal variations)?

The reviewer is correct that this approach implicitly assumes that the range of variability in tau*c observed at a given slope is comparable to the variability in threshold introduced by flow history effects. However, the full range of tau*c observed at the Erlenbach is equivalent to the range of thresholds across the flume and field data compiled in Fig. 1A. Adding the Erlenbach range to Fig. 1A shows this nicely. We added the following to the main text to point this out and clarify:

Added at Line 100:

> *We acknowledge that many other factors beyond slope may influence the possible range of τ_c^{*} variability a given sites, including grain size distributions (e.g., Parker, 1990) and differences in relative roughness. In the absence of independent τ_cmin^{*} and τ_cmax^{*} constraints that could be used to describe a particular field site, the compilation should reasonably represent the range of possible values.*

Added at (now) Line 617:

> *The temporal variability in τ_c^* observed at the Erlenbach is equivalent to the full range of τ_c^* observed in flume and field data for equivalent slopes (Fig. 1A).*

Calibration of gamma: Similar to my comments on tau*c_min and tau*c_max, I think (again, maybe a matter of preference) the material on gamma calibration (Figure S2 and associated text) might be better suited to be placed in the main manuscript rather than as supporting material. Since the calibration of gamma affects all other parameter values and is a major part of the proposed equations, I think it likely belongs in the main text? I think you have space (?) within the journal page limitations to move this text to the main document and Figures S1 and S2 could be combined into different panels of the same figure in the main text? I also really appreciated the discussion about the uncertainties associated with the gamma calibration using the sand data/equation of Paphitis and Collins (2005). I understand that grain size does not enter directly into their equation but it will enter into your conversion from utc to tau*c. What grain size did you use in this conversion, some representative grain size from their experiments, the grain size from the Erlenbach, or some other size? To further help make your case that using sand data to calibrate gamma is relevant for gravel bed channels, I wonder if you could also frame this analysis not only in terms of particle Reynolds

number values but also tau*c values? For example, are your back calculated values of tau*c (from utc from their equation) within the range of tau*c values typically reported for (hydraulically rough flow) gravel bed rivers? If these tau*c are in the reported range for gravel rivers, I think this would further support your reasoning of applying this sand-based equation to gravel beds since all you really care about is tau*c and the time derivative of tau*c anyway?

We have taken the reviewer's suggestion and combined Figures S1 and S2 into a new figure, now displayed as Figure 2 in the revised text. We have also added the details of the calibration of gamma to the main text as suggested by the reviewer, starting at Line 150.

To answer the reviewer's question regarding grain size: We are using reported values of critical shear velocity before and after low flow conditions reported by Paphitis and Collins (2005) for their coarse sand experiments (D=0.774 mm). Mathematically the grain size does not actually come into this calculation or conversion, because their equation is nondimensional, giving the ratio of critical shear velocity after flow conditioning to the critical shear velocity without flow conditioning. To convert from a shear velocity ratio to a shear stress ratio we only had to square the shear velocity ratio. While their results may depend on the grain size they used, the conversion from their nondimensional relation to ours does not. We tried to clarify this in the incorporation of the calibration into the main text.

Revised text starting at Line 150:

We calibrated γ using experimental data from Paphitis and Collins (2005) to reduce the number of free parameters in the model. Paphitis and Collins (2005) conducted experiments using fine, medium and coarse sand, in which they systematically varied both the conditioning time (E_D, the duration of flow below the threshold condition) between 5 and 120 minutes, and the ratio of shear velocity (u_τ) to initial critical shear velocity ($u_{\tau ci}$). The ratio of shear velocities they explored was between 70% and 95% of critical, which corresponds to conditioning flow shear stresses between 50% and 90% of the initial critical shear stress. In their study, Paphitis and Collins (2005) present the following equation to describe their experimental data:

$$\frac{u_{\tau c(t)}}{u_{\tau ci}} = 1.05\left[1 - 0.01e^{(-0.005E_D)}\right] + \left[0.005 + 0.1\left(\frac{u_\tau}{u_{\tau ci}} - 0.7\right)\right] ln(E_D) + 0.06\left[10^{-7\left(0.97 - \frac{u_\tau}{u_{\tau ci}}\right)}\right] \quad (3)$$

for $0.7 \leq \frac{u_\tau}{u_{\tau ci}} \leq 0.95$ *and* $E_D \leq 120$ *minutes, where* $u_{\tau c(t)}$ *is the critical shear velocity following low flow conditioning. This function fits the experimental data with a correlation coefficient of 0.83 (i.e.* R^2*=0.69).*

To calibrate γ, we calculate $u_{\tau c(t)}/u_{\tau ci}$, for a range of E_D and u_τ, using values of $u_{\tau ci}$ for their coarse sand (D=0.774 mm) experiments, corresponding to a critical shear velocity of $u_{\tau ci}$=0.0195 m/s. We then square $u_{\tau c(t)}/u_{\tau ci}$ to convert to a Shields stress ratio. We then numerically calculate the partial derivative of Equation 3 with respect to time, $\partial\tau_C^*/\partial t$. Figure 2B shows a nonlinear regression (using Matlab's cftool) of the strengthening term in our model (Equation 1) to $\partial\tau_C^*/\partial t$ calculated from Equation 3.

This regression provides a best-fit estimate of γ, including empirical regression uncertainties. For below-threshold conditions explored in the Paphitis and Collins (2005) experiments, the weakening term in Equation 1 is zero. Given this, the calibration of γ described here is not influenced by other model parameters, in particular, the weakening exponent ε. Our reported 95% confidence interval on γ only represents the empirical regression uncertainty when fitting our function to Equation 3. Therefore, it is likely that a somewhat wider range of γ may be able to fit the range of the experimental data from Paphitis and Collins (2005), and the true range of possible values may be somewhat larger than 2.5+/-0.32. We hypothesize that a range of γ =2-3 may be possible.

We use the data and fitting function of Paphitis and Collins (2005) to calibrate γ because it is the most complete and internally consistent dataset that we are aware of with sufficient constraints to describe the evolution of τ_c^* as a function of both transport capacity and time. Nonetheless, a possible limitation of applying these experimental data to calibrate our model is that the Paphitis and Collins (2005) experiments were conducted with unimodal sand. Specifically, boundary Reynolds numbers in their experiments are transitional between hydraulically smooth and hydraulically rough flow. For the coarsest grains they use (D_{50}=0.0774 mm), boundary Reynolds number $Re_w = u_\tau k_s/\upsilon \approx 15$, where k_s is a roughness length scale assumed to be D_{50}, and υ is the kinematic viscosity of water. If k_s was instead assumed to be a multiple of D_{50} (such as k_s=3.5D_{84}) then Re_w would be closer to the hydraulically rough flow criteria of $Re_w \geq 100$. It is also worth noting that grain size did not explicitly factor into Equation 3, beyond its implicit control on $u_{\tau ci}$. The insensitivity of their results to grain size suggests that the results may not depend significantly on grain size or on hydraulically rough flow being fully developed. Further, converting critical shear velocities from their coarse sand experiments to critical Shields stress yields τ_c^*=0.031, consistent with critical Shields stresses reported for gravel data sets (Fig. 2a).

Application of equation: The discussion mentions that the proposed equation could be used to improve critical Shields stress estimates at other field sites as long as some site specific calibration is conducted. I think it might be useful here to discuss what level of calibration would be needed for practical application of this approach? You conducted extensive calibration of the equations because you had measured tau*c over time in the Erlenbach. In a river in which tau*c is not known and a user would instead like to predict tau*c, what type of calibration would be necessary? Would someone need to measure tau*c over time for a certain period of time? Would calibration for one year of data on tau*c be sufficient or would many years of data be needed? If many years of data are needed, would this intensive data requirement potentially limit the application of this approach to very well studied rivers? Or would only relatively limited data be needed for calibration, which would allow for a more broader application of this approach? I think some discussion on this would be really helpful to understand how someone might apply your approach and the potential data-based limitations in using the approach?

Thank you for the thoughtful comment. Calibration requirements will vary depending on the intended application of the equation. The key question is whether the user wants to estimate the long-term variability in critical Shields stress, possibly by generating a probability distribution over a long discharge record, or if they want to refine estimates of bedload transport for individual events. Each application would require different calibration data based on its scope and required accuracy. While we cannot outline every hypothetical case in the main text, we have added a paragraph acknowledging potential calibration requirements and the need for further study on the sensitivity of model parameters to individual sites. We decided to focus on a similar approach to the analysis that we present in the paper and leave allusion of event-scale parameterization to the final paragraph.

If we want to reliably estimate the distribution of tau*c, a useful starting point is the "rule of thumb" from central limit theorem that approximately 30 independent samples are typically needed to capture the distribution of normally distributed data. Given that the Erlenbach tau*c data appear to follow a normal distribution, the model may require calibration based on at least 30 events to encompass the expected range of variability at a site. However, this assumes independence between measurements, which we know is not strictly true. Another approach could be to calibrate the model based on the timescale over which tau*c remains correlated. Masteller et al. (2019) found that at the Erlenbach, tau*c exhibits "memory loss" after approximately 10 events, suggesting that this number of events could be sufficient to calibrate the model to capture the trajectory of tau*c over time. However, a major challenge in model calibration will always stem from hydrologic variability and river self-organization itself. Small floods that strengthen the bed occur frequently, whereas weakening events which exceed the bankfull condition are rarer. As a result, robust calibration of weakening parameters will be significantly more difficult than for strengthening. Again, these approaches would be most applicable for studies focusing on only the initiation of motion, as done in our paper. We acknowledge that the requirement of 10-30 observations only to calibrate the model may be labor intensive, so more work is needed to better determine the potential variability in the k1, k2, and epsilon parameters introduced in the model. However, the variability of some of these parameters with flow history, bed slope, and grain size distributions can be more readily explored in flume experiments, as we suggested in the closing paragraph.

We've added the following paragraph to the discussion to clarify these points:

> *The calibration that we have performed here leverages an extensive dataset of direct measurements of τ_c^*. While similar datasets are available for a small subset of rivers (cites), most gravel bed rivers lack time series data of τ_c^*. However, we find that the range of temporal variability observed in our calibration dataset is consistent with the existing data compilations of τ_c^* across a range of slopes (Fig. 2a) suggesting that these data compilations may provide reasonable preliminary constraints of minimum and maximum bounds on τ_c^* in future applications of the model. Reach-averaged starting values of τ_c^* could be estimated based on bed grain size and bankfull geometry. Additional model calibration will vary depending on the*

intended application of the equation. To characterize long-term variability in τ_c^*, calibration of the model over approximately 30 transport events may be needed to reliably capture the expected variability of τ_c^*, assuming a normal distribution, as observed at the Erlenbach by Masteller et al. (2019). However, this commonly used minimum sample size assumes independent observations, which does not apply here. An alternative approach is to calibrate the model based on the number of subsequent events over which τ_c^* remains correlated. Masteller et al. (2019) also found a loss of correlation between τ_c^* values after 10-13 transport events. This number of events may be sufficient to calibrate the model to capture the trajectory of τ_c^* over time. We recognize that requiring 10–30 observations may not always be feasible, and future studies should assess the necessary level of calibration for different applications. Further research is needed to evaluate the potential variability of the k1, k2, and ϵ parameters, especially their sensitivity to discharge distributions, riverbed grain size, and bed slope. Controlled flume experiments could offer a systematic approach to investigating this variability.

Minor comments by line number/figure number etc.

36-39. “For example, hysteresis in bedload transport rates is often observed between the rising and falling limbs of individual floods (Hsu et al., 2011; Mao, 2018; Mao et al., 2014; Pretzlav et al., 2020; Reid et al., 1985; Roth et al., 2017). Dynamic threshold evolution over the duration of a flood event is implied by the observed change in bedload transport rate.”I am not fully sure that you can state that all changes in bedload transport during an event are caused by threshold evolution? Changes in bedload transport rates during an event (hysteresis) could be caused by other factors. Transport rates could also change over an event because the flow hydraulics within a channel or the morphology have evolved during an event, which will both alter the applied shear stress without necessarily having to change the threshold of sediment motion. Similarly, many studies on hysteresis attribute changes in bedload transport rates to changes in sediment supply (e.g., landsliding) to the river during an event. Although sediment supply can influence the threshold of motion, it can also influence other variables that control the bed load transport rate such as bed grain size, channel bed roughness, channel topography etc. Hysteresis in bedload is likely caused by a variety of factors and I think you probably need to reword this text slightly?

Fair point – reworded to: "Dynamic threshold evolution over the duration of a flood event **is one mechanism** that can results in the observed changes in bedload transport rate".

50-54 and other locations. Thanks for this citation (!) but Yager et al., 2012 didn't investigate changes in the critical Shields stress after floods as implied here? They attributed increases in bedload transport rates after floods to a higher sediment supply during floods from landsliding and showed that this sediment supply would alter the applied flow shear stress by changing channel morphology? I think this might be a more relevant reference for when you are discussing sediment supply effects in the introduction and discussion rather than as cited in this location and others later on?

Fair enough. In our reading of the paper, we interpreted the rapid reduction in sediment availability following an extreme event and the subsequent slowdown in in the changes in sediment availability (Fig. 1c) to imply a decrease in transport efficiency with time since the extreme event as the bed went from a relatively weak state to a relatively strong state. We removed the reference in cases where it was mis-referenced.

70-75. I think you might want to briefly discuss and review here the other developed equations for the temporal evolution of critical shear stresses as a function of flow properties? I think this could more specifically set the stage for missing component of these equations that you are explicitly trying to address here (e.g., none calibrated with field data, none include both strengthening and weakening)? I think this could better highlight the novelty of what you have done. For example, you could mention Ockelford et al. (2019) used flume experiments to develop an empirical equation for the temporal strengthening of the critical shear stress using the duration of a certain flow magnitude. Also, a brief discussion of Paphitis and Collins (2005), and their flume based equation for strengthening based on flow conditions, could also be mentioned here?

Thanks for the suggestion! We added a brief description of these models in the final paragraph of the introduction. It now reads:

> *While empirical evidence exists for systematic, flow strength-dependent temporal variations in thresholds for motion, few equations exist that quantify feedbacks leading to threshold evolution which can be incorporated into existing bedload transport models. Relatively simple model formulations have been proposed to describe temporal bed strengthening as a function of the duration of bed exposure to a constant, inter-event flow magnitude and an initial τ_c^* based on experimental data (e.g. Ockelford et al., 2019; Paphitis and Collins, 2005). However, because these models only focus on inter-event strengthening effects, they cannot capture decreases in τ_c^*. Johnson's (2016) model predicts τ_c^* evolution as a function of sediment supply and allows for both strengthening and weakening effects. Nonetheless, this model is an incomplete*

description of τ_c^* evolution because it does not account for riverbed strengthening or weakening directly caused by the flow. Notably, to our knowledge, none of these equations have been used to describe field observations of temporally varying τ_c^*. Our goals in the present work are (i) to propose a new model in which τ_c^* evolves as a function of flow magnitude and encapsulates some memory of past shear stresses as reflected in the changing state of the riverbed, and (ii) to evaluate whether the model can broadly capture annual strengthening and weakening trends as a function of discharge, as observed in Erlenbach field data (Masteller et al., 2019).

Equation (2) and lines 99-100. I understand how B=0 when tau*c=tau*c_max but I don't understand under what conditions B is equal to one because I think (?) the equation becomes undefined when tau*c=tau*c_min? Can you please explain under what conditions B=1? Also, can you explain how you deal with the situation when tau*c=tau*c_min?

We apologize. Our equation had a significant typo which we overlooked, and have now corrected. The equation now correctly gives 1 at tau*=tau*min, and 0 at tau*=tau*max.

Tau*c_max and tau*c_min represent bounds of our model. Because of the feedback term, B, tau*c can approach, but cannot reach these values. Equation 1 is only defined within these limits. We have modified the text to clarify this starting at Line 95:

In this treatment, we assume that τ_cmin^* and τ_cmax^* are limits that the threshold can approach but does not reach due to the feedbacks implemented by the B parameter (Equation 2). Equation 1 is only defined between these bounds.

Figure S1 caption. Can you explain how field and flume data are equally weighted when conducting the regression? How are you assigning weights to the data to offset the greater number of datapoints in flume experiments? Are the labels of the figure correct because the caption says there are 3.5 more datapoints from flume experiments but in the figure, it looks like there are more field data (blue diamonds) than flume data (red circles)? The figure caption also says the best fit exponent is 0.36 but in the figure legend, the exponent appears to be 0.7 in all equations; can you please make these consistent?

Thanks for spotting the error in our figure legend. The labels were flipped. We have updated them in our revised figure. We have also corrected the best-fit exponent to 0.7 in both the figure and the text.

The regression was done using nonlinear curve fitting in Matlab, with all the field-derived data points together having equal weight in the regression as all the flume-derived data points. In other words, because there were ~3.5 times more flume data points than field data points, each field data point was weighted 3.5x more strongly in the regression compared to each flume point. Overall, the field and flume points each contribute half of the weight in the regression. We have added the following to the text starting at Line 100 to help clarify:

> *Field and flume data were weighted equally in the nonlinear best-fit regression, removing possible bias from there being ~3.5 times more flume data points. Minimum and maximum bounds were determined visually to accommodate almost all data points, assuming a consistent best-fit exponent (0.7). This approach assumes that the total observed variability in threshold at a given slope is comparable in scale to the temporal variability in threshold induced due to flow history efforts. We acknowledge that other factors may influence variability in threshold between sites, including grain size distributions (cites) and differences in relative roughness (cites). However, this approach allows the model to explore the maximum potential variability in threshold due to temporal variations in bed shear stress.*

170-200. In the results, a mix of median parameter values (k1, k2 etc.) for the annual best fits and mean variable values for the average best fit are used. Is there a reason for mixing the use of medians and means of parameter values? I think it might be better to use one of these consistently (median or mean parameter value) or explain why you are using medians for the annual best fits instead of means, which would be more directly comparable to the average best fit parameter values for all years combined?

Sorry for the mix-up – this was an inconsistency so thank you for catching it! We can see how that can be confusing! The differences between the means and medians are very slight, so we will go back through the document and report means for consistency in the revised version of the MS. We also recognise that some additional ambiguity may be coming from our use of "average" to describe the best-fit parameter combinations that, yield the minimum MAE when MAE is averaged across all sample years. Here we are using an average so that each year is weighted equally (as described in Line 250). We are also using a Mean Absolute Error (MAE), so admittedly we are using the terms average, mean, and median a lot in this section. To try and address/reduce some of this, we have replaced "average" with "combined" to indicate that the combined MAE is reflective of the model runs that best describe the entire dataset.

203-205. But is the lower range of values for k2 and epsilon because the model is less sensitive to these parameters or could it also be because you only had three years in which weakening (which these parameters represent) was dominant, which I think you kind of imply in the discussion? Can you please clarify here?

Apologies but we're not sure what the reviewer means when they say "the lower range of values for k2 and epsilon"? Best-fit values for k2 and epsilon both spanned the entire parameter range. The original text reads:

"In contrast, mean MAE values are higher and less variable when binned by k_2 (MAE = 0.108-0.124) and ε (MAE = 0.113-0.117), suggesting that annual model performance is less sensitive to variations in these parameters (Fig. 2b,c)."

We don't expand on this statement here as to keep our results separate from our interpretation. The previous paragraph describes the differences in median MAE for the best-performing models between strengthening and weakening years.