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

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? 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)?

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?

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?

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?

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?

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?

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?

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?

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?

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?