Author response to referee #1

The authors thank Anonymous referee #1 for their detailed comments on the manuscript. We will take the comments into account when revising the manuscript. In this document we provide responses to each of the referee's comments (formatted as italics in indented paragraphs and grouped together when appropriate).

> *Line 10: The error comparison for different regions is as such not useful in an abstract. It does not compare to achievable quality in the regular case (all axes working) and it does not relate the error to the measured fields (percentual) – e.g. a deviation of 2 nT in solar wind has a different impact as 12 nT in the high field magnetosphere. Provide more information or less.*

We agree and will remove that part from the abstract and expand the discussion section in the revised manuscript accordingly.

> *Line 25: Please state why this deviation occurs (mechanical alignment/boom etc.)*

According to Auster et al. [2008], there is no alignment requirement between the instrument and the spin axis. Fig. 1 shows the evolution of the angles α_{S1} and α_{S2} throughout the mission, demonstrating that they consistently deviate from 90°. We will add a short explanation in the revised manuscript.

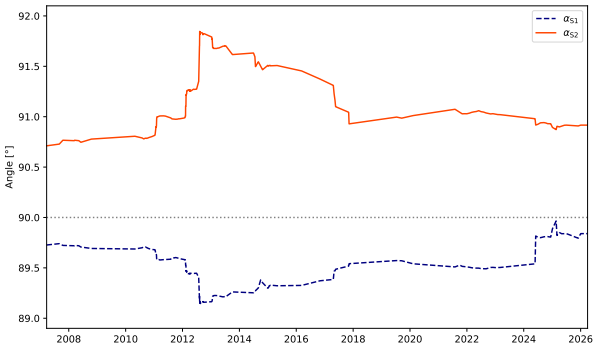

Figure 1: Angles α_{S1} and α_{S2} between the spin plane sensors and the spin axis over the course of the mission.

> *Line 50:*
>
> - *The assumption for accurate spin plane offsets is unclear. The calculation of spin plane offsets requires precise elevation angles, otherwise a constant contribution of the spin axis field (to be expected over many field regions) could be mistaken for spin axis offsets. Precise elevation angles in turn require knowledge of the spin axis field, but this field is in turn recovered by the method below with considerable error.*
> - *There is a danger of a circular dependency, so a simple assumption should not be used without analysis (or presented without justification, if the analysis was done, but is not included)*
> - *It is clear that there will be an error and that the situation is not perfect. Nevertheless, the assumption without further information is not enough. An IGRF-comparison could provide an estimate (i.e. rather than correlating spin axis variations to spin plane variations, one could correlate IGRF variations to spin plane to find thetas.*
> - *It might be that the variations in theta are small enough to be irrelevant, but this should be brought to support the initial claim of spin plane offsets.*

We agree that there is the danger of circular dependency and that this should not be phrased as it is in the current version. In lines 230 to 243 in the section Discussion of Limitations we address this point, although without analysis. First, we must emphasize that T89+IGRF (or another model field) is

required to obtain an estimate of the spin plane offsets. To demonstrate that these estimates are reliable, we will include a comparison between the spin plane offsets calculated using the standard calibration procedure and those calculated using the model field as a spin axis sensor measurement.

> *Line 51: The angle alpha is not very clear for the outside world. A drawing would help.*
> *Line 132 and 140: The factor alpha has not been explained well, see above.*

We agree and have therefore expanded Fig. 1 by adding a third panel illustrating the angles $\alpha_{S1,S2}$ between spin axis and spin plane sensor axes $S1$ and $S2$ (see Fig. 2).

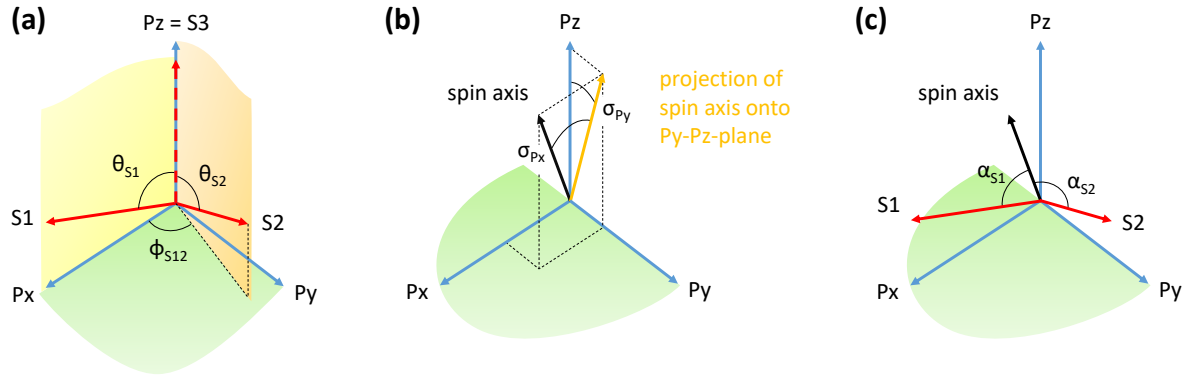

Figure 2: Updated sketch of sensor coordinate system. The panel c) shows the angles α_{S1} and α_{S1} between spin axis and spin plane sensors S1 and S2, respectively.

> *Line 53: The tradeoff is unclear. 2017 is not "close". This needs more precise justification. No big problem, just needs to be done.*

We agree that a justification is necessary. Since 2021, we have occasionally observed disturbances in the magnetic field data that fall outside the scientifically meaningful range but within the calibration range. To avoid these anomalies, we limit our analysis to data collected before 2021 and have arbitrarily selected the year 2017. Examining the evolution of the calibration parameter (see Fig. 3) suggests that 2018 or 2019 might provide better results with more constant calibration parameters.
However, redoing the analysis with data from 2017 to 2020 reveals no significant deviations among these years (see Fig. 9, Fig. 10 and 11 at the end of the manuscript). We therefore use all data from 2017 to 2020. In the revised manuscript we will justify the chosen time range and redo the analysis with data from 2017 to 2020.

> *Line 85: Provide a citation for constant sensor-internal orthogonality. The statement is true enough, but should be backed up. Past calibration values could help.*

We agree and therefore calculated the angles between the sensor axes θ_{S1}, θ_{S1}, and ϕ_{S12}. In Fig. 3 we present the evolution of the calibration parameter over the whole mission time. All parameter, beside σ_{px} and σ_{py}, are very stable over the duration of the mission. σ_{px}, σ_{py}, and ϕ_α are the angles that change when the spin axis direction changes. We will add Fig. 3 to the revised manuscript.

> *Line 95:*
>
> - *The choice of the wording "spin tone" is very unfortunate. It probably originates from the fact that the contribution of BST would result in a spin tone if despun.*
> - *What is considered here is the "true" magnetic field in the spin axis (+ errors) that is amplified and used as spin axis field. Even equation 11 is considering a "spinning" part (A+B) and a non-spinning part (C+D), but the latter is then referred to as spin tone.*
> - *In that sense the word "spin tone" is the name of what A+B if they were demodulated, but they are not and used for something else.*

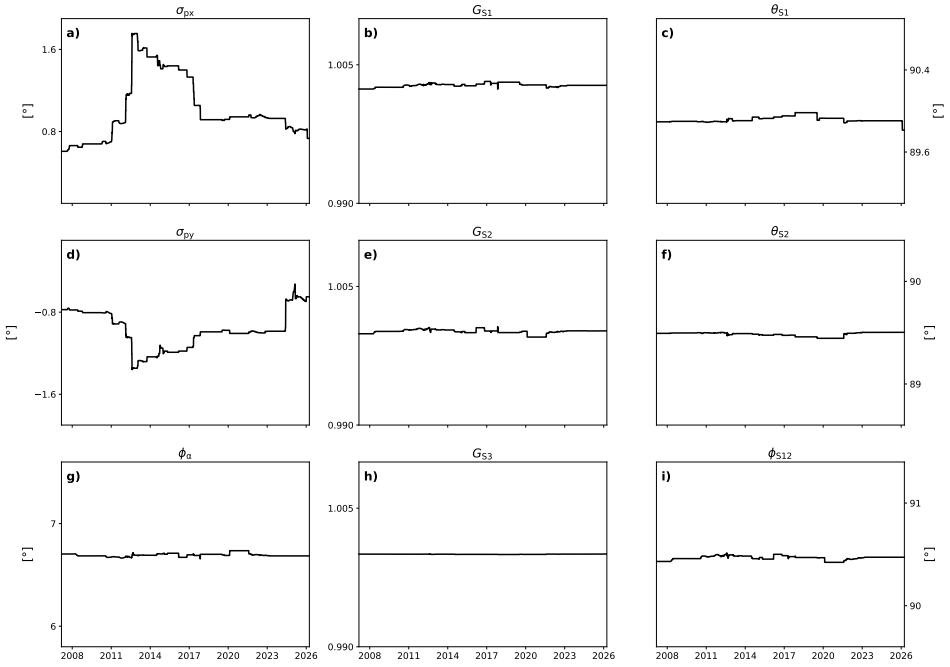

Figure 3: Evolution of calibration parameter: σ_{px}, σ_{py}, and ϕ_{α} are the rotation angles around the orthogonalized sensor axes (c.f. with blue coordinate system Fig. 1 in the manuscript or Fig. 2 in this reponse). G_{S1}, G_{S2}, and G_{S3} are the gains along the sensor axes. θ_{S1}, θ_{S1}, and ϕ_{S12} are the internal sensor angles as described in Fig. 1 in the manuscript.

- *(My personal experience was that I first skipped the index ST, just had a look at the equation and things were quite clear. After that I checked what ST means and got confused and started to look for what I have misread, which was nothing)*
- *Please find a new term to make this more clear.*

We agree that the wording is unfortunate. We will replacing the wording "spin tone" with "projected spin axis contribution" and rename the variable from B_{ST} to B_{pSA}. This should make things clearer.

Line 100:

- *The fit is not with respect to Braw – Braw is the desired result. The fit is done with respect to A/B/C/D*
- *It is also not clear whether the parameters ws and phi are part of the variable parameters. They probably are, with some limitations on prior knowledge.*

We agree and will change the wording accordingly. We will also clearly indicate that A, B, C, D, and φ are free fitting parameter, whereas ω_{s} is provided by the state file of THE.

Line 104:

- *Phi is not the phase; it is the angle of the external field that is projected onto the spin plane. This appears as a phase shift in the equation, but does not correspond to the usual term of phase shift, i.e. a fixed time or angle shift. It is a variable, data dependent shift. Please change wording to avoid confusion*
- *Line 104 uses phis, while eq. 11 uses only phi. Also the use of omega/omegas is inconsistent across equation and text.*

We agree and will rewrite the sentence. We will also fix the issue of inconsistency.

Line 119: The method for averaging was chosen in time units rather than samples. The impact and justification for this was not given. Samples would generate overcomplete/incomplete spins, while time generates a variable number of samples. The choice was not explained, but will have some (minor) impact.

We agree and compared the different methods. First, we used fixed time windows and disregarded any window containing less than 90% or more than 110% of the expected $2.5 \cdot T_{\mathrm{spin}} \cdot f_{\mathrm{FGL}}$ (≈ 166) samples. Second, we calculated the fits with windows that contain a fixed number of samples and rejected any window where the time difference between the first and last point exceeded $2.55 \cdot T_{\mathrm{spin}}$. The differences of the recovered values of the two methods in Fig. 4a) and Fig. 4b) shows that the outcome of the fitting process is very similar (3 years of data from 2017 to 2020: Median around 0 nT and around 1 nT in width). Fig. 4c) illustrates that the number of samples per window in time units matches the constant sample number quite well, although there are some intervals with a few more points when using a constant time window.

In the revised manuscript we will continue using the constant time window approach and discuss the

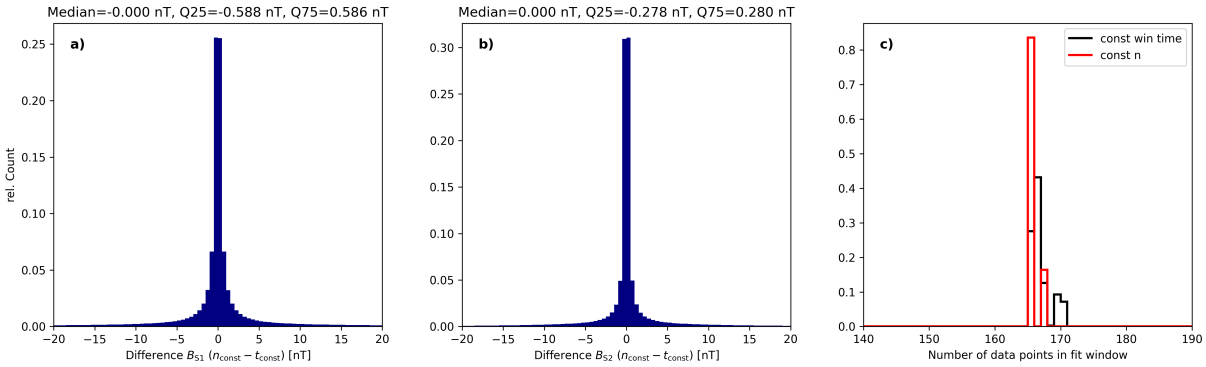

Figure 4: Comparison of recovered spin axis values for both spin plane sensor components B_{S1} (a) and B_{S2} (b) between fits with a fixed number of samples and fits with fixed window length. Panel c) shows the distribution of samples per fit window for fixed numbers in red and for fixed window length in black.

minor discrepancy compared to using a fixed number of samples.

Method in 3.1

The method used in chapter 3.1 is not well explained. The authors are using the following measures to improve the quality of their result:

- *Assume a first order least square fit for the magnetic field and considering everything else as error signal.*
- *Using the fit onto 2.5 spin periods, shifting by 1 spin period (overlap 40%).*
- *Using a weighted moving average over 2 minutes. Weight increases with constant part of the spin plane component and decreases with the error signal (i.e. true error and higher frequency contributions).*

Nevertheless, the formulation of equation Eq. 13 is not clear. BST is defined as a function of t in Eq. 11. Equation 13 also uses BST(t) but it is obvious that this is not the same definition – Here BST is averaged over 2 minutes, while it is only defined for 2.5 spins in Eq. 11.
*I assume that the authors average the central values of the respective length-2.5 periods (i.e BST(1.25*tspin)), but this is not stated anywhere. Please clarify.*
Assuming equal weights (will be discussed later), the application of a 2-minute weighted average would result in a bandwidth reduction 1/120 Hz, i.e. even though one averaged value per spin is provided, the included bandwidth is much lower. While this corresponds to the abstract statement of providing one estimate per spin, this is still generating the wrong impression of a higher bandwidth. This fact is only stated at the very end (line 264). Nevertheless, it seems that also the authors do not consider this reduction of bandwidth, see comment to Figure 5.

We agree that we have to explain the method in more detail. As correctly pointed out, we used t in Eq. 11 and Eq. 13 although they are not the same. Therefore we will replace t with τ in Eq. 11 confined

from 0 to $2.5T_{\mathrm{spin}}$. In a similar manner we will change Eq. 12. The referee is also right that we are using the central average $B_{\mathrm{pSA}} = B_{\mathrm{pSA}}(\tau = 1.25T_{\mathrm{spin}})$. We will expand the description of the method and will fix the aforementioned inconsistencies. Regarding the bandwidth reduction: We agree that we have to communicate this more clearly. We will add the remark in the abstract, in the method section and in the discussion.

The other thing is the topic of weighted averages.

- *The authors use a weight that decreases with the error signal. The error signal is the residual between fit and measured signal and contains (a) components with higher frequencies and (b) disturbance/noise and (c) low frequency signals that are not covered by the model of eq. 11 – the latter is the actual fit error the authors are looking for. The presence of higher frequencies is not an error, it is by choice of the bandwidth. The attempt to extract one spin "model" over 2.5 spins means that the actual bandwidth that is relevant for the fit is clearly less than the raw Nyquist of 2 Hz in FGL. As a result, the resulting fit quality estimate fmin is not only giving the quality of the fit, but also the amount of higher-frequency contribution. The latter is not really an error, it is just an artificial residual that is generated by the choice of bandwidth. It would be more consequent (and it would most probably deliver better results) to reduce bandwidth first and remove the higher-frequency components. This is also justified by the fact that later averaging is massively reducing this bandwidth anyway, so there is no loss as long as the upper bandwidth limit is sufficiently above fspin.*

We agree that f_{min} contains more than the actual error and appreciate the suggestion to use a low-pass filter before fitting. We applied a zero-phase low-pass filter with different cutoff frequencies. In Fig. 5 we present exemplary raw data and low-pass filtered data with different cutoff frequencies.

We selected to continue with a cutoff frequency of $2f_{\mathrm{spin}}$ as this is still resembling the amplitude

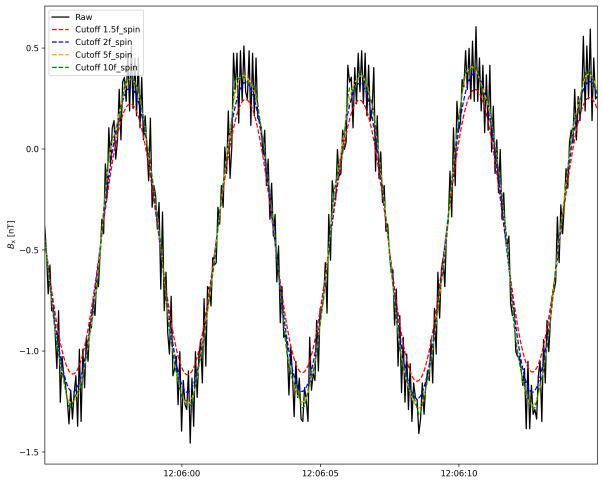

Figure 5: Comparison of raw data (black) and low-pass filtered data with different cutoff frequencies ($1.5f_{\mathrm{spin}}$ in red, $2f_{\mathrm{spin}}$ in blue, $5f_{\mathrm{spin}}$ in yellow and $10f_{\mathrm{spin}}$ in green) at 2017-01-15.

of the raw signal (contrary to $1.5f_{\mathrm{spin}}$), while ignoring the higher frequent oscillations (contrary to $f_{\mathrm{cutoff}} > 5f_{\mathrm{spin}}$). When we compare the deviation from the original measurements, reducing the bandwidth with the low-pass filter is improving the result a little bit but does not affect the end result a lot. This is visible in Fig. 7 and Fig. 8 when we compare the widths of the distributions (more discussion and explanation in the next answer). In the revised manuscript we will add the low-pass filtering as a new step to the implementation.

- *The other topic is the weighting with the value of A, i.e. the constant spin plane field. While of course such a field could increase some errors, it might also be correlated with the spin axis field – i.e. high spin plane and axis fields come at the same time. While the first degrades results, the second is actually improving results. IMO the weights should be*

revisited. In addition there might also be a correlation between field magnitude and field variation, i.e. fmin. This would require analysis before specifying an error criterion. One could also discuss the justification of this scaling. The quality measure would only increase the weight of high field regions in the case there are also low field regions within the same 1-minute interval. In all other cases, the field is (typically) consistently either high or low over 1 minutes, so the scaling will not have any effect.

- *Generally, the application of such an error criterion over such a long time (1 minute or 15-ish spins) is not advisable. Generally, as also done in most filtering applications, the weight of a sub-estimate should be reduced with the time distance to the currently averaged point. While its quality of a point with high time distance may be higher, its validity for the current time is typically lower. The used approach here could (theoretically) mean that a single spin with high field could dominate a complete 1-minute interval. There are 2 options out of this time problem:*
  - *Scale with time distance and/or*
  - *Decrease output cadence to 1 minute instead of Tspin (even then additional distance scaling might be appropriate)*

We have revised the weighting and agree that the factor A^2 has no significant influence and does not change substantially within 1-minute intervals. We used three different weighting schemes ($A^2/f_{\min}$, $1/f_{\min}$, and $1/f_{\min}$, as well as a scaling with the time distance) and compared them with one another. The results show that using $A^2/f_{\min}$ or $1/f_{\min}$ lead to very similar results (see Fig. 6, the blue histogram). Adding the time scaling changes the results only slightly more (see Fig. 6, the orange and black histograms).

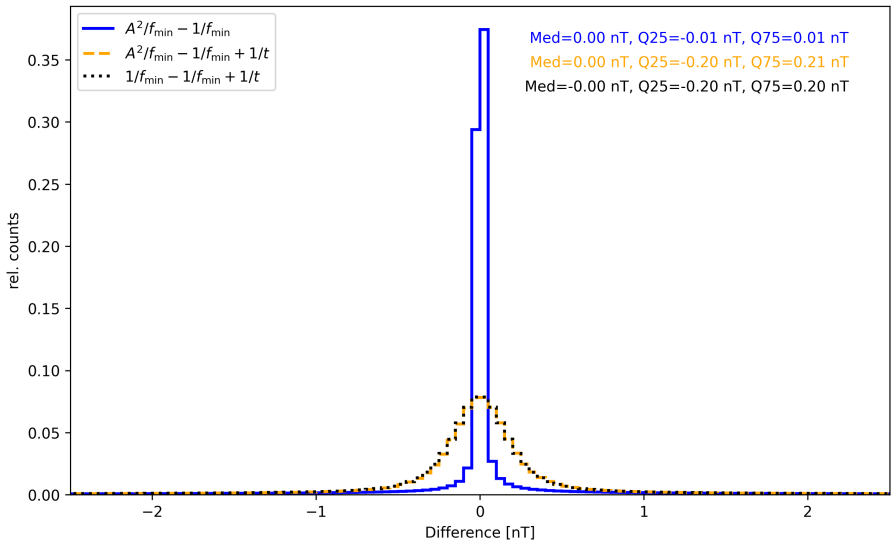

Figure 6: Comparison of different weights and averages of recovered data from 2017 to 2020. The blue line compares the weights $A^2/f_{\min}$ with $1/f_{\min}$, the dashed orange line $A^2/f_{\min}$ with $1/f_{\min}$ and additional time distance scaling, and the black dotted line $1/f_{\min}$ with $1/f_{\min}$ and additional time distance scaling.

Although when we compare the averaged FGS original data (see comment to reduction of bandwidth of original measurements) with the recovered data using $A^2/f_{\min}$ or $1/f_{\min}$ and the scaling with time distance, the comparison histograms dont differ too strongly (see Fig. 7 and 8). One can argue that the option with time scaling behaves a bit better due to lower interquantile ranges.

Line 150: The impact of alphaS1 and S2 is relevant and the method is correct, but there is no error estimate and no method is presented.

We agree and will expand the paragraph. We can get an estimate for $\Delta\alpha$ by comparing our calculated α_{fit} values with the values directly from the calibration file α_{cal} from previous years when we could still

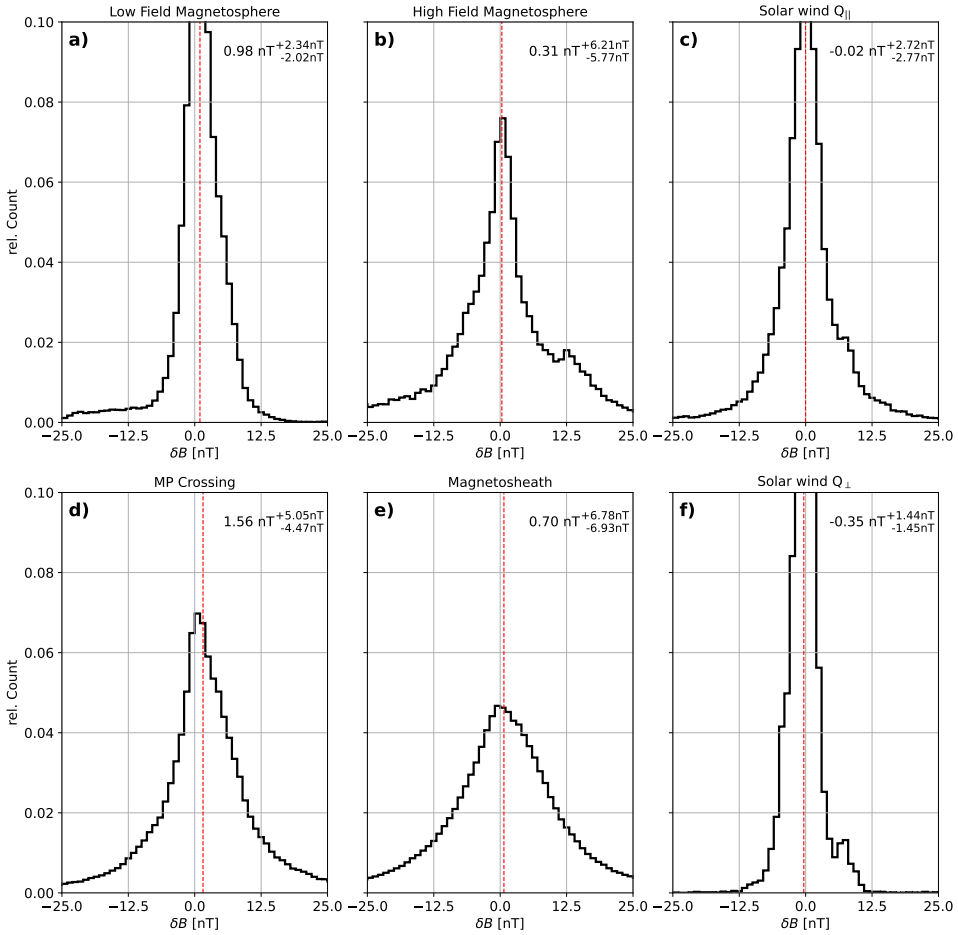

Figure 7: Histograms of the deviation δB between recovered (B_{r}) and original (B_{FGS}) spin-axis measurements in six different regions. Here we used A^2/f_{min} as weight and applied a simple moving average to both recovered and original data.

use the routine calibration methods. We are using the median difference between these two from 2017 to 2020 (we used these years, since they are close to current time and before the first signs of degradation appeared, as explained above):

$$\Delta\alpha \approx \mathrm{median}\left(|\alpha_{\mathrm{cal}} - \alpha_{\mathrm{fit}}|\right). \tag{1}$$

This results in uncertainties of $\Delta\alpha_{\mathrm{S1}} = 0.045°$ and $\Delta\alpha_{\mathrm{S2}} = 0.058°$.

Line 159:

- *There is no real error/sensitivity analysis. The impact of having wrong calibration parameters is not discussed. The impact of amplifying the tiny spin axis part within the spin plane with a value from several tens to 100 (e.g. for elevation=1°) is not discussed – this would e.g. also scale sensor noise and have a relevant impact on the expectable quality of the result. It cannot become any better than that.*

- *As a result, the authors only present an empirical error of the given data set, which is an upper bound for a given data set. Nevertheless, it is not known if this error is different for other data sets and what the sensitivity of these bounds relative to other calibration parameters is. I am aware that not all of this should be written in the paper, but there is some room to improve.*

We agree that we should expand the discussion about uncertainties. First, we will address the calibration limits caused by sensor noise. As the reviewer correctly noted, sensor noise is also scaled by $1/\cos\alpha (\approx 50)$, which is why we will derive the lower limit of our recovery uncertainty due to the sensor noise. In addition, we calculated the influence on α of the variations in all calibration parameters. To

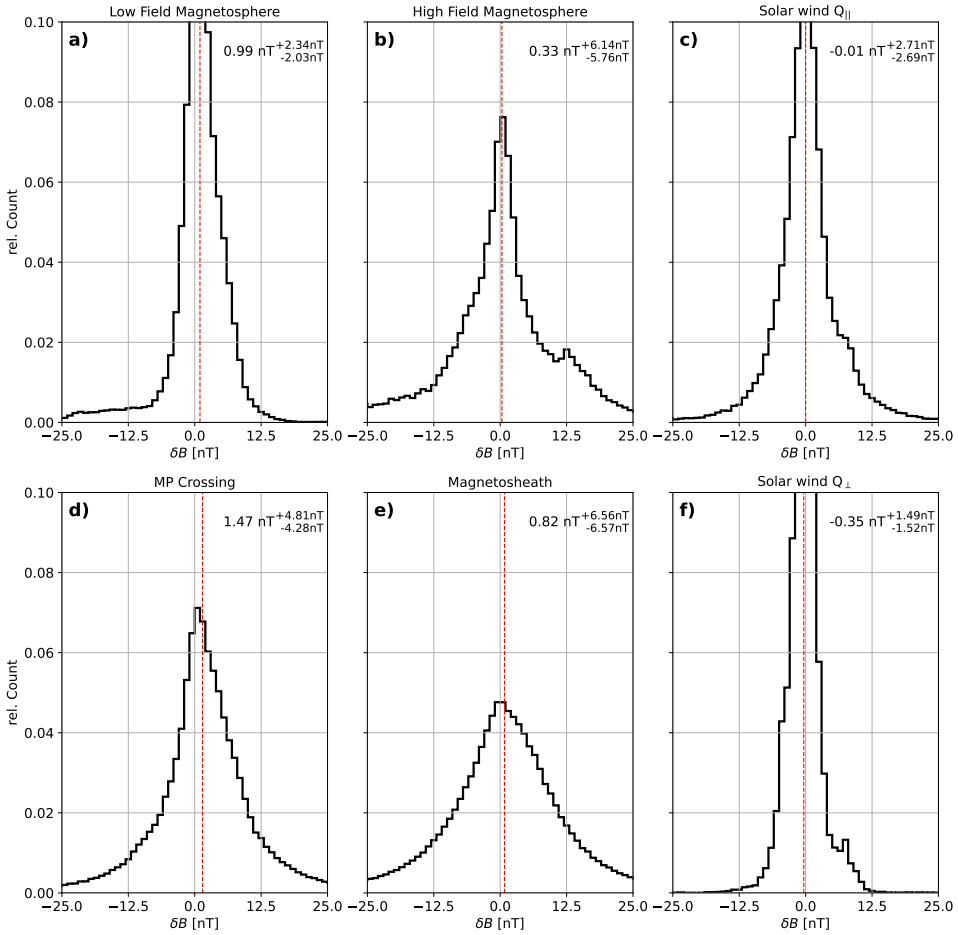

Figure 8: Histograms of the deviation δB between recovered (B_r) and original (B_FGS) spin-axis measurements in six different regions. Here we used $1/f_\mathrm{min}$ as weight and applied a moving average scaled with time distance to both recovered and original data.

do this, we determined the median as well as the 25th and 75th percentiles of all 9 parameters over the entire mission duration. We then calculated the deviations $\Delta\alpha_\mathrm{S1}$ and $\Delta\alpha_\mathrm{S2}$ between the angles calculated using the median values and those calculated using the 25th and 75th percentile values. The largest deviations resulted from the deviations in σ_px and σ_py (which is clear when looking at Fig. 3) and resulted in deviations of $\Delta\alpha_\mathrm{S1} = 0.015°$ and $\Delta\alpha_\mathrm{S2} = 0.018°$.

Figure 3:

- *The chosen time + amplitude scales of the plot are not suitable to actually demonstrate the quality of the recovery. An overview plot is great, but detail plots would be needed.*
- *In addition, the choice of unfiltered FGS vs. the recovered signal suggests differences that are in fact caused by the higher bandwidth of FGS – it would be necessary to reduce the bandwidth of FGS to the one of Br before doing this comparison. This would most probably show better results.*
- *The same applies to figure 4*

Figure 4:

- *It might be worth to reduce bandwidth and to use time lags for comparing the 2 spacecraft*

We agree that comparing the recovered data with FGS data of different bandwidth is not suitable. We will reduce the bandwidth of FGS data for THE and THA by averaging in the same manner as described above for the recovered data. In addition, we will apply time lags for the THA comparison and we will

add more panels with details to Fig. 3 and Fig. 4 in the revised manuscript.

Figure 5:

- *Also this figure has the problem of different bandwidth between Br and FGS. The shown errors are a mixture of actual error and high frequency content. With the general inverse power laws, the existence of higher frequency signals with lower amplitude is more probable and so the error distribution is potentially highly related to the high frequency components rather than the actual error.*
- *This plot has identical scaling on the x axis. This choice is valid, but it is not really giving a quality estimate for the residual – e.g. in solar wind the error will be small per se, as there is not a lot of signal anyway. Even an opposite signal might result in an error of 4 nT.*

In addition, the SNR of the tiny bit of spin axis field in the plane axes is probably quite bad, it is probably not that much above the noise. It might be worth to add some "SNR limit line" that helps to differentiate between errors in recovery and the limitations of noise that is just unavoidable.

We agree that we also need to reduce the bandwidth for this comparison. We like the idea of adding a SNR limit line and will revise the plot accordingly. In addition, we will discuss not only the absolute, but also relative errors in the revised manuscript.

Line 283:

- *This statement should not be given without proper analysis (e.g. impact on currently used calibration method, errors and parameter accuracy). One example is that every axis offset would require a mixture of Hedgecock and spin calibration output – but those require different external magnetic field conditions and are therefore executed on different data sets and times.*
- *Suggestion for a change: "The impact of such an intentional tilt would require detailed analysis though, as it might impact the sensitivity and errors of calibration during normal operation scenarios."*

We agree and appreciate the suggestion. We will revise the statement.

1 Comparison of Recovery for Different Years

References

Auster, H., Glassmeier, K., Magnes, W., andW. Baumjohann, O. A., Constantinescu, D., Fischer, D., Fornacon, K., Georgescu, E., Harvey, P., Hillenmaier, O., Kroth, R., Ludlam, M., Narita, Y., Nakamura, R., Okrafka, K., Plaschke, F., Richter, I., Schwarzl, H., Stoll, B., Valavanoglou, A., and Wiedemann, M. (2008). The themis fluxgate magnetometer. *Space Sci. Rev.*, 141:235–264.

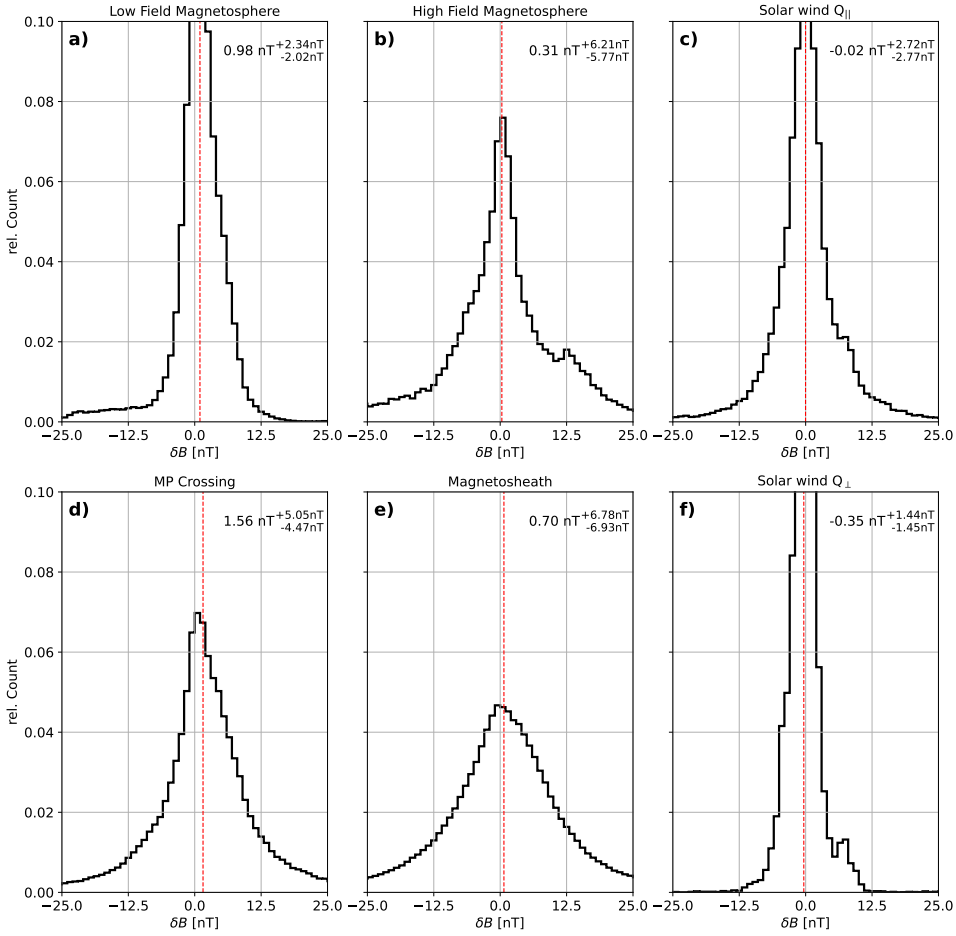

Figure 9: Histograms of the deviation δB between recovered (B_{r}) and original (B_{FGS}) spin-axis measurements in six different regions for the year 2017. Here we use A^2/f_{min} as weight and applied a simple moving average to both recovered and original data (a more detailed discussion of various weighting schemes will follow later).

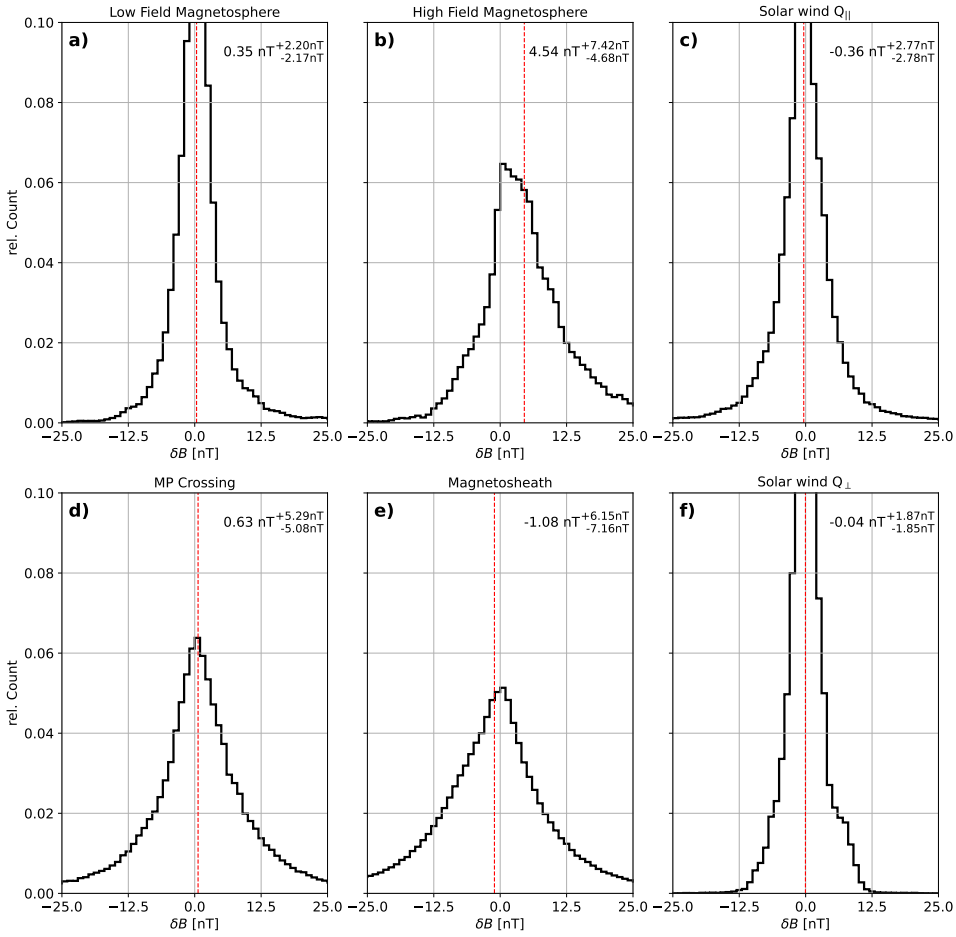

Figure 10: Histograms of the deviation δB between recovered (B_{r}) and original (B_{FGS}) spin-axis measurements in six different regions for the year 2018. Here we use A^2/f_{min} as weight and applied a simple moving average to both recovered and original data (a more detailed discussion of various weighting schemes will follow later).

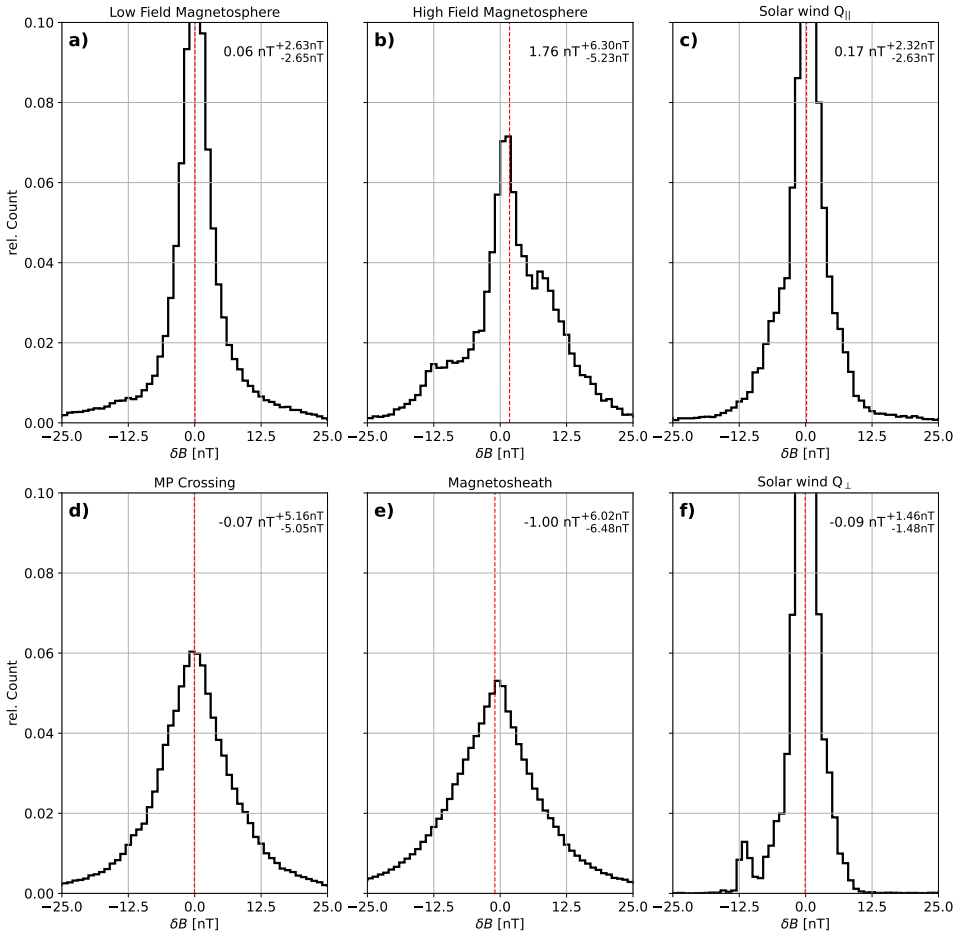

Figure 11: Histograms of the deviation δB between recovered (B_{r}) and original (B_{FGS}) spin-axis measurements in six different regions for the year 2019. Here we use A^2/f_{min} as weight and applied a simple moving average to both recovered and original data (a more detailed discussion of various weighting schemes will follow later).