RC1

*Line 43; You state that the system differs from other systems in two ways; towing design and fluxgate magnetometers.*

*I did find other systems using fluxgate; for example, Bartington instruments offer a solution (https://www.bartington.com/archaeology-forensics/), which also use the same magnetometers as you do.*

*Concerning the towing design, do you find your design superior to other system designs like the Bartington "Non-Magnetic Cart"?*

This sentence has been slightly rewritten. It differs from *most* other systems in this way. Indeed there are a few systems (Bartington, Ferrex) that utilize fluxgate and towing. However, these were not designed to be towed, but rather to be pushed by hand. The vehicle attachment is more ad hoc, and FAR too close to the vehicle. It is the separation of the electronics from the sensors as well as the distance of the towing vehicle that results in superior performance with respect to towing. I would not make the blanket statement that this system is 'better,' but has some advantages over the others.

*Line 46: Low noise compared to what?*

Good point; that sentence isn't particularly meaningful. It has been removed. The sensor noise is low compared to electronic and motion noise.

*Line 71: Cesium is only one kind of vapour used for optically pumped magnetimeters. You could use "alkali vapor" instead to include all types of optically pumped magnetometers using vapour.*

Agreed. The change has been made.

*Line 90: What is the argument for using the 1m fluxgate separation? I would expect archaeological structures to be relatively shallow targets.*

The reason is two-fold: the availability and cost of 1m gradiometers and the desire to not create our own separation due to the required precision of mounting two separate vector sensors, and second is that the system was initially designed to look for drain pipes with a particular signal strength, for which 1m was completely adequate. There could potentially be some benefit to reducing the separation for weaker, shallower targets, but this study has not been done. I added a comment to this effect in the discussion.

*Figure 1(a): both the IMU and the two antennas are relatively close to the magnetometers; this may cause some magnetic distribution. Furthermore, wires are passing close to the sensors. Since you don't mention it, I assume it's not an issue but have you tested/measured it?*

This is an issue brought up by all reviewers. The sensor positions were tested quite extensively, and are placed in as close to a null position as possible. We made the decision to not put the sensors on

poles due to engineering requirements, so their position is a tradeoff between noise suppression and construction limitations. The AC noise is aliased out and is essentially a non-issue. The DC noise due to DC currents and induced magnetic fields in the background are relatively small: less than 2 nT at our magnetic latitudes. This manifests as a DC signal in the sensors which is removed during the bias correction process. We note that even if the GPS units were mounted on poles, the power cable would still need to run up to the sensor, potentially producing a magnetic signature as well. A paragraph has been added in section 2.2.1.

*Figure5(a); It is hard to differentiate between the colours. It would be more informative if you came with an estimate of how low a degree you need.*

*-and-*

*Line 328: What accuracy and resolution of an IMU would be sufficient for your application?*

The precision of the IMU depends entirely on the requirements of the specific survey and which component you are looking at. To get the contribution to the vector components of the magnetic field to be approximately 1 nT, the angle needs to be known to approximately 0.01 degrees (with a variety of assumptions)—this is in the discussion. However, this value is for computing the vector components of each sample independently; the requirement is less strict if more advanced signal processing is done to estimate the vector projections during data reduction.

*Line 244: 2000nT at 6 degree.*

I am not sure what this comment means.

RC2

We thank the reviewer for insightful comments which have strengthened the paper. We have addressed each comment below.

*I think the paper would be strengthened at the outset by better describing and linking the application to the tool: why the minimum resolution specifications of 6 nT and 8 nT/m? These are rather high values, and without context it's difficult to imagine why they were chosen. The anomaly magnitudes are mentioned in section 4.1, but they should be briefly introduced earlier.*

Agreed. Some commentary that the 8nT/m noise requirement was based on the anomalies for which we were looking and while driving at 20kph has been added for context. Of course the system's noise floor is MUCH lower while stationary.

*The paper does suffer somewhat from a lack of clear application of the technology. The point of the paper is to describe the system, so it is understandable that a full description of applications (through processing and interpretation of anomalies) is not a major focus. None-the-less, there are important points that need to be clarified.*

1. *First, what is the range of usefulness of the system? Is 6 nT a reasonable standard for archaeological sites in general, or only some sites? Are there other applications for which this system is designed (e.g. infrastructure projects, geological applications)? Are there physical limitations (where it can collect data)? Etc. A short paragraph could address these questions.*
2. *Second, the extent of the downstream analytics consists of the total field anomaly and the total vertical gradient. One could get essentially the same results with a total field gradient magnetometer setup, and the precision would be much higher and with less issues of bias, temperature sensitivity, etc. (one could use an array of QuSpins or MFAMs, for example), so why use vector magnetometers if the results aren't being used? If it is because of the steady march towards full tensor gradiometry, that's fine, and you do state this several times throughout the paper, but perhaps emphasize it when discussing the case studies.*

This comment is appreciated. We have added a paragraph in the discussion, highlighting where the noise floor is most relevant and the physical limitations of the system, e.g. turning radius. Secondly, to get total field and TVG, you absolutely could use a set of total field sensors if desired. However, the greater expense, and to my experience power drain (though the reviewer has pointed out this may no longer be the case) of total field sensors is an issue. More importantly, we see this system as a good step towards full-tensor gradiometry, and a physical demonstration of the requirements for an IMU system or other way of measuring attitude in order to achieve this full-tensor goal.

I was unable to find information on OuSpins, but it seems that the MFAMs are not on the market yet, at least according to the Geometrics website which was updated at an unknown date. Nevertheless, they will be an exciting addition to the toolbox.

*The GPS antennas seem to be place too close to the fluxgates; I would expect noise from them. If you've tested the noise envelope for the antennas (changing orientation at all four compass points, pullaway tests, etc.) briefly include this information (or cite a previous white paper etc.). If you haven't tested this, I would strongly advise it.*

This issue was brought up by multiple reviewers so I have duplicated the comments here. The sensor positions were tested quite extensively, and are placed in as close to a null position as possible. We made the decision to not put the sensors on poles due to engineering requirements, so their position is a tradeoff between noise suppression and construction limitations. The AC noise is aliased out and is essentially a non-issue. The DC noise due to DC currents and induced magnetic fields in the background are relatively small: less than 2 nT at our magnetic latitudes. This manifests as a DC signal in the sensors which is removed during the bias correction process. We note that even if the GPS units were mounted on poles, the power cable would still need to run up to the sensor, potentially producing a magnetic signature as well. A paragraph has been added in section 2.2.1.

*It was not immediately clear that the gradient system used by the authors is a commercial system. This is significant and should be made clear, because a commercial system is plug and play and has published specs. This lowers the bar for others who might be interested in building a system, knowing they don't have to build their own gradient system in addition to everything else. As written, the interpretation could be (and is what I initially assumed) that the individual sensors were purchased*

*and a custom gradient system created in-house. Maybe add a picture of the Barington system, showing where the sensors are located?*

It is stated in section 2.1 which sensors are used, but a note that they are commercially available has been added in the introduction. The sensors are visible in Figure 1a; a sentence has been added to section 2.1 calling out how they are positioned.

The sensors, however, are not exactly plug and play. You can purchase a logging system from Bartington which is essentially ready for use, at the expense of flexibility, and if memory serves, will not accept 8 sensor inputs. One step removed from their logging system is to use their available API, which simplifies communication. However, the problem remains of how to record 8 RS485 streams at 230 Hz. This can be solved with additional hardware. More significantly, there is an issue with the timing of these sensors: essentially the data streams are timestamped with an integer millisecond value relative to power-up of the individual instrument. However, there is no guarantee that the timing interval is an integer value depending on the chosen sample rate, and the system firmware reports the time as (rounded sampling rate) x (sample number), which causes extreme drift in timing. As a consequence, we dispense entirely with the API and communicate with the sensors directly to allow us to time stamp the data streams ourselves. Even more issues arose as the data packets sometimes were returned out of order, so significant effort was spent in achieving proper timing of each sensor. This problem is of course specific to this sensor generation, so other instruments would not suffer from this issue. At the end of the day, though, getting 8 sensors to communicate with a PC with proper timing is certainly possible, but non-trivial.

*Fluxgates can be quite temperature sensitive. Did you test the temperature sensitivity? I didn't see any discussion of this in the Barington grad-13 manual. It's worth mentioning whether you or Barington has tested this – provide an opinion on the instrument's temperature stability.*

Past the initial warm-up period, the sensors were steady with respect to temperature. However, we fully recognize that the temperature in Denmark is usually quite stable during the day given the humidity. We added a short comment, though our observations should not be treated as statistically significant.

*Section 2.4*

1. *Please do cite Reid's 1980 Geophysics short note paper, which can be found here Aeromagnetic survey design: SHORT NOTE (reid-geophys.co.uk)*

Rather than in 2.4, we have added the reference in 2.2.1.

*Section 3*

1. *The authors seem to have developed an allergy to Fourier-domain modeling – I don't think it's the problem that they do, and it is a tried and true way of processing signals. True, you need to buffer your area of interest because there are edge effects, but one can plan for that. While you don't need to work in the fourier domain, one benefit is less noisy derivatives, which might be significant if gradients are important. Noise*

*management for calculating the derivatives (e.g. total gradient) should be briefly explained if Fourier methods are not being used.*

2. *The suggested workflow needs to be modified with caution – upward continuation (which is suggested) should not be used on a line by line basis, because in 1D upward continuation assumes no out of plane changes, which is not the case for the archaeology surveys. The data would need to be interpolated to 2D before applying Fourier filters.*

*achoo* In all seriousness, one can absolutely construct gradients, apply filtering, or whatever else in the Fourier domain, and we stress that *it is a totally appropriate and fine thing to do*. We have chosen to minimize Fourier operations to avoid the need for interpolation or padding. Again, these can be done—in fact the first iteration of the processing software was replete with Fourier domain processing. These were later removed as the particular signals we were looking for did not require those operations, leaving us with a minimal workflow. We again stress that the workflow here is the minimum viable—the user can add in any and all processing techniques as appropriate. Some commentary has been added to the beginning of section 3 to that effect.

Regarding the calculation of the total gradient, we note that it is not the total gradient (I assume the reviewer means the proper way of saying the commonly-misused analytic signal in 2D) shown in the paper, but the total vertical gradient ($\partial|\mathbf{B}|/\partial z$), which is much less susceptible. To calculate the total gradient or any of the gradient components, a Fourier method is appropriate, or a Lanczos differentiation to keep in the spatial domain. These are discussed in the discussion section in the context of calculating the gradient components.

Regarding #2: the authors 100% agree. While one could construct an upward continuation in the spatial domain using all lines, there is absolutely no good reason to do so. The absolute raw data are available to the user, so there is essentially complete freedom to implement whatever processing desired.

*Section 3.1*

1. *Fluxgates can be very sensitive to temperature. They can also drift throughout the day, like a gravimeter does (this could be related to temperature fluctuations). Add a paragraph noting how you have addressed any hourly drift that may be associated with the fluxgate system (or that you have measured it and it doesn't appear to be a significant problem).*
2. *Note that at 20 ms data stream integration, at 20 kph and 230 Hz collection rate, that's 5 samples per distinguishable location, so that effectively limits the sampling rate of the mag. Still should not be a problem in the spatial domain, 20 kph is 5 m/s, so even if there are only effectively 60 samples per second, that's ample samples for 5 m along a line/*

Number 1 was addressed in an earlier comment. Regarding 2, there is no question this is highly oversampled. Position is effectively continuous in the first derivative (acceleration is discontinuous of course) so one could argue either way whether we establish datapoints at positions between 'distinguishable locations.' The higher sampling can give us better statistics on our noise, rather than trusting an averaged datapoint from the sensor, can improve the powerline harmonic removal, and if Fourier operations are to be implemented in

the time rather than spatial domain, improve the spectral resolution. Either way, whether 230 or 60 Hz, the data along line are still oversampled relative to the 50cm line spacing.

*First, the bias correction; the concern here is that different lines have different signals, even though they may be close in space, e.g. one sensor passes over a very local source (a bolt that dropped out of a tractor 20 years ago, say). You want to keep the true signal (bolt) and eliminate the instrument bias (eventually you want to get rid of the bolt, of course, but not at this stage because it's part of the "true" signal). In this case longer line segments would be better than moving windows, because it allows greater opportunity for these sorts of local sources to "even out". The effect of these local sources is greater the smaller the window, and you want to minimize their effect. And if outliers are a problem, I would first try removing the median – it will be robust to outliers. I would try an experiment at some point – compare your method with simply removing the global median from each line, and see how the results differ. I think of the problem this way: for a given line measurement, M, there is signal S and bias B, M=B+S. The signal can further be broken down into its median and signal, Sm+S0. For two lines, M1=B1+Sm1+S01 and M2=B2+Sm2+S02, the true comparison is Sm1+S01 to Sm2+S02. If the difference in the signal medians is small (Sm1 - Sm2 ~= 0), as might be expected from lines that are not very far apart (I think you'd need some strong regional gradients for this not to be true), then subtracting a robust measure of central tendency (median is a pretty good one) should get you close to comparing the signals you want to compare, without the need for iteration or removal of the longer wavelengths.*

*Second, the long-wavelength removal; this is not necessary, and not desirable, if you can remove the bias without removing the longer wavelengths. Keeping the longer wavelengths preserves more of the "true" signal (from subsurface sources), and allows subsequent analysts the option of keeping it or removing it. Why filter out signal if it is unnecessary? Let the analyst do that (or not) as they see fit.*

Longer line segments are indeed desired for the reasons stated. We experimented with a variety of lengths, settling on half-overlapping 30m windows. This is large enough to avoid the example given with the bolt, but shorter enough to utilize multiple independent bias estimates. Interestingly, if the iterative approach is redefined to eliminate a single data point at each iteration, the solution converges to the median of each window.

The windowed approach is necessary as lines are not always perfectly straight, so the apparent bias can be changing over the course of the line. We experimented with mean and median removal across the whole line, and found better performance with the windowed approach. Using a polynomial across windowed bias estimates provides a good tradeoff of a robust bias removal at the expense of long wavelengths. Given that the bias is removed on a line-by-line basis, these wavelengths are lost across lines anyway.

The need to compute this bias highlights the main issue with fluxgate measurements. If we could easily flip the system upside down, we could directly estimate the bias and obviate the need for this process. Unfortunately the bias can change with proximity to large magnetic sources so we would need a field procedure to directly do this, which is of course not feasible.

A global method of bias correction, calculating biases across lines, would improve the estimates and help keep longer wavelength signal, but has not been explored in detail. Given the design requirements that the system be sensitive to isolated targets in the near-surface, we chose to allow the loss of longer wavelength signal. Given time (or more precisely funding) we would like to

explore that option—there is no question that we would like to keep those signals in general if possible.

*Nice to have the intro paragraph giving some background on the archaeology. Figure 6 would benefit from identifying some of the anomalies seen on the map e.g. suspected iron forge, suspected building, etc.*

Some description has been added, taking care not to overstep my interpretative ability as a geophysicist and not an archaeologist.

1. *It would be helpful to list the magnitude and perhaps shape of the anticipated anomalies. Presumably they are > 6 nT and produce a gradient > 8 nT/m…?*
2. *Would also be helpful to estimate how long this would have taken to perform a total field walking survey, for comparison. I imagine the time improvements, as well as data density, would be significant, and would make your point about efficiency abundantly clear.*
3. *Would be helpful to know if you found anything archaeologically worthy. Also, why is there a magnetic contrast in the sediments such that the permafrost is outlined? This seems strange. I realize this is not the point of the paper, but the examples stands out as odd without further explanation. Could the permafrost be creating rough terrain, such that the instrument changes elevation or bumps, and this is propagated into the data?*

Frankly, we don't know the expected anomaly magnitudes. Given the breadth of potential sources, it could be anything from below detectable to hundreds of nT.

We added a comparison in Ørregård, since another group had actually performed a walking survey over the area. We acquired much more data than shown, and the comparison was a half day vs nearly 3 weeks.

Some more information has been added regarding Aggersborg. We did not find anything conclusive, though we did identify a few potential areas for excavation.

More information has been added to the interpretation. There is no surficial expression (topographic or otherwise) for these polygons. They are periglacial features in soft sediment resulting from permafrost polygons. In each of the 'cracks,' an ice wedge had formed, and later filled in with transported sediment as the ice melted.

Further changes were made to the manuscript based on the supplement.