We would first like to thank Robert Mahon for their thoughtful review. In the below document, the reviewers comments are in black; our responses to reviews are in blue italics.

The authors present a systematic comparison of bedform bedload measurement techniques using a unique dataset. Using field data, as opposed to flume data as is often the case, the authors are able to investigate some of the complexities associated with systems evolving under unsteady flow conditions. The ultimate outcome of this paper can inform decisions on both multibeam sampling and processing strategies as well as the placement of single beam echosounder instrumentation on rivers to monitor bed- load flux. Thus the results of this paper are broadly relevant to river managers as well as to academic geomorphologists.

The overall flow and structure of the manuscript are quite clear. Figures are well placed into the manuscript context and are appropriate for fully describing the nature of the work. While I have no concerns that fundamentally call into question the nature of the science being done, there are a number of points which the authors could clarify or analyses that could be bolstered by more complete discussion. These comments are below:

I would like to see a description of the methods used to extract height and wavelength data from the BTT toolbox as it is a fundamental operation to the analysis in the paper. There are several methods for calculating these parameters, each of which have their respective advantages and disadvantages so it would be good for the authors to describe why the calculations employed in this toolkit are appropriate to their system.

We have added the following to our description of the BTT: "After the BEPs are detrended, the BTT determines the zero upcrossing (i.e. points at which the profile positively crosses zero) and zero downcrossing (i.e. points at which the profile negatively crosses zero). The locations of crests and troughs are determined in the original BEP as follows: a crest is located at the maximum value between a zero up- and zero downcrossing; and vice versa, a trough is located at the minimum value between a zero down- and zero upcrossing. Bedform height is calculated at the vertical distance between crest and downstream trough. Bedform wavelength is calculated as the distance between two successive crests. For a more detailed explanation of the BTT please refer to van der Mark et al. (2008) and van der Mark et al. (2007)."

Were bed elevation surveys corrected for apparent dilation as a function of the time between start and end of each multibeam survey? If not, was this considered and determined to be a negligible effect? See McElroy dissertation 2009, p. 44 (URI: http://hdl.handle.net/2152/1511).

The average difference in time between surveys was around 10 minutes and departures from this were also order 10 minutes, therefore, while we acknowledge that the effect noted by the reviewer and McElroy (2009) is real, we determine it to be a negligible effect in the present study.

A figure demonstrating the cross-correlation results would be good to show, as a lot of discussion is based on issues resulting from velocity calculations.

See figure 3A.

In Page 5 Line 6 the method for estimating wavelengths for the singlebeam experiment is described as the daily average from the repeat multibeam. I wonder if this introduces potential for extra accuracy for this method that may not be possible in a situation in which a single beam fixed echosounder would be employed.

Yes, this is most likely the case. Our single beam flux estimates are probably more accurate than what one might get using a different estimate of bedform wavelength.

I would suggest more discussion of when a situation would arise where you have a measurement or a daily average of bedform wavelengths but only a single beam profile to estimate flux from. An alternative formulation might be to estimate wavelength using a height-wavelength relationship such as Bradley and Venditti, 2017 as this might be a more realistic representation of a likely application (i.e. a deployed single beam sensor established for continuous monitoring).

Analyses of bedform fields throughout the Colorado river in Grand Canyon reveals that the Bradley and Venditti (2017) relations are a poor fit to observations. Because the bedform field is likely not in equilibrium with the flow due to daily fluctuations in discharge, i.e. because the dunes are always adjusting to flow, correlations between instantaneous bedform height or length and flow are not as robust as they otherwise would be. A better approach is to use mulitbeam measurements at multiple flows to develop a site-specific model for bedform dimensions as predicted by flow.

What did the manual process entail for determining bedform velocity? Were you picking crests and tracking them? Looking at the slopes of the forms in the $\eta(x,t)$ field (e.g. in Figure 2D)? It would be critical to determine whether the manual method itself includes any potential sources of bias in order to interpret its relation to the cross-correlation results.

We picked crest locations and tracked them. We have added this information to the main text.

I wonder if other methods for calculating bed velocity might be more appropriate than the cross-correlation method for this application, particularly given the unsteady flow conditions investigated. One example from Ganti et al., 2013 (doi:10.1002/jgrf.20094), their eq. 5 to compute the local velocity based on dividing the temporal change in local elevation by local slope at all points on the bed.

The above mentioned method from Ganti et al. (2013) would most likely not apply to this data set. While computing local velocities is appropriate in flume experiments where the change in time between each successive bed elevation profile is 45 seconds, applying this method to field data with both a coarser spatial and temporal resolution would likely result in large errors (likely larger than those associated with the cross-correlation method).

Were any physical bedload samples collected during the multibeam campaigns to compare with the ranges of flux measurements?

No physical bedload samples were collected. We rather doubt the reliability of measurements from a bedload sampler lowered from a cableway suspended high above the water surface through 7m of the water column, into a field of dunes up to a meter high moving up to a meter per second.

Some discussion is warranted of whether the bedform bedload equation of Simons et al., is even geometrically appropriate in situations where bedform growth/decay is occurring. I don't believe they considered this in their original work, and I am not aware of any later publications that show the validity of this method for non-steady bedform fields.

Aside from the assumption that dunes are appropriately triangular, there is nothing in the Simons et al derivation that suggests it is not appropriate for application in time-discerete fashion such as here. That is to say, despite the growth and decay in dunes, the instantaneous bedform flux as predicted by their instantaneous geometry and celerity is appropriate and has been applied before to unsteady flows. While it is true that inferring wavelength from time-series of bed elevation measurements is made more difficult by spatially accelerating and decelerating dunes, the issue is a more in the implementation of the theory governing Simons et al, rather than the theory itself.

Along similar lines I would encourage the authors to consider incorporating, or at least explaining the inappropriateness for their application, the insights from Guala et al. (2014, their Section 4 paragraph 2 in particular; doi: 10.1002/2013JF002759) in joint averaging of the elevation and velocity values.
We agree the study of Guala et al is pertinent so we have added the following to the revised Discussion: "Guala et al. (2014) demonstrate a frequency dispersion in the relationship between dune celerity, Vc, and wavelength, λ, because small dunes tend to move faster than larger ones. This doesn't bias our computed bedload fluxes from multibeam data since we use time-series of bedform statistics from η(x,y,t), however it does place limits on any calculation of equivalent statistics from η(t) because it requires assuming a model that relates average Vc with average λ, or rather that the functional form between them doesn't vary in time, which may not be strictly true."

While somewhat outside the scope of the review of the paper itself, I should note that the license type given to the dataset and code hosted in the SEAD repository is potentially quite restrictive to some river management uses and researchers, given that it does not allow commercial use or any derivatives. This may be less important for the data itself, but it may heavily limit the use of this work to have code that cannot be modified. A share alike restriction, for example, would make this more accessible.
We have updated the license to allow for commercial use and derivatives.

Line Comments: The following line specific comments are non-critical to the science of the manuscript and are meant to help improve readability or clarity.
We have corrected the below Line Comments in the manuscript.

Page 1 Line 2: "remains elusive" is relatively non-concrete and feels dismissive of the wealth of literature and practice on field-scale bedload measurement techniques spanning half a century or more.
Page 1 Line 14: references are missing at "(e.g. ?)"
Page 1 Line 20: References such as Simons et al., 1965 and others don't explicitly derive from the exner equation, per se. They are derivations of mass conservation but not necessarily predicated on Exner's formulations.

Page 2 Line 1: Simons wasn't the first to show this, as written. For example, Bagnold 1941, Chapter 13 derives a similar formulation, albeit with some geometric inaccuracies. I suggest simply removing the word "first" from the sentence.

Page 2 Line 9: remove comma after ". . .discharge conditions,"

Page 2 Line 12: is there a reference for ". . .bedload flux estimated from translating dunes remains one of the most accurate. . ."?

This claim has previously been made by Wilbers and Ten Brink (2003) and Nittrouer et al. (2008) who said the measurements came with "… relatively high accuracy so long as the dune geometry and translation distances are large relative to the positioning error" which is the case here. The relative disadvantages of direct sampling are well documented (e.g. Holmes and Holmes, 2010), for example direct sampling disturbs the flow and therefore the rate of bedload, and the sampler must be placed squarely on the bed surface to adequately sample, therefore the presence of dunes causes error.

Page 4 Line 14: ISDOTTv2 is not a familiar/common tool since it is not public. If you wish to include this statement, it would be good to describe what that tool is and why it would be useful here. Otherwise I would suggest removing it.
We have removed this statement.

Page 4 Line 16: please describe the "missing triangles" correction.

Page 4 Line 22: consider rewording as it states a 1965 reference is based on a 2005 reference.

Page 7 Line 14: ". . .for growing (shrinking) dunes is 1.2 (0.75)." I suggest rewording to ". . .for growing and shrinking dunes is 1.2 and 0.75, respectively."

Figures: for figures 1, 2 and 4 there are abbreviations used which would be helpful to have defined in figure captions so the reader doesn't have to remember or find from the text. BEP, RMB, SB and MSB are all used. Additionally, Xcorr and RMSE are used but not defined in captions or in the text body.