Bias Corrected Estimation of Paleointensity (BiCEP): An improved methodology for obtaining paleointensity estimates

Brendan J Cych[1,1], Matthias Morzfeld[2,2], and Lisa Tauxe[3,3]

[1]Scripps Institution of Oceanography
[2]UCSD
[3]University of California, San Diego

November 30, 2022

Abstract

The assumptions of paleointensity experiments are violated in many natural and archaeological materials, leading to Arai plots which do not appear linear and yield inaccurate paleointensity estimates, leading to bias in the result. Recently, paleomagnetists have adopted sets of "selection criteria" that exclude specimens with non linear Arai plots from the analysis, but there is little consensus in the paleomagnetic community on which set to use. In this paper, we present a statistical method we call Bias Corrected Estimation of Paleointensity (BiCEP), which assumes that the paleointensity recorded by each specimen is biased away from a true answer by an amount that is dependent a single metric of nonlinearity (the curvature parameter $\vec{k}$) on the Arai plot. We can use this empirical relationship to estimate the recorded paleointensity for a specimen where $\vec{k}=0$, i.e., a perfectly straight line. We apply the BiCEP method to a collection of 30 sites for which the true value of the original field is well constrained. Our method returns accurate estimates of paleointensity, with a higher level of accuracy and precision than the strict CCRIT selection criteria, and with higher accuracy and similar precision to the modified PICRIT03 criteria. The BiCEP method has a significant advantage over using these selection criteria because it achieves these accurate results without excluding large numbers of specimens from the analysis.

Bias Corrected Estimation of Paleointensity (BiCEP): An improved methodology for obtaining paleointensity estimates

Brendan Cych[1], Matthias Morzfeld[1], Lisa Tauxe[1]

[1]University of California, San Diego

Key Points:

- Empirical evidence suggests that paleointensity estimates for non-ideal specimens are biased.
- BiCEP is a method for estimating paleointensity for ensembles of specimens, correcting for bias
- BiCEP produces accurate results when applied to data where the true field strength is known.

Corresponding author: Brendan Cych, `bcych@ucsd.edu`

Abstract

The assumptions of paleointensity experiments are violated in many natural and archaeological materials, leading to Arai plots which do not appear linear and yield inaccurate paleointensity estimates, leading to bias in the result. Recently, paleomagnetists have adopted sets of "selection criteria" that exclude specimens with non linear Arai plots from the analysis, but there is little consensus in the paleomagnetic community on which set to use. In this paper, we present a statistical method we call Bias Corrected Estimation of Paleointensity (BiCEP), which assumes that the paleointensity recorded by each specimen is biased away from a true answer by an amount that is dependent a single metric of nonlinearity (the curvature parameter $\vec{k}$) on the Arai plot. We can use this empirical relationship to estimate the recorded paleointensity for a specimen where $\vec{k} = 0$, i.e., a perfectly straight line. We apply the BiCEP method to a collection of 30 sites for which the true value of the original field is well constrained. Our method returns accurate estimates of paleointensity, with similar levels of accuracy and precision to restrictive sets of paleointensity criteria, but accepting as many sites as permissive criteria. The BiCEP method has a significant advantage over using these selection criteria because it achieves these accurate results without excluding large numbers of specimens from the analysis. It yields accurate, albeit imprecise estimates from sites whose specimens all fail traditional criteria. BiCEP combines the accuracy of the strictest selection criteria with the low failure rates of the less reliable 'loose' criteria.

Plain Language Summary

Paleomagnetists perform experiments on rocks and pottery sherds (among other things) to estimate the strength of the ancient Earth's magnetic field (the paleointensity) through time. These make assumptions that are frequently violated, leading to bias. Quantitative metrics (selection criteria) attempt to screen out 'bad' data. If a particular experiment fails the criteria, the results are ignored. However, there is a lack of agreement as to which set of criteria are the most important and what is considered a failure. One of these criteria quantifies the deviation from the fundamental assumption of linearity between the ancient and laboratory magnetizations. We present a new Bayesian method called Bias Corrected Estimation of Paleointensity (BiCEP), in which we assume that the estimated paleointensity depends on this deviation. We can then use this dependency to correct the paleointensity made on an ensemble of specimens with differing deviations from ideal behavior. BiCEP allows us to calculate accurate estimates of the ancient magnetic field, without ignoring results from non-ideal specimens. We test BiCEP on paleomagnetic data for which the original field strength is well constrained. BiCEP recovers the field strength with similar accuracy to stricter sets of criteria, but gets results for a greater number of sites.

1 Introduction

Estimates of the strength of the ancient Earth's magnetic field are currently made by performing experiments that compare the natural remanent magnetization (NRM) acquired by a specimen while cooling in the Earth's field, to a remanence known as thermal remanent magnetization (TRM) acquired by the specimen while cooling in a known laboratory field. Such experiments include the Königsberger-Thellier-Thellier (KTT) family of experiments (Königsberger, 1938; Thellier & Thellier, 1959), the Shaw family of experiments (Shaw, 1974), and the multi-specimen family of experiments (Hoffman et al., 1989), among others. All of these experimental families make assumptions about the relationship between the magnetic field and the remanent magnetization which may or may not be applicable (see the review by Tauxe & Yamazaki, 2015). In this paper, we will focus on the KTT family of experiments.

KTT type experiments involve a double heating protocol in which a specimen is heated two or more times to a series of temperatures up to the Curie Temperature. At each temperature, the specimen is cooled in two different fields. This has the effect of replacing the NRM with a TRM acquired in a known laboratory field. Data from KTT-type experiments are normally represented by the Arai diagram (Nagata et al., 1963), which plots the NRM magnetization remaining at each temperature step against the magnetization imparted in the laboratory (often referred to as partial TRM or pTRM). The ratio of these two magnetizations, as represented by the slope of the best fitting line to the Arai plot data, is generally taken to be the ratio of the two magnetizing fields (ancient, B_{anc} and laboratory, B_{lab}).

KTT-type experiments rely on several assumptions which are frequently violated in paleointensity experiments. These include thermochemical alteration of specimens which may lead to the production of new magnetic minerals, and an assumption known as reciprocity, which requires that the blocking temperature (the temperature below which grains retain their magnetization after an external field is removed) is the same as the unblocking temperature (the temperature above which grains equilibrate with the external field).

The reciprocity assumption of Thellier and Thellier (1959) is fundamental to Néel's theory for uniaxial single domain grains (Néel, 1949). Néel theory assumes that the electronic spins within magnetic grains are fully aligned, and that the alignment is in one of two directions along an energetically favorable 'easy' axis. In zero field, there is no preference for either direction, but in the presence of a field there is a slight preference for the direction along the easy axis with the smallest angle to the applied field. If the reciprocity assumption is met, then the energy required for the magnetization to change directions along the easy axis is always the same regardless of whether the specimen is cooled from higher temperature (blocking) or heated from room temperature (unblocking) and the two temperatures are identical.

By assuming that electronic spins within magnetic grains are fully aligned, Néel theory fails to take into account a term in the magnetic energy of grains which causes deviations from full alignment, resulting in structures such as the vortex state of, e.g., Williams and Dunlop (1989). Although this effect is present in nearly all magnetic grains, it is insignificant over short length scales (10s of nm) and so uniaxial single domain theory may be a reasonable approximation for smaller, elongate grains. Specimens in paleointensity experiments contain mixtures of grains with different sizes and shapes and a specimen used for paleointensity is likely to include grains for which the applicability of single domain theory does not hold.

Failure of reciprocity and other fundamental assumptions embedded in the KTT family of experiments (laid out by e.g., Thellier & Thellier, 1959) provides a challenge for those analyzing paleointensity data. Paleomagnetists generally use a set of selection criteria which reject an intensity result if the NRM and pTRM data behave in a way which deviate from single domain theory (linear on the Arai plot, see Figure 1a) by more than some arbitrarily chosen threshold value. This is because data that contain a large proportion of non single domain-like grains or which otherwise violate the assumptions of the experiment are likely to give biased results (Tauxe et al., 2021). Selection criteria generally operate in a binary way, with specimens either being 'accepted' or 'rejected' from the estimation of the site mean, where 'site' is the collection of specimens assumed to have cooled in identical external magnetic fields (say, a lava flow or ceramic fragment).

Figure 1 gives a demonstration of biased results in specimens from prepared magnetite powders of increasing grain size that were magnetized in a 60 μT field (Krása et al., 2003). If all assumptions of Thellier and Thellier (1959) were obeyed, we would expect the best fitting lines to data on Arai plots to give a range of values distributed closely about a mean of 60 μT. As the grain size of the powder increases, the Arai plot becomes more curved and the best fitting line to the Arai plot yields a progressively lower intensity estimate. As all the paleointensities estimated from the curved plots are below the expected value, the estimate for the ensemble can be biased, with the high temperature segment having an even lower mean value, and the low temperature segment having a high mean value. The data of Tauxe et al. (2021) also demonstrate downward curved Arai plots in natural samples are biased so this problem may effect many of the results compiled in paleointensity databases like the MagIC database (Tauxe et al., 2016) or PINT (Biggin, 2010).

The curvature of an Arai plot can be quantified using the curvature criterion ($\vec{k}$) of Paterson (2011) (see also Paterson et al., 2014). Curvature is calculated using the reciprocal of the radius of a circle fit to scaled Arai plot data (see Section 2.2.1). While there is no theoretical basis for a circular fit (as opposed to the linear fit, which is firmly rooted in Néel theory), it is a useful approximation that we will exploit in this paper.

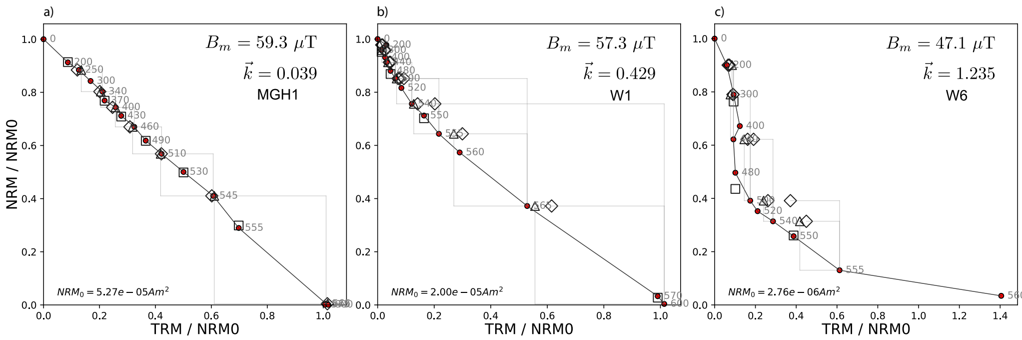

Figure 1. Arai plots from prepared magnetite powders given a TRM in a 60 μT field (Krása et al., 2003). The curvature criterion, $\vec{k}$ (Paterson, 2011) and specimen level paleointensity estimate B_m estimated from fitting a line to the entire Arai plot are plotted on the figure as text. The grain size of the magnetite powders increases from left to right. The coarser grains have non ideal domain state, leading to curved Arai plots and estimates of paleointensity which are biased to lower values than the expected 60 μT. a) Nominal grain size of 23 nm. b) Mean grain size of 70 nm. c) Mean grain size of 12.1 μm.

The practice of using binary (pass/fail) selection criteria is problematic for many reasons. Paleomagnetic specimens generally contain magnetic carriers which span a range of grain sizes and may or may not conform to the assumptions of the method. In addition, micromagnetic simulations (e.g., Williams & Dunlop, 1989; Nagy et al., 2017) demonstrate that the change in magnetic domain state with grain size is a continuum, and so one individual grain's behaviour may be more or less ideal than any other's. With binary pass/fail criteria, the distinction between 'good' and 'bad' data must be assessed with an arbitrary threshold value, which does not reflect the range of behaviors within both groups. Consequently there are a large number of selection criteria in common use (over 40 in Paterson et al., 2014), most of which have some empirical rationale, but there is little agreement on which set to use or their threshold values.

In this paper, we describe a new approach for paleointensity estimation that treats the quality of paleointensity data as a continuum as opposed to the binary 'in' or 'out' approach using selection criteria. We assume that paleointensities become more biased as specimens' magnetic behaviors become more non-ideal and their Arai plots become less linear. By allowing the data interpretation for specimens to be based on the shape of their Arai plots, we are able to obtain unbiased estimates of paleointensity without the need for many specimen level (binary) selection criteria. We call this method the 'Bias

Corrected Estimation of Paleointensity' or BiCEP. In the next section, we develop a Bayesian approach to obtain accurate paleointensity estimates with realistic uncertainties, using $\vec{k}$ as a metric of bias, and show how to combine data at the site level. In Section 3 we compare results from the BiCEP method to those of more traditional selection criteria based approaches. We discuss the results in Section 4 and summarize our conclusions in Section 5. Accompanying this paper, we release a Graphical User Interface (GUI) which can apply the BiCEP method to MagIC formatted data. Links and instructions on how to access the code can be found in Appendix 6.3.

2 Methods

2.1 Accounting for bias in paleointensity experiments

Paleomagnetists determine the paleointensity for a site by performing a Thellier-type double heating experiment on multiple specimens from that site. According to the theory for single domain grains (assuming no alteration of the specimen during heating), the ratio of NRM lost to pTRM gained is the ratio of the ancient field to the laboratory field. If the specimen conforms to theory, the Arai plot data will fall along a line the slope of which is equal to the ratio of ancient to the laboratory field (see Figure 1a).

We expect that the field strength predicted by the slope of the line on the Arai plot for each specimen (here called B_m) will be distributed about the true (expected) ancient field (B_{exp}) at the site with a Gaussian distribution. However, rarely do a set of specimens from a site all produce linear Arai plots that are easily interpretable. For example, interpretation of data from specimens with magnetic grains exhibiting non single domain magnetic domain states produce non-linear Arai plots which violate the assumptions of the method (e.g., Dunlop & Özdemir, 2001). Fitting lines to the data on such Arai plots often produces estimates of paleointensity which are biased (see Figure 1c, Krása et al., 2003), which in turn would bias site level estimates.

Paleomagnetists generally approach non-ideal data by using certain quantitative criteria chosen to eliminate results suffering from one or more pathologies (Paterson et al., 2014). If a particular criterion calculated for a specimen fails to meet some threshold value, then the specimen is excluded from the analysis. In this paper, we present an alternative approach in which we allow for specimens to behave in a non-ideal (non-linear) fashion when considering how specimen intensity estimates are distributed about a site mean and weight the contribution of individual specimen estimates according to linearity. Under such a scheme, we start by predicting a bias for each specimen, and the specimens with the smallest predicted bias most strongly determine the paleointensity at that site. In this way, biased specimens do not strongly affect our site intensity estimate, as they are down-weighted, yet provide useful constraints on the uncertainty.

To predict the amount of bias a specimen is likely to have, we require a proxy for bias in paleointensity experiments. For this we use the curvature criterion $\vec{k}$ of Paterson (2011) (see Section 2.2.1). There are several reasons that make this criterion a useful proxy for bias in paleointensity experiments:

- Specimens that are highly linear have, by definition, low values for $|\vec{k}|$ and will generally give unbiased paleointensity estimates (e.g., Cromwell et al., 2015).
- By contrast, specimens with higher $|\vec{k}|$ yield biased paleointensities, with the magnitude of the bias generally increasing with the magnitude of $|\vec{k}|$ (e.g., Tauxe et al., 2021).
- $|\vec{k}|$ has an empirical correlation with magnetic grain size (Paterson, 2011).

Site	Citation	Material	Lat.	Long.	Year	B_{exp}	M
1991-1992 Eruption Site	Bowles et al. (2006)	lava flow	9.8	-104.3	1991	36.2	53
hw108	Cromwell et al. (2015)	lava flow	19.9	-155.9	1859	39.3	23
hw123	Cromwell et al. (2015)	lava flow	19.1	-155.7	1907	37.7	12
hw126	Cromwell et al. (2015)	lava flow	19.7	-155.5	1935	36.4	13
hw128	Cromwell et al. (2015)	lava flow	19.3	-155.9	1950	36.2	26
hw201	Cromwell et al. (2015)	lava flow	19.4	-155.0	1990	35.2	12
hw226	Cromwell et al. (2015)	lava flow	19.6	-155.5	1843	39.9	11
hw241	Cromwell et al. (2015)	lava flow	19.5	-155.8	1960	36.0	18
BR06	Donadini et al. (2007)	brick	60.1	24.9	1906	49.7	3
P	Muxworthy et al. (2011)	lava flow	19.3	-102.1	1943	44.6	36
VM	Muxworthy et al. (2011)	lava flow	40.8	14.5	1944	43.8	18
BBQ	Pick and Tauxe (1993)	submarine lava flow	9.8	-104.3	1990	36.2	12
rs25	Shaar et al. (2010)	synthetic	N/A	N/A	N/A	30.0	5
rs26	Shaar et al. (2010)	synthetic	N/A	N/A	N/A	60.0	5
rs27	Shaar et al. (2010)	synthetic	N/A	N/A	N/A	90.0	10
remag-rs61	Shaar et al. (2011)	synthetic	N/A	N/A	N/A	40.0	6
remag-rs62	Shaar et al. (2011)	synthetic	N/A	N/A	N/A	60.0	6
remag-rs63	Shaar et al. (2011)	synthetic	N/A	N/A	N/A	80.0	5
remag-rs78	Shaar et al. (2011)	synthetic	N/A	N/A	N/A	20.0	4
kf	Tanaka et al. (2012)	lava flow	65.7	-16.8	1984	52.0	3
Hawaii 1960 Flow	Yamamoto et al. (2003)	lava flow	19.5	-155.8	1960	36.0	22
SW	Yamamoto and Hoshi (2008)	lava flow	31.6	-130.6	1946	46.4	19
TS	Yamamoto and Hoshi (2008)	lava flow	31.6	-130.6	1914	47.8	53
ET1	Biggin et al. (2007)	basaltic lava	37.8	15.0	1950	43.3	3
ET2	Biggin et al. (2007)	basaltic lava	37.8	15.0	1979	44.1	2
ET3	Biggin et al. (2007)	basaltic lava	37.8	15.0	1983	44.2	4
Synthetic60	Krása et al. (2003)	synthetic	N/A	N/A	N/A	60.0	7
LV	Paterson et al. (2010)	Lithic Clasts	-23.4	67.7	1993	24.0	45
MSH	Paterson et al. (2010)	Lithic Clasts	46.2	-122.2	1980	55.6	19
FreshTRM	Santos and Tauxe (2019)	remagnetized/synthetic	N/A	N/A	N/A	70.0	24

Table 1. Table of sites used for analysis in this study, including original study locations, latitude, longitude and year of magnetization (where applicable), expected field at that location (B_{exp}), number of specimens used for analysis at that site M. Lat.: site latitude (°N). Long. site longitude (°E. N/A: Not Applicable (Synthetic). B_{exp} is either a known laboratory field, from the International Geomagnetic Reference Field (IGRF, Thébault et al., 2015 or in two cases (hw226,hw108) using the Arch3k.1 model of Korte et al., 2009

To predict bias, we can use a method by which we minimize the misfit to a model assuming that B_m is linearly related to $\vec{k}$ for all specimens. In other words, we say that:

$$B_m = B_{exp} + c\vec{k}_m + \epsilon \tag{1}$$

where m is an index reflecting the specimen number, ϵ is an error term and B_{exp} is the true value of B. Effectively, our model just becomes a linear fit between the specimen estimate B_m and $\vec{k}$, the y-intercept of which is the true value of the field B_{exp} and c is a slope constant. While there is no theoretical justification (yet) for why B_m would be related to $\vec{k}_m$, although it has been observed empirically (by Paterson, 2011 using the data in Figure 1, and more recently by Tauxe et al., 2021), a linear model is the simplest one to relate the two. We demonstrate in Section 3.3 that more complex models with a quadratic and cubic fit relating B_m to $\vec{k}_m$ perform worse than the linear model when predicting the paleointensity for sites for which the paleointensity is well constrained (historical lava flows or laboratory remanences).

Arai plot curvature is not the sole cause of bias in paleointensity experiments. In some cases, specimens with Arai plots which do not have high $|\vec{k}|$ but are still non linear (e.g., 'zig-zagged' as in, e.g., Yu et al., 2004), may still cause bias in paleointensity experiments. To counteract this, we use a Bayesian method of calculating $\vec{k}_m$ and B_m which provides an uncertainty for both of these parameters. The benefit of this approach is that specimens whose Arai plots are not well fit by a line or an elliptical arc have less influence on the linear fit. Therefore, the specimens with the lowest uncertainty in $\vec{k}$ are generally the most linear, and will have the most influence on the linear fit. Yet, for each specimen, there is a trade off between minimizing the circle fit at a specimen level and the linear fit between B_m and $\vec{k}$ for specimens from the same site, an issue we will deal with in Section 2.2.3.

Figure 2 shows results from our method (detailed in Section 2.2) applied to several sites for which the true value of B_{anc} (here, B_{exp}) is either calculated from the International Geomagnetic Reference Field (IGRF, Thébault et al., 2015) or Arch3k.1 (Korte et al., 2009) for historical flows, or known as the NRM is a laboratory TRM imparted to the specimens. Following Equation 1, the uncertainty in the intercept value of these linear fits gives us the uncertainty for our site value of B_{anc}. In this way, we can obtain an unbiased estimate of B_{anc} without relying on arbitrary binary (accept/reject) criteria to exclude specimen results.

In the following, we detail how the specimen level circle fit $\vec{k}$ and site level paleointensity for unknown values for B (here called B_{anc}) can be calculated. We then compare the efficacy of several different versions of our model to classical selection criteria. We do this using a data compilation from 30 sites updated from Paterson et al. (2014) and Tauxe et al. (2016) for which B_{exp} is well constrained (see Table 1 for details concerning the original publications of the data).

2.2 Statistical Methodology

2.2.1 Estimating curvature

Paterson (2011) proposed a least squares fit of circles in Arai plot data. The parameter $\vec{k}$ of Paterson (2011) is defined as the reciprocal of the radius of a best-fitting circle through the data. It is positive if the circle center is to the upper right of the Arai plot data (concave up, Figure 3a) and negative if the circle center is below and to the left of the Arai plot data (concave down, Figure 3b).

Before fitting to the Arai plot data, Paterson (2011) scales the pTRMs by the maximum pTRM to ensure that the paleointensity data are independent of the laboratory field. For estimating $\vec{k}$, we also subtract the minimum remaining NRM (NRM_{min}) for specimens for which full demagnetization has not been completed and we subtract the

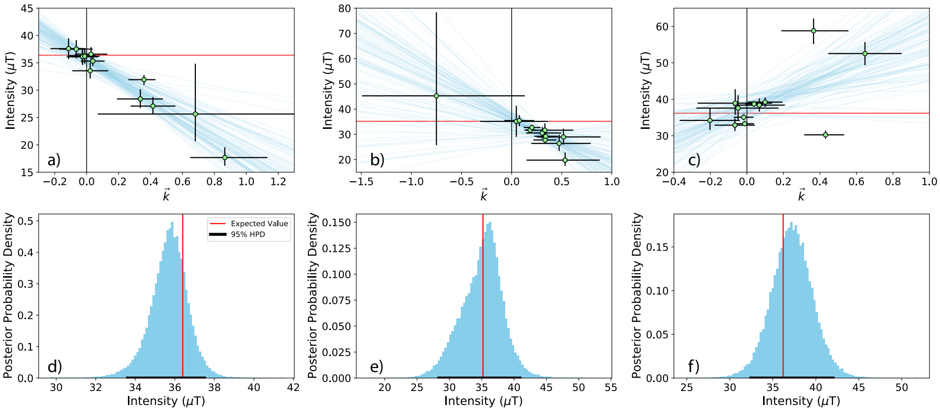

Figure 2. Example of results from the BiCEP method for several sites used as examples in this study. Lines (in blue) are fit to the values of B_m and $\vec{k}$ for each specimen (blue dots, with uncertainties as black lines). The values of linear fits at $\vec{k} = 0$ (blue histograms) provide an unbiased estimate of the expected paleointensity value at the site from the known field (red lines). a,d) hw126. b,e) hw201. c,f) BBQ. See Table 1 for sampling and citation details and Section 3 for comparison with the expected field values, B_{exp}.

minimum pTRM (pTRM_{min}) for specimens for which the low temperature steps were excluded from the analysis (e.g., because of viscous remanent magnetization).

For the BiCEP method, we define two sets of data vectors x and y:

$$x_n = \frac{\mathrm{pTRM}_n - \mathrm{pTRM}_{min}}{\mathrm{pTRM}_{max}}, \qquad y_n = \frac{\mathrm{NRM}_n - \mathrm{NRM}_{min}}{\mathrm{NRM}_0}, \tag{2}$$

where n is the index of the data point. Because scaling should be by the total (original) TRM (the NRM), we also exclude specimens whose NRM_{min} is more than 25% of the initial NRM. This is justified by the assumption that the experimenter did not carry out demagnetization to fully replace the NRM. Then, to fit a circle with center x_c, y_c and radius R to the data, we try to minimize the squared perpendicular distance d_n^2 (Figure 3a) of all the n data points to the circle edge:

$$\sum_{n=1}^{N} d_n^2 \quad \text{where} \quad d_n^2 = (\sqrt{(x_n - x_c)^2 + (y_n - y_c)^2} - R)^2. \tag{3}$$

In a total least squares fit, Equation 3 would be our objective function that we would minimize. To fit circles to the Arai plot using a Bayesian method, we use Bayes' formula (Equation 4). This formula allows us to assign a probability distribution to the values of different parameters (in this case, $\vec{k}_m$ and B_m), rather than just finding the 'best' value of the parameters. In a Bayesian context, we can simply assume that the data have some Gaussian noise distribution with some unknown standard deviation σ and apply Bayes' formula (e.g., Gelman et al., 2004):

$$P(\text{Parameters}|\text{Data}) = \frac{P(\text{Data}|\text{Parameters})P(\text{Parameters})}{P(\text{Data})}, \tag{4}$$

where the left hand side is the probability of the parameters given the data and the right hand side is the probability of the data given the parameters times the probability of the

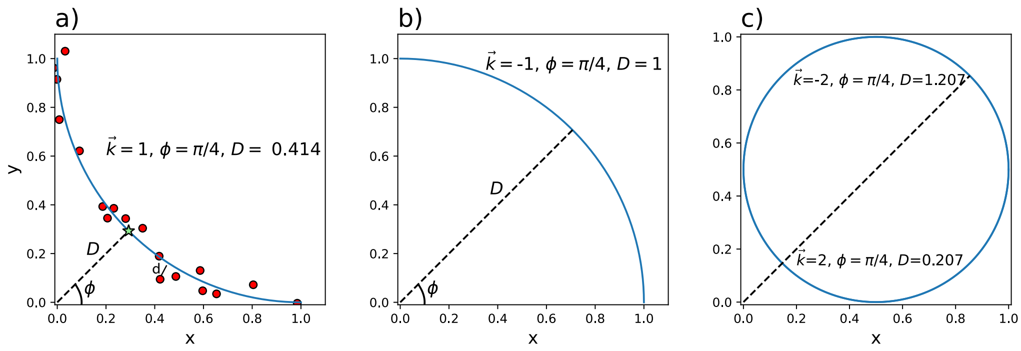

Figure 3. Example circles with different values for parameters $\vec{k}$ and D with the same ϕ, showing how these parameters define a circle. a) Positive $\vec{k}$. Red dots are example data, and the green star is the intersection of D, ϕ with the circle edge (see text for definitions). d is the distance of an individual data point from the best-fit curve (blue). b) Negative $\vec{k}$. Note that in this case, ϕ could take any value as the circle center is at the origin, making the definition of ϕ meaningless in this case. c) Example showing how two sets of the parameters $\vec{k}, \phi, D$ can describe the same circle.

parameters, normalized by the probability of the data. In our case, the parameters are x_c, y_c, R and σ and our data are x and y so we rewrite Equation 4 as:

$$P(x_c, y_c, R, \sigma | x, y) = \frac{P(x, y | x_c, y_c, R, \sigma) P(x_c, y_c, R, \sigma)}{P(x, y)}. \tag{5}$$

The term $P(x, y|x_c, y_c, R, \sigma)$ is known as the "likelihood" and is based on the probability of generating the observed data from a given set of parameters using the assumed Gaussian distribution. The term $P(x_c, y_c, R, \sigma)$ is known as the "prior" and is a probability distribution for values of x_c, y_c, R and σ we consider to be reasonable before we see any data. We consider the priors on these parameters to be independent of one another, so we could rewrite this as $P(x_c)P(y_c)P(R)P(\sigma)$. The term $P(x, y)$ is known as the "evidence", and is simply a normalizing constant that makes the "posterior" probability distribution, $P(x_c, y_c, R, \sigma|x, y)$, integrate to 1. In our application, we can simplify the relationship by ignoring the normalization. Furthermore, we can say from the definition of the Gaussian distribution that:

$$P(x, y | x_c, y_c, R, \sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp\left(\sum_{n=1}^{N} -\frac{d_n^2}{\sigma^2}\right). \tag{6}$$

Now we have an expression for our posterior probability distribution:

$$P(x_c, y_c, R, \sigma | x, y) \propto \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp\left(\sum_{n=1}^{N} -\frac{d_n^2}{\sigma^2}\right) P(x_c, y_c, R) P(\sigma). \tag{7}$$

Because the actual noise distribution of the Arai plot data is quite complicated (Paterson et al., 2012), we do not know the value of σ, so we use the uninformative prior $P(\sigma) \propto \frac{1}{\sigma}$; in other words, the smaller σ, the more likely the result. We can then substitute this prior into Equation 7 and integrate out σ to obtain:

$$P(x_c, y_c, R | x, y) \propto \left(\sum_{n=1}^{N} d_n^2\right)^{-N/2} P(R, x_c, y_c) \tag{8}$$

where N is the total number of measurements considered.

The set of parameters x_c, y_c and R is not easy to solve for, because Equation 3 has multiple local minima (see Chernov and Lesort (2005) for a more detailed discussion). Consider the simple case of a specimen with a linear Arai plot; in even this simplest case, there are four minima, as both R and x_c, y_c will be either positive or negative and very large. To avoid this complexity, we can use instead a change of parameters similar to that of Chernov and Lesort (2005) which Paterson (2011) used as a basis for the circle fitting protocol. Based on this, we define a set of three new parameters which avoid the problem of multiple minima.

Firstly, we require a point on the Arai plot which can be related to a unimodal distribution. We know that linear data will plot along the edge of a circle (the tangent), so if we draw a line from the origin toward the center (x_c, y_c) (not shown), this will touch the edge of the circle at some distance D (green star in Figure 3a). The angle to the horizontal of this line we call ϕ and we can directly estimate the $\vec{k}$ parameter of Paterson (2011) using Equations 9,10,11. We can then establish equations for transforming between these two sets of parameters (see Appendix 6.1 for a more detailed derivation):

$$x_c = \left(D + \frac{1}{\vec{k}}\right)\cos(\phi), \tag{9}$$

$$y_c = \left(D + \frac{1}{\vec{k}}\right)\sin(\phi), \tag{10}$$

$$R = \frac{1}{|\vec{k}|}. \tag{11}$$

Despite this transformation, the circle fitting equation can still have multiple minima, even with $\vec{k}, D, \phi$ as our parameters, as the line connecting the origin to the horizontal touches the circle edge in two locations (see Figure 3c). However, we can use prior distributions to avoid this.

Chernov and Lesort (2005) define a function of the data d_{max} to define the region of possible values for $\vec{k}$:

$$d_{max} = \max_{i,j}\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{12}$$

Additionally, we define distance from the origin to the centroid of the data, d_{cent}:

$$d_{cent} = \sqrt{\bar{x}^2 + \bar{y}^2} \tag{13}$$

Using this function, we can assume that $D < 2d_{cent}$ and $|\vec{k}| < N/d_{max}$ and can define priors for our parameters:

$$P(D) \sim \text{Uniform}(0, 2d_{cent}), \tag{14}$$

$$P(\phi) \sim \text{Uniform}(0, \pi), \tag{15}$$

and

$$P(\vec{k}) \sim \text{Uniform}(-N/d_{max}, N/d_{max}). \tag{16}$$

Using these priors gives us a posterior with a single maximum in most cases, which makes the problem much easier to solve computationally.

We can now apply a Bayesian approach to estimate $\vec{k}$ for all temperature steps for a given specimen m. It is frequently useful to choose a subset of the temperature steps

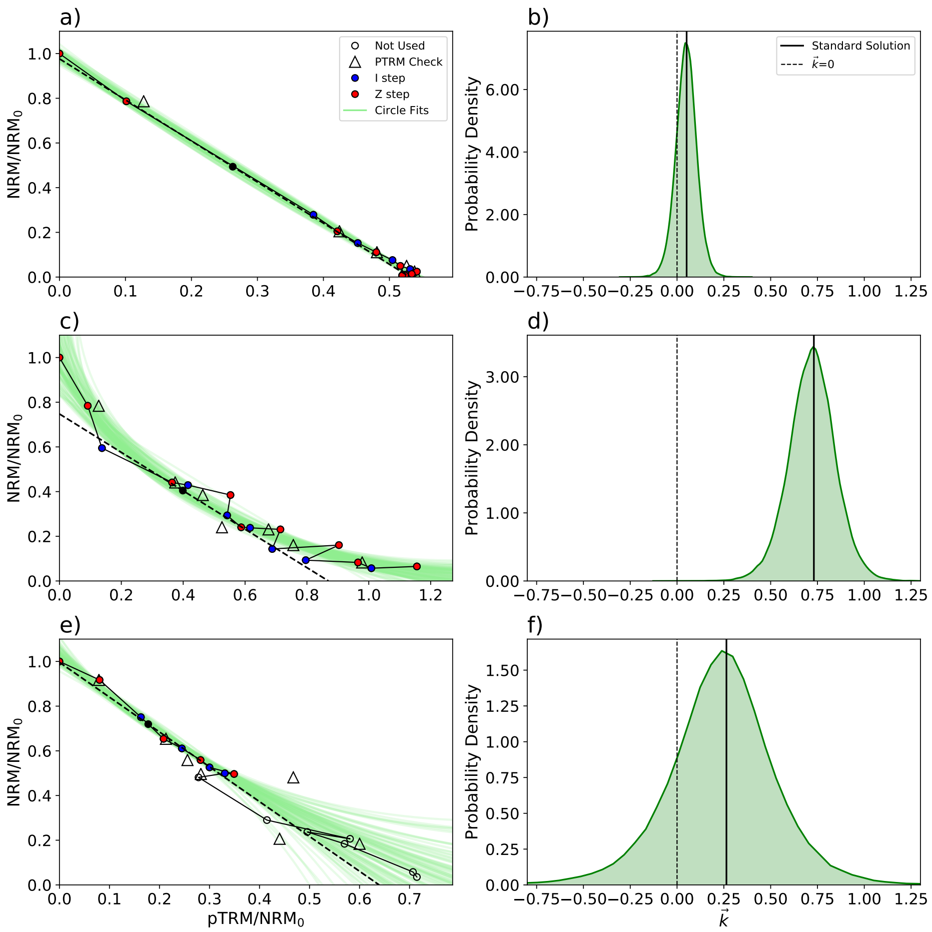

Figure 4. Examples of circle fits to Arai plots (left column) and approximate probability densities of $\vec{k}$ (right column). Dashed lines in the left hand plots are the tangents to circles with the median values for ϕ and D. We use tangents to the circle to get an estimate for B_m as outlined in Section 2.2.2. Triangles in a), c), e) are repeated lower temperature steps (pTRM checks) that indicate alteration of magnetic minerals during the experiment when offset from the original measurements (red dots). a) Specimen hw126a1. A fit to a straight line yields a precise $\vec{k}$ distribution with a maximum close to zero (b). c) Specimen hw126a7. A curved Arai plot with a high amount of scatter/zigzagging (left) results in a higher uncertainty in the value of $\vec{k}$ (d). e) Specimen hw126a6. Arai plot for a specimen that underwent thermochemical alteration at high temperature. A circle fit to just the low temperature steps results in a high uncertainty in the value of $\vec{k}$ (f). Note that we do not exclude any measurements due to thermal alteration in our results section, and that this is only done here for illustrative purposes.

(e.g., if there is evidence for multiple components of the NRM or heating related alteration, as detected by repeated lower temperature pTRM steps). When using a subset of steps, we scale by the maximum pTRM for all temperature steps and the NRM at room temperature; in this way we can predict the curvature for the part of the Arai plot that is missing. This means that interpretations based on a small fraction of the Arai plot will have large uncertainties in the value of $\vec{k}$. Therefore, our circle fit can prioritize interpretations using the largest fraction of the NRM.

Figures 4a,c,e show circle fits sampled from the posterior distributions for specimens from site hw126 (site level results shown in Figure 2a). The probability densities of all the $\vec{k}$ values for each specimen are plotted in Figures 4b,d,f. The plot demonstrates how a straight Arai plot (Figure 4a) produces a narrow posterior about $\vec{k} = 0$ (Figure 4b), while a curved one (Figure 4c) produces a posterior which does not contain $\vec{k} = 0$ (Figure 4d). In the example with failed pTRM checks at higher temperatures (offset triangles in Figure 4e), we exclude the data points represented by open circles and use a linear segment with only a portion of the results, the posterior distribution of $\vec{k}$ has a larger uncertainty on the value, translating to a larger uncertainty in the bias for that specimen. We do not advocate for any particular method of checking for alteration, and do not exclude any measurement steps in our results section. However, our circle fitting routine allows for measurement steps to be excluded and accounts for the increased uncertainty in doing so.

2.2.2 Obtaining a specimen level paleointensity estimate

Analogous to the case in which paleointensity estimates are made using the slope of a fitted line to the Arai plot data, we can obtain a similar "slope" value for a circular arc fit to the data. Consider the case in which the edge of the circle forms an exact line ($\vec{k}$=0, see Figure 4a). In this case, the slope of the line can be given by the tangent to the circle at the point where it intersects a line drawn from the origin (0,0) to the circle center (Figure 3a). In other words, the "slope" of the Arai plot can be estimated as $\cot\phi$, which gives the tangent to the circle. We can then turn this into an intensity estimate B_m using the formula:

$$B_m = \frac{B_{lab}\cot(\phi)}{\mathrm{pTRM}_{max}}, \tag{17}$$

where B_{lab} is the laboratory field used to impart a pTRM to the specimen. If a specimen is corrected for anisotropy, cooling rate, or non-linear acquisition of TRM, we apply this correction to Equation 17.

We now have a way of obtaining estimates for B_m and $\vec{k}_m$ for each specimen. We use the methodology laid out in Sections 2.2.1 and 2.2.2 to plot the median value of the posterior for these parameters (with error bars) in Figure 5a, and examples of circle fits in Figures 5c, e, g. For specimens with values of $\vec{k}$ that are approximately 0 (Figure 5g), the B_m values are quite accurate. There appears to be a bias for specimens with large $\vec{k}$, with the amount of bias increasing as $\vec{k}$ increases. In this example, large positive values of $\vec{k}$ lead to a large underestimates of B_m while negative values of $\vec{k}$ lead to overestimates of B_m (although small in this example).

2.2.3 Obtaining a site level paleointensity estimate

The main problem with the method presented thus far is that we still do not have a way of obtaining an estimate for B_{anc}, the unknown value at the site level. However, in Figure 5a there appears to be a dependence between $\vec{k}_m$ and B_m as suggested earlier, with most of the specimens showing a quasi-linear relationship (the only exception being the point labeled e) whose Arai plot is shown in Figure 5e) and suggests there is a great deal of uncertainty in the value of $\vec{k}$ itself. Because of this, we can modify our model slightly by imposing the extra restriction that B_m must be linearly dependent on $\vec{k}_m$ (with noise) using Equation 1 (substituting B_{anc} for the unknown value of B_{exp}).

Previous papers have assumed that B_{anc} for selected specimens follows a Gaussian distribution and we can also make this assumption here. In the following, we will show how this modification can shift results from specimens that are offset from the linear relationship toward the line (as in the point labeled 'f' in Figure 5b) and produce models (shown as blue lines) that estimate all of our B_m. We can then use the resulting models to estimate the probability distribution for B_m as:

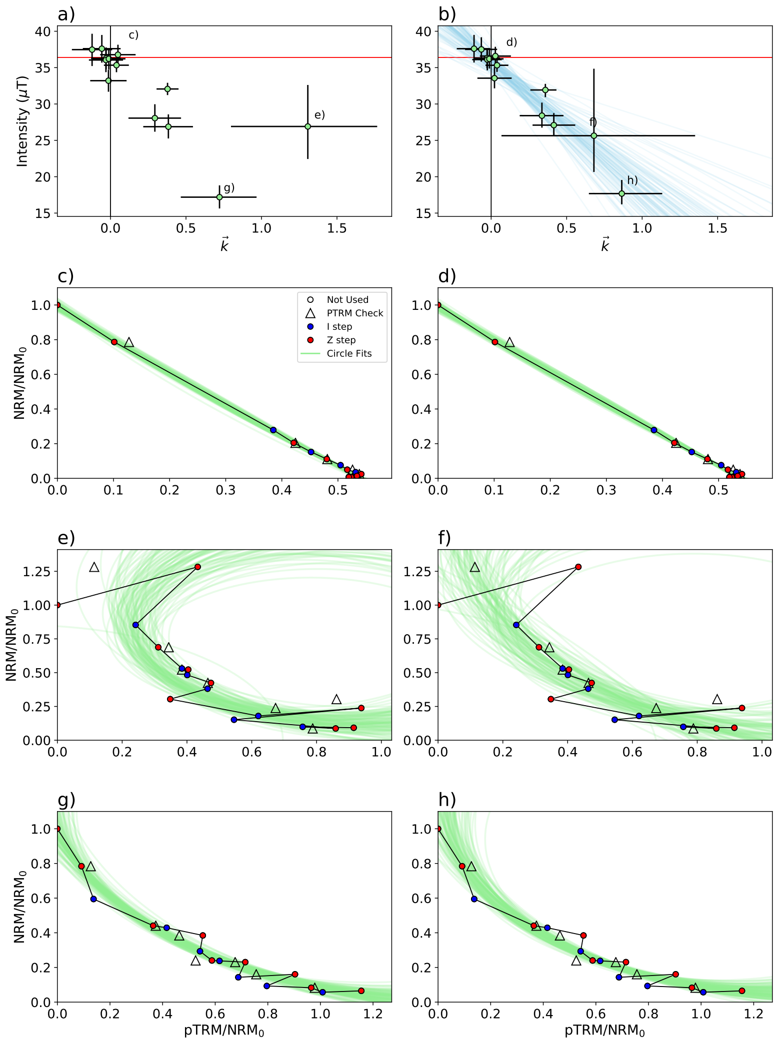

Figure 5. Examples demonstrating how the predicted $\vec{k}$ and B_m for each specimen are modified for a site by using a hierarchical model (Equation 14). The left column shows draws from the posterior for an "unpooled" model where we estimate B_m and $\vec{k}_m$ independently. The right column shows draws from the posterior for the BiCEP method where we assume a linear relationship between B_m and $\vec{k}_m$. a) Red horizontal line is B_{exp} (hw126, see Table 1). 95% credible intervals for $\vec{k_m}$ and B_m are plotted using black error bars, with the medians as green points. b) Representative draws from the posterior distribution are plotted as blue lines assuming that the individual specimen values B_m follow the relationship stated in Equation 14. Note that the higher curvature specimens with large uncertainty in $\vec{k}$ follow a linear trend away from B_{exp}. c),e),g): [Symbols same as in Figure 4.] Arai plots of particular specimens are shown with circle fits sampled from the posterior of the unpooled model shown in a) and plotted in green. In d), f), h), same specimens as in c), e), g) but using the posterior of the BiCEP model in b). Note that there is little change in the specimen in d) for which a close fit to the data is possible, but in f) and h) the curvature (and intensity) of the specimen are modified to fit the line better.

$$P(B_m|k_m, B_{anc}, \sigma_{site}, c) = \frac{1}{\sqrt{2\pi\sigma_{site}^2}} \exp\left(-\frac{(B_{anc} + c\vec{k}_m - B_m)^2}{2\sigma_{site}^2}\right). \tag{18}$$

Now we can combine our expressions for B_m and $\vec{k}_m$ (Equations 17, Sections 2.2.1 and 2.2.2) with the new constraint of a linear relationship between B_m and $\vec{k}_m$ (Equation 18). This allows us to obtain an expression for the site level intensity estimate B_{anc}:

$$P(B_{anc}, \sigma_{site}, c, B_m, k_m, D_m|x_m, y_m) \propto \\ P(x_m, y_m|k_m, D_m, B_m)P(B_m|k_m, B_{anc}, \sigma_{site}, c)P(B_{anc}, \sigma_{site}, c)P(D_m, k_m). \tag{19}$$

Equation 19 may look complicated, but we defined each of the terms already. The benefit of this treatment is that we can obtain $P(x_m, y_m|k_m, D_m, B_m)$ from our circle fitting in Equation 8 (see also Appendix 6.1). We defined $P(B_m|k_m, B_{anc}, \sigma_{site}, c)$ in Equation 18. The values of $\vec{k}$ and B_m for each specimen are needed to fit both of these terms. This means that specimens with large scatter in their Arai plots (those which have Arai plots that are not fit well by a line or a circle) are more strongly affected by the site level fit B_{anc}, and therefore by the specimens with more linear (or circular) Arai plots. Conversely, those specimens with a small uncertainty in $\vec{k}$ or B_m are tightly constrained by the Arai plot fit and so have more control over the fit at the site level.

The other two terms on the right side of Equation 19 ($P(B_{anc}, \sigma_{site}, c)P(D_m, k_m)$), are priors. $P(D_m, k_m)$ were defined in Equations 14 and 16 respectively. Now, we need to define priors for $P(B_{anc}, \sigma_{site}, c)$. For this purpose, we use a poorly constrained prior for the slope, c, where $P(c) \propto 1$. Although this is not a probability distribution, the resulting posterior distribution for B_{anc} is always a real probability distribution if the number of specimens is greater than one. We use a uniform prior between 0 and 250 μT for $P(B_{anc})$ as intensity values can never be negative and in databases such as the MagIC database (Tauxe et al., 2016) or the PINT database of Biggin (2010) rarely (if ever) exceed 250 μT. For $P(\sigma_{site})$ we use a normal distribution with zero mean and standard deviation of 5 μT, truncated to always be positive.

Figure 5b shows our median estimates for B_m and $\vec{k}_m$ after applying the linear restriction. Here, there is a tradeoff between fitting the Arai plot data with the circle, and fitting the linear trend at a site level. The effect of the linear fitting is apparent when compared to estimating $\vec{k}_m$ and B_m for each specimen in isolation, which is shown in Figure 5a. With the linear restriction, the $\vec{k}$ and B_m of specimens are "pulled" closer to a linear trend by modifying the Arai plot fits; specimens with more uncertain $\vec{k}_m$ are more strongly affected (e.g., specimen labeled e) and f) in Figure 5a and b). The specimens with highly linear Arai plots (for which we have small uncertainty in $\vec{k}_m$), the circle fits (see g and h) are mostly unchanged. Despite this modification of the circle fits to the Arai plots by the linear model, the circle fits to those specimens do not look unreasonable.

2.3 Metrics of success

In order to 'ground-truth' the method, we rely on a compilation of paleointensity data updated from that of Paterson et al. (2014) and Tauxe et al. (2016). This compilation has data from 30 sites for which B_{anc} is well constrained (hence we use B_{exp}), either through the IGRF, or because the specimens were given TRMs in a known lab field before the Thellier experiment. One exception to this is for hw226 and hw108, lava flows erupted in Hawaii in 1843 and 1859, prior to the range included in the IGRF. For these sites, we used the Arch3k.1 model of Korte et al. (2009). A list of sites used here is given in Table 1. Instead of choosing a range of temperatures for each site, we simply use every temperature on the Arai plot for all specimens.

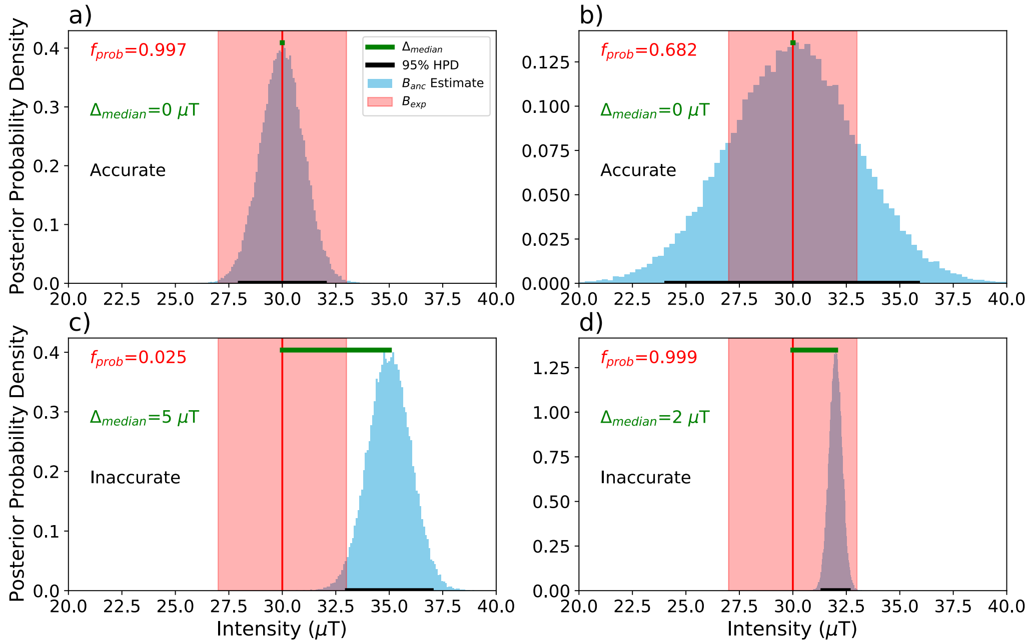

Figure 6. Examples of accuracy and precision metrics used in this study with simulated Gaussian distributions of B_{anc} for illustration. a) An accurate and precise estimate, b) An accurate but imprecise estimate, c) An inaccurate and imprecise estimate. d) A slightly inaccurate and highly precise estimate. Accuracy check used for n_{acc} checks whether the black line intersects the expected value (B_{exp}). f_{prob} is the area of the blue histogram that lies within the red shaded area. Δ_{median} is the length of the green line.

Because we have to estimate multiple parameters for each specimen, our method involves a high dimensional optimization problem. Therefore, we generate the estimates for B_{anc} for a given site using a Markov chain Monte Carlo (MCMC) method which approximates the posterior distribution by generating pseudosamples from it (see Appendix 6.2 for details). MCMC techniques are frequently used to solve high dimensional problems of this kind.

For each site, we quantify the effectiveness of the BiCEP method using several metrics, f_{prob}, Δ_{median} (see Figure 6 for graphical representation), $\bar{f}_{prob}$, and n_{acc}:

1. f_{prob}: We report the median value of our posterior distribution and the 2.5^{th} and 97.5^{th} percentile of the Monte Carlo sample (95% credible interval) as error bars. To quantify the effectiveness of our method, we look at the proportion of the posterior distribution that lies within 3 μT of the expected value of B (B_{exp}) and call this proportion f_{prob}.
2. $\bar{f}_{prob}$: the mean value of f_{prob} over all sites included in the study. A value of 1 is the best possible value and means all our results are accurate and precise to better than 3 μT.
3. Δ_{median}: the difference (in μT) between the median value of the MCMC sample (see Section 6.2 for explanation) and B_{exp}. The median value of Δ_{median} is $\tilde{\Delta}_{median}$. Values of $\tilde{\Delta}_{median}$ close to zero are best.
4. n_{acc}: the number of sites for which B_{exp} lies within our 95% credible interval. A related parameter, f_{acc} is the fraction of results that are accurate (n_{acc}/n_{sites}),

where n_{sites} is the total number of sites analyzed. We expect this number to be 0.95 in ideal circumstances.

We use these metrics to compare the BiCEP results to those obtained by several different sets of selection criteria: CCRIT (Cromwell et al., 2015), Paterson's modified PICRIT03 (here called PICRITMOD) and SELCRIT Criteria (here called SELCRITMOD, Paterson et al., 2014). For this exercise, we also calculated these two criteria with the addition of the curvature criterion of $|\vec{k}| < 0.270$, which we refer to as PICRITMODk and SELCRITMODk. We apply these criteria using the standard deviation optimization method in Thellier GUI. Most sets of commonly used selection criteria rely on an assumption of a Gaussian probability distribution for the site level estimate B_{anc}, which allows us to calculate these same metrics.

For our analyses of our success metrics, we exclude sites that contain fewer than three specimens. For fair comparison, we do not exclude sites from our analyses with traditional selection criteria which have high standard deviation, as we do not do this for BiCEP. If a site fails to produce an estimate of B_{anc} for any reason (for example, selection criteria passed less than two specimens), we assume the prior distribution of a uniform distribution between 0 and 250 μT. This allows us to compare methods directly, with a penalty applied for excluding sites. An excluded site will have f_{prob}=0.012, whereas a site with a highly inaccurate result can have f_{prob} of 0, so exclusion is considered only slightly better than an inaccurate result in this scheme. We discuss the results of this comparison in Section 3.1.

2.4 Width of prior and order of fit

Here we consider several alternative contingent models in order to explore our choices for $P(\sigma_{site})$ and assumptions about the relationship of B_m and $\vec{k}$. In addition to using a standard deviation of 5 μT for $P(\sigma_{site})$, we use standard deviations of 10 μT and 20 μT. The effect of this is hard to conceptualize, but wider priors will prioritize fitting circles to the individual specimens over fitting the linear relationship between B_m and $\vec{k}_m$ at a site level. The practical effect of this is wider posteriors for sites where the number of specimens is small.

So far, we have assumed *a priori* that B_m is linearly dependent on $\vec{k}_m$. Because there is no theoretical reason why this should be the case, we test models for which the relationship between B_m and $\vec{k}_m$ is described by a quadratic polynomial and a cubic polynomial. We would expect a higher order model to more closely fit the individual $\vec{k}_m$ and B_m values, but with a loss of precision due to the more complicated model.

Results for our method, as well as for two sets of selection criteria, are given in Table 2. For each model, we calculate $\bar{f}_{prob}$, $\tilde{\Delta}_{median}$ and f_{acc} for comparison. In this table, our models are named for the value of the standard deviation of $P(\sigma_{site})$ as well as the order of the fit. Our preferred model is referred to as "Linear 5 μT", and this is the model used in this paper where otherwise unspecified.

2.5 MCMC sampler diagnostics

MCMC samplers are only ever an approximation of the posterior distribution, and the number of Monte Carlo samples needed to make an accurate approximation is not the same for every site, or every run of the sampler. To determine whether we are accurately sampling the posterior distribution, we look at three diagnostics which are also described in Appendix 6.2:

1. $\hat{R}$: (Gelman & Rubin, 1992) quantifies convergence between chains in the MCMC method. This parameter is required to be between 1.1 and 0.9 for the sampler to converge.
2. n_{eff}: the effective MCMC sample size. We are using 30,000 Monte Carlo samples and n_{eff} should be large (> 1000) to have a good representation of our parameters.
3. f_{div}: the proportion of divergent transitions f_{div} in the MCMC sample. This should ideally be zero, but it does not appear to cause large problems for the estimate of B_{anc} if it is non zero (see Section 6.2).

The diagnostics n_{eff} and $\hat{R}$ are produced for each of our parameters (each of our $B_m, \vec{k}_m, D_m$ and B_{anc}, σ_{site}). When reporting these values, we look at the worst value of $\hat{R}$ (furthest from unity) and the value of n_{eff} for B_{anc}. If $\hat{R}$ >1.1, we replace the distribution on B_{anc} with a uniform distribution between 0 and 250 μT (the prior). The results of the MCMC sampler are presented in Section 3.4.

3 Results

3.1 Comparison of BiCEP to Selection Criteria

In this section, we compare the BiCEP to several sets of selection criteria (see Section 2.3). The full set of results for all sites can be seen in Figure 7, and are summarized in Supplementary Data Set S1.

Model Name	$\bar{f}_{prob}$	$\tilde{\Delta}_{median}$ (μT)	f_{acc}	Number of Sites
Linear, 5 μT (BiCEP)	0.63	1.7	0.85	25
Linear, 10 μT	0.62	1.7	0.85	25
Linear, 20 μT	0.61	1.7	0.85	25
Quadratic, 5 μT	0.56	1.7	0.81	25
Quadratic, 10 μT	0.55	1.6	0.85	25
Quadratic, 20 μT	0.55	1.8	0.85	25
Cubic, 5 μT	0.45	2.6	0.85	22
Cubic, 10 μT	0.45	2.3	0.85	24
Cubic 20 μT	0.44	2.5	0.85	24
CCRIT	0.47	1.9	0.88	22
PICRIT (Modified)	0.56	2.2	0.77	23
PICRIT (Modified with $\vec{k}$)	0.61	1.9	0.69	21
SELCRIT (Modified)	0.53	2.9	0.58	25
SELCRIT (Modified with $\vec{k}$)	0.58	2.3	0.58	23

Table 2. Results comparing the models used in this study to results using CCRIT (Cromwell et al., 2015) as well as PICRITMOD and SELCRITMOD (Paterson et al., 2014), both with and without the $\vec{k}$ criterion. See details in text and Figure 6 for explanations of the different parameters presented here. Results are sorted by the number of specimens in the site used to make the estimate using our method.

Figure 7 shows the 95% credible intervals for each method, normalized by the expected value at the site. The median values of our results are generally similar to those found by our selection criteria. BiCEP yields the largest number of accurate and precise results, with CCRIT being generally less precise and slightly less accurate. PICRITMOD and SELCRITMOD are generally less accurate and precise than BiCEP, however

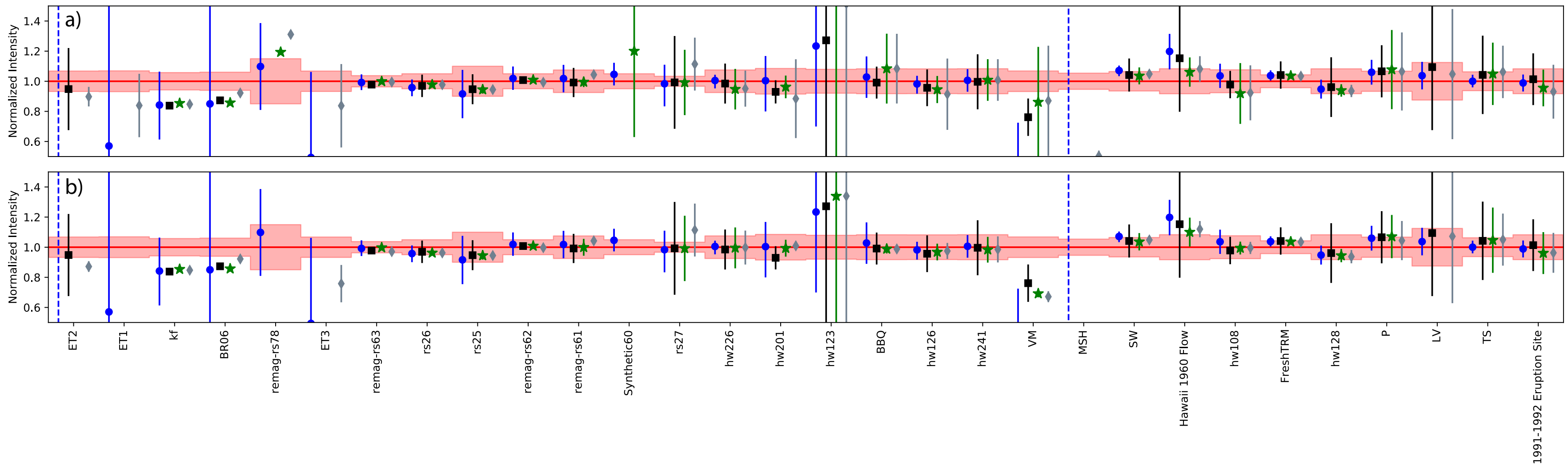

Figure 7. In a) for our collection of sites, we plot (Table 1) paleointensity estimates using BiCEP (blue circles) with 95% confidence interval compared to results using CCRIT (black squares), Paterson's modified PICRITMOD (green stars) and SELCRITMOD (grey diamonds). In b), the same information is plotted, but green stars represent results from PICRITMODk and grey diamonds represent results from SELCRITMODk; the $\vec{k} < 0.270$ criterion has been applied to PICRITMOD and SELCRITMOD. A dashed blue line indicates a site where the sampler failed with $\hat{R} > 1.1$, so the prior distribution (a uniform distribution between 0 and 250 μT was used. The results are normalized to the expected field value for each site (black line, by definition 1) and the pink shaded region represents $\pm$ 3 μT. Sites are ordered by the number of specimens used by BiCEP for paleointensity analysis in that site with the number of specimens increasing to the right. Results are summarized in Supplementary Data Set S1.

introducing the curvature criterion for PICRITMODk and SELCRITMODk improve the accuracy and precision significantly. Both PICRITMODk and SELCRITMODk boast highly precise estimates for passing sites, with similar levels of accuracy to BiCEP. However, this improved accuracy and precision is achieved by excluding more sites, which penalizes these methods using our success metrics.

Sites in Figure 7 are sorted by the number of specimens used by BiCEP for the analysis. Unique to our method, sites with low numbers of specimens (M) have wide credible intervals and sites with high M have narrow credible intervals, so the estimate of B_{anc} becomes more precise as more specimens are measured. This is because calculating the credible interval for a B_{anc} is more similar to calculating the standard error of the mean than the site level standard deviation, which is done for our traditional selection criteria. The increasing precision on B_{anc} leads to some sites with high M having estimates of B_{anc} which are seemingly too precise. These estimates are still generally only a few μT away from the expected value, however, and we discuss potential reasons for this in Section 4.4.

Our results in Table 2 indicate that BiCEP is the method that yields the largest number of accurate and precise results, having a higher $\bar{f}_{prob}$ and lower $\tilde{\Delta}_{median}$ than all of our sets of selection criteria. For selection criteria which include a curvature criterion, much of this improvement comes from BiCEP's inclusion of accurate results for two sites, remag-rs78 and Synthetic60. If we look exclusively at the sites which passed each criterion, PICRITMODk and SELCRITMODk achieve higher levels of precision for those sites (higher $\bar{f}_{prob}$ than BiCEP if only passing sites considered), with PICRITMODk achieving similar levels of accuracy to BiCEP (similar $\tilde{\Delta}_{median}$ for passing sites). This higher level of precision is likely an outcome of using the standard deviation optimization procedure, and is probably not reflective of the true uncertainty judging by the low f_{acc} for both PICRITMODk and SELCRITMODk. CCRIT still achieves lower $\bar{f}_{prob}$ and higher $\tilde{\Delta}_{median}$ than BiCEP even if only passing sites are considered, indicating a slightly lower accuracy and precision overall. Our two selection criteria which do not include a curvature criterion (PICRITMOD and SELCRITMOD) have a larger number of passing sites, including remag-rs78 and Synthetic60, but still have reduced $\bar{f}_{prob}$ and $\tilde{\Delta}_{median}$. Ultimately it seems that BiCEP offers the best of both worlds, passing at least as many sites as the more permissive criteria, and achieving higher accuracy and more realistic precision than the more restrictive criteria.

3.2 Width of the prior

To investigate the role of the prior distribution ($P(\sigma_{site})$), we apply the BiCEP method on the data compilation using a variety of values for its standard deviation (see Table 2). The main effect of varying σ_{site} is that for smaller values, the estimates of B_m and $\vec{k}_m$ for specimens are "pulled" closer to the line being fitted at a site level (see Figure 5a,b). For our estimate of B_{anc}, this means that sites with fewer specimens will be more precise, as it is unlikely that specimen B_m will deviate strongly from the mean. For sites with many specimens, there is little effect as σ_{site} is well constrained by the data.

From Table 2, we see that changes to $P(\sigma_{site})$ seem to have little influence on the effectiveness of the model, as all our f_{acc} values are the same for our linear model regardless of the prior distribution used. We can also see graphically in Figure 7 that our precision is low for sites with small number of specimens (M). Because of this, we favor the version of the model with a 5 μT standard deviation on $P(\sigma_{site})$, as models with higher standard deviations reduce precision without capturing any more sites within their 95% credible intervals.

3.3 Order of polynomial fit

The results for our test sites (Table 2) demonstrate that increasing the order of the polynomial fit decreases the precision of the estimate as demonstrated by reduced values of $\bar{f}_{prob}$. This is expected as there are more parameters to be estimated with the same number of data. The level of accuracy is not significantly improved by increasing the model order. The best quadratic model produced a $\tilde{\Delta}_{median}$ of 1.6 μT, which is not a significant improvement over the value of 1.7 μT for the best linear model to account for the reduction in precision. The number of passing sites is reduced for the cubic model, indicating that the sampler is struggling to fit this model. Consequently, the cubic model produces more inaccurate and less precise results. For this reason, we assume a linear relationship between B_m and $\vec{k}_m$.

3.4 Sampler Diagnostics

Site Name	Worst $\hat{R}$	n_{eff}	f_{div}
1991-1992 Eruption Site	1.00	59741	0.00
hw108	1.00	77959	0.00
hw123	1.01	11687	0.00
hw126	1.00	36130	0.01
hw128	1.00	78978	0.00
hw201	1.00	10641	0.01
hw226	1.00	7139	0.05
hw241	1.00	66565	0.00
BR06	1.01	451	0.00
P	1.00	62252	0.00
VM	1.05	1447	0.00
BBQ	1.00	63082	0.00
rs25	1.00	5614	0.00
rs26	1.00	11866	0.00
rs27	1.00	22211	0.00
remag-rs61	1.00	26746	0.00
remag-rs62	1.00	16916	0.00
remag-rs63	1.00	3788	0.00
remag-rs78	1.00	12388	0.00
kf	1.02	2712	0.00
Hawaii 1960 Flow	1.00	60184	0.00
SW	1.00	36390	0.00
TS	1.00	56518	0.00
ET1	1.01	995	0.00
ET2	6.93	6	0.03
ET3	1.01	424	0.00
Synthetic60	1.00	36572	0.01
LV	1.02	5931	0.08
MSH	2.78	24	0.45
FreshTRM	1.00	81007	0.00

Table 3. Sampler diagnostics (see Section 2.5 for an explanation of each diagnostic) for each site using the BiCEP method.

The sampler diagnostics for each site are given in Table 3. Indicators of poor MCMC sampler performance (worst $\hat{R}$ >1.1, low n_{eff}, high f_{div}) tend to occur at sites with four or fewer specimens, or for specimens where the Arai plots are extremely scattered and the sampler struggles to fit them. In the latter case, it may be possible to exclude these specimens by looking at which specimen level parameters have high $\hat{R}$, as this indicates that fitting a circle to these specimens is inappropriate. We did not exclude specimens on this basis in our analysis, however, we include an option to do this in the BiCEP GUI software (see Appendix 6.3).

The prevalence of high $\hat{R}$ for sites with low numbers of specimens indicates that to get a strongly reproducible answer from this method, paleomagnetists ought to measure five or more specimens per site. In practice, most studies already do this in order to have enough specimens that pass the chosen selection criteria, yet many specimens may be excluded from analysis. Here, we can use all of the specimens measured so there may be no additional burden.

3.5 Summary of Results

After testing all of our contingent models, we prefer the model which assumes the relationship between B_m and $\vec{k}_m$ is linear, and which uses a 5 μT standard deviation on $P(\sigma_{site})$. This model performs better than classical sets of selection criteria, either passing a greater number of sites (than CCRIT, PICRITMODk, SELCRITMODk) or having significantly higher accuracy and precision (than PICRIT, SELCRIT). Our precision increases for sites for which the number of specimens is large, similar to calculating the standard error of the mean when using selection criteria. Unlike selection criteria, the BiCEP method does not require exclusion of large numbers of specimens to obtain an accurate result, which leads us to prefer it over those methods.

4 Discussion

4.1 Advantages of BiCEP compared to selection criteria

BiCEP has significant advantages over the classical selection criteria approach. Firstly, we obtain paleointensity estimates for all sites with at least three specimens, including some which do not contain any specimens that would pass classical selection criteria (see Figure 7). In most cases, our estimates have similar or higher accuracy than the selection criteria approach (evidenced by $\tilde{\Delta}_{median}$ and Figure 7), and this is accomplished while only excluding specimens from the analysis which were not fully demagnetized. In some cases, our method yields results even if none of the selection criteria accept any specimens or are inaccurate. For example, for sites remag-rs78 and Synthetic60, our strict criteria (CCRIT, PICRITMODk, SELCRITMODk) produce no results, and our more permissive criteria (PICRITMOD, SELCRIT) produce less accurate (and in the case of Synthetic60, much less precise) results than BiCEP.

Secondly, the increasing precision of our paleointensity estimate as the number of specimens increases allows for an improved workflow when compared to classical paleointensity criteria. Instead of needing a minimum number of specimens to pass our selection criteria, we can keep measuring specimens until we reach a desired level of precision. We discuss this workflow in more detail in Section 4.3. The property of increasing precision with number of specimens is inherent to Bayesian models and can also be found in the method of Kosareva et al. (2020), although their method does not include the bias correction found in our method.

Thirdly, the BiCEP method propagates the uncertainties from a specimen to the site level. Specimens with more scattered (or non linear, or non circular) Arai plots will have less influence over the specimen mean than those with highly linear Arai plots. In addition to this, the BiCEP method foregoes the need for criteria which are concerned with the length of the line on the Arai plot used to make an interpretation, like the NRM

Fraction (e.g., FRAC of Shaar & Tauxe, 2013). Using a set of temperatures with small FRAC will cause an increase in the uncertainty in $\vec{k}$ (see Figure 4e, f), which will cause this specimen to have less effect on the estimate of B_{anc}, without excluding it from the analysis entirely. We discuss this further in Section 4.5.

4.2 Predictive ability of the method

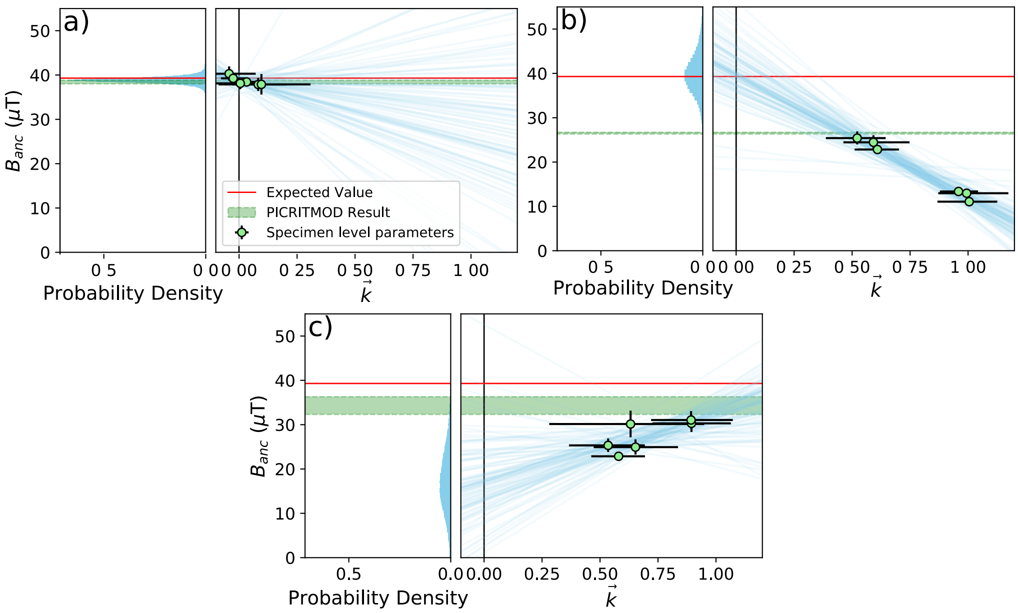

Figure 8. Example of the BiCEP method applied to three subsets of 6 specimens from site hw108 ($B_{exp} = 39.3$ μT). The left column in each subplot shows histograms of the BiCEP results, and the right column shows plots of specimen $\vec{k}$ vs B_{anc} with the BiCEP line fits. Light green shaded regions with dashed edges represent the 2σ interval of the PICRITMOD estimate for these subsamples. In a) there is a small range of $\vec{k}$ and B_{anc} values which leads to an imprecise estimate of c, but an accurate and precise estimate of B_{anc}. In b) there is a large range of values on $\vec{k}$, but all specimens have high $\vec{k}$. This leads to an estimate with a relatively precise estimate of c, and an accurate but imprecise estimate of B_{anc}. In c) there is a reasonably small range of values on B_{anc}, and the relationship between B_{anc} and $\vec{k}$ is not linear, but BiCEP attempts to find a linear model. This leads to an imprecise and inaccurate estimate of both c and B_{anc}.

Although our results are promising, it is worth noting that traditional selection criteria also perform well for the majority of our sites. To see if the BiCEP method offers accurate results with poorer quality data, we subsampled results from site hw108, which had a range of good and poor quality specimens. Figure 8 shows the results of BiCEP applied to three different subsets of six specimens taken from this site, along with the results of the PICRITMOD criteria applied to this site (in green). It is worth noting that only the specimens in Figure 8a would pass the CCRIT criteria which gave a highly accurate result (within 1 μT), or any of our more restrictive criteria.

We identify three behaviours for which BiCEP results deviate from a linear model with high precision on the slope and intercept. Figure 8a shows a subset of specimens for which the range of $\vec{k}$ values of the specimens is very small, and so the uncertainty of the slope of the linear relationship between $\vec{k}_m$ and B_m (c) is high. In this case, however, because these specimens all have $\vec{k}$ close to zero, the estimate of B_{anc} is accurate and precise. Figure 8b shows a different subset of specimens for which the range of $\vec{k}$ values is large, but there are no $\vec{k}$ values close to zero. This results in an estimate of B_{anc} which is still accurate, but imprecise due to the uncertainty in extrapolating the linear relationship between $\vec{k}_m$ and B_m back to zero. The PICRITMOD result for this subset returns an average value which underestimates B_{exp} by around $\sim$13 μT or $\sim$30%, and criteria using the curvature criterion return no values, as all specimens have curvature values higher than the threshold. The high uncertainty in B_{anc} might still be considered a problem, but this result indicates that measuring more specimens would likely yield a more precise result.

Figure 8c shows a set of specimens where the range of $\vec{k}$ is low, so the $\vec{k}_m$ versus B_m relationship is not particularly linear. BiCEP attempts to find a linear trend with these data, and extrapolates back to a B_{anc} which is both highly inaccurate and imprecise. This might be considered a problem for BiCEP, but it is possible to detect such behavior as the uncertainty on both B_{anc} and the slope relating the B_m versus $\vec{k}_m$ (c) are large. This indicates to us that we can use a metric of the uncertainty in both the slope and intercept of the linear fit in BiCEP to decide whether a site level result is accurate or not. This leads us to a laboratory workflow which uses BiCEP results to decide if a site is acceptable, might be acceptable with further work or is unlikely to give a reasonable result.

4.3 Workflow with BiCEP

Figure 9 plots the 95% credible interval on B_{anc} as a percentage against the 95% credible interval on c as a proportion of the median $\tilde{B}_{anc}$ for all sites where $\hat{R} > 1.1$. The sizes of the points on the plot represents the number of specimens per site (M), with squares representing sites with $M < 5$. The colors show the percentage deviation from B_{exp} using BiCEP, with redder colors for more inaccurate results. With the exception of two sites (VM and hw123), as the number of specimens increases, sites trend towards the bottom left region of this plot, indicating an increase in precision. This has diminishing returns as the number of specimens increases above five. The increase in precision is also accompanied with an increase in accuracy. Almost all sites with an estimated precision on B_{anc} better than 40% have median values within 20% of B_{exp}. Our outlier sites VM and hw123, which are imprecise despite having large numbers of specimens (M=12 and 18 respectively), are also inaccurate. This indicates that the width of the 95% credible intervals is a useful statistic for diagnosing inaccuracy in the BiCEP method.

We have divided Figure 9 into four regions (labeled A-D). Sites in region A have high precision on both B_{anc} and c and are representative of the results for the majority of sites in this study; sites in this region are highly accurate. Sites in region B have high precision on B_{anc} (better than 40%, which for a Gaussian distribution would be equivalent to a standard deviation of $\pm$10%) but low precision on c (95% credible interval on $c/\tilde{B}_{anc} > 1$). These sites are usually analogous to the example shown in Figure 8a, with low Arai plot curvature and similar intensities for all specimens. Sites in region C have high precision on B_{anc} but a low precision on c. These sites may have a large number of curved specimens which follow a linear trend that can be extrapolated back to the correct B_{exp}, and are analogous to our example in Figure 8b. Region D is representative of the worst constrained estimates, with low precision on B_{anc} and c. Sites in this region may have highly inaccurate estimates of B_{anc}, often with low M. If these sites have high M, they may be similar to our example in Figure 8c in which a linear relationship between B_{anc} and c is not well determined, and the average $|\vec{k}|$ is large, leading to an inaccurate estimate of B_{anc}.

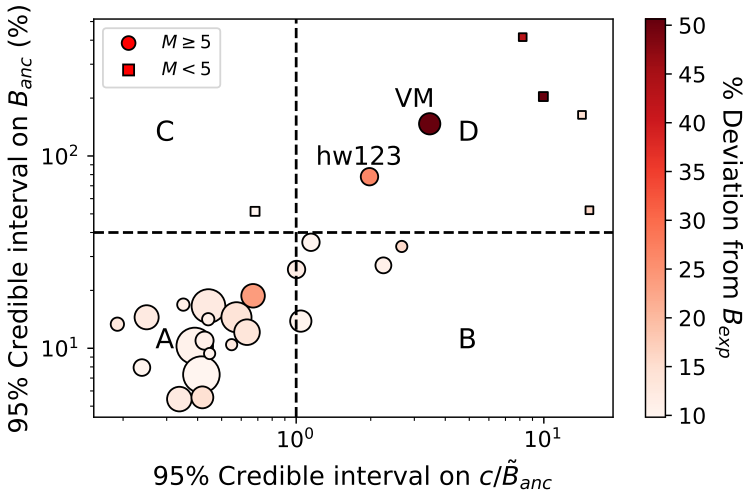

Figure 9. Plot of the 95% credible interval on B_{anc} against the 95% confidence interval on c (slope between intensity estimate and $\vec{k}$), normalized by the median B_{anc} for all sites with $\hat{R} < 1.1$. Circles indicate sites for which the number of specimens $M \geq 5$, and squares indicate sites where $M < 5$. Colors indicate the deviation of the median value of the estimate from the expected site value (B_{exp}) as a percentage. The size of markers is used to represent M. The horizontal dashed line indicates a value of 40% for the full width of the 95% credible interval on B_{anc}, which for a Gaussian distribution would correspond to a standard deviation of ±10%. The vertical dashed line represents a value of 1 for the 95% confidence interval of $c/\tilde{B}_{anc}$. Suggested workflow for sites in regions: A) or B), accept the site or continue measuring if improve precision is desired. C) Continue measuring specimens, as improved precision is likely. D) If $M \geq 5$ stop measuring the site as further effort is likely to be futile. Otherwise continue measuring specimens until $M = 5$.

Considering the region in which a particular site plots leads to a workflow based on the likelihood of success. In general, sites with very low numbers of specimens, ($M =$ 2 or 3), will begin in region D, and migrate to regions C, B or A as M increases to around five. If a site has migrated to region A or B after five specimens have been measured, then we likely have an accurate and precise estimate of B_{anc}, and we can finish measuring specimens (or continue to measure if a higher level of precision is desired). If a site has migrated to region C, it is likely that our estimate is accurate and that our uncertainty in B_{anc} can be reduced by increasing the number of specimens. If our site remains in region D after five specimens have been measured, the site level estimate may be inaccurate, and measuring more specimens would be unlikely to reduce the site level uncertainty.

Because the regions in Figure 9 define a workflow based on measuring five specimens, we wanted to test whether our methodology could identify sites which will remain in region D after a large number of specimens were measured. We randomly subsampled 100 sets of 5 specimens from sites hw108, VM and hw123 and calculated B_{anc} and c using BiCEP. Site hw108 was chosen because contains specimens which exhibit a wide range of behaviours, with a large number of specimens having high $\vec{k}$, but yields an accurate result. Sites VM and hw123 were chosen because these are our sites which remain in region D after measuring a large number of specimens. Our results for these three sites are given in Supplemental Figure S2. Our subsampled hw108 obtained results in region

D 5 times out of 100, whereas our subsampled VM and hw123 obtained results in region D 94 and 90 times respectively. This indicates that sites which remain in region D after measuring 5 specimens are likely to remain there after measuring many more, and so measuring more specimens is usually a futile effort.

4.4 Overly precise estimates of B_{anc}

The BiCEP method has a lower f_{acc} than CCRIT, despite having a similar degree of accuracy when using a metric like $\tilde{\Delta}_{median}$. The reason for this is that the increasing precision on the BiCEP estimate leads to estimates which are highly precise when M is large. This is the case shown in Figure 6d.

Labeling sites with extremely high precision in the estimate as inaccurate may be misleading, as we have not taken into account uncertainties in the value of the expected fields at the sites in this study. For example, using differences between the observed directions and the IGRF, Yamamoto and Hoshi (2008) quoted the expected value at the site "SW" as 46.0±2.6 μT, which is just consistent with the 95% credible interval for our specimen (48.2-49.7 μT). Because of this, we prefer to use $\bar{f}_{prob}$ as a metric of how well a model performs as it allows for a few μT of uncertainty in the expected field value. Additionally, Yamamoto and Yamaoka (2018) suggested that the IZZI-Thellier results for sites SW and TS may be biased slightly high due to acquisition of a thermo-chemical remanent magnetization (TCRM), which is not detectable by our method. Yamamoto et al. (2003) also invoke a TCRM mechanism to explain the paleointensity overestimate for the Hawaii 1960 Flow, which is another of their sites for which we overestimate the expected intensity (see Figure 7 and Supplementary Data Set S1). We note that Cromwell et al. (2015) also sampled the 1960 flow (hw241 which targeted the fine grained flow top) and all selection criteria resulted in accurate results, with BiCEP producing the tightest confidence interval.

4.5 Exclusion of measurement level data

It is frequently possible to improve the accuracy and precision of results by finding the 'best' set of temperature steps to use in the intensity interpretation. Two situations frequently occur for which this might be justified. The first is the case in which thermochemical alteration occurs at high temperature (e.g., Figure 4e). For such specimens, the low temperature measurements can be used to make a paleointensity estimate (colored dots in the figure). Figures 4e and f show how our method can be used on a reduced range of temperature steps on the Arai plot at the cost of precision. The plot of circle fits (green lines in Figure 4e) demonstrates that the Arai plot interpretations are poorly constrained and can continue in any direction after the last temperature step chosen. This results in a higher uncertainty in the curvature associated with this (Figure 4f). The second case in which a portion of the data could be excluded from the calculation, would be when the magnetization has multiple components (Figure 10a). In such a case, a paleointensity estimate can only be made using the small range of temperature steps that correspond to the characteristic component. We currently do not have an objective method to choose which set of temperature steps on the Arai plot to use. We suggest that decisions about which data points to include should not be made based on the original in-field or zero field Arai plot measurements (dots in the Arai plots), but rather exclusively on deviating pTRM checks (triangles in, e.g., Figure 4e) or other indicators of alteration for the first case and on the directions of the magnetization vector (it must trend to the origin and be well defined) in the second case, e.g., Figure 10a.

Caution should be used when excluding a particular temperature steps for reasons other than this. If the set of temperature steps chosen does not represent the characteristic component of magnetization, this can alter the outcome of the BiCEP method, especially if a large part of the Arai plot is excluded. Additionally, excluding more points on the Arai plot tends to increase the chance that a specimen will cause $\hat{R}$ failure. As such, we recommend using as many points on the Arai plot as possible unless done for one of the reasons stated above.

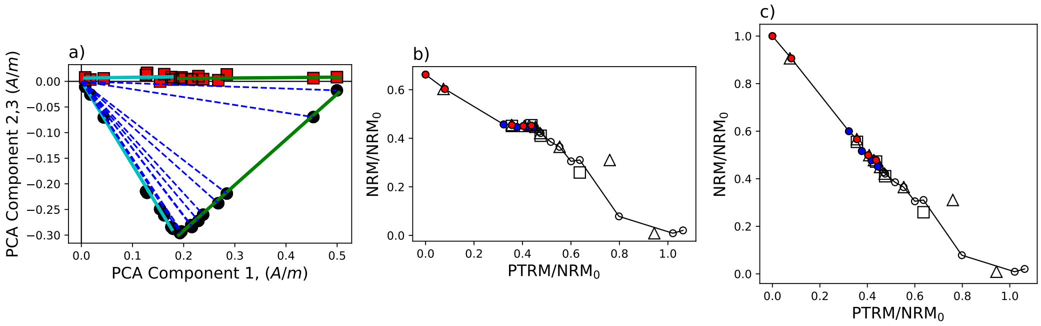

Figure 10. a) Example of vector endpoint diagram for specimen FB2-B1 from Lisé-Pronovost et al. (2020). The magnetization is rotated so that the principal component of the TRM direction for all steps lies along the x axis. Green line fit to the low temperature component and cyan line fit to the high temperature component. b) Arai plot and c) "corrected" Arai plot for a specimen from the data shown in b). NRM values for the low temperature component (filled circles) are usually calculated by taking the magnitude of the vector endpoint (blue dashed lines in the vector endpoint diagram in a). In c), these NRM values are calculated by vector subtracting the high temperature component (cyan line), taking the magnitude of our new NRM vectors (distance along green line), and adding the magnitude of the low temperature component (length of cyan line). Both b) and c) are scaled by the total NRM distance along both components (total distance along both green and cyan lines).

4.6 Application to multi-component magnetizations

We test an application of the BiCEP method on data with multi-component directions as shown in Figure 10a using the data of Lisé-Pronovost et al. (2020). The data are from Scottish firebricks which were used in a foundry in Australia. The date and location of firing are both well constrained, hence we have a reasonably well constrained value for B_{exp}. The bricks all contained a low temperature component associated with the Australian field. Some also displayed a high temperature component associated with the original firing in Scotland as shown in Figure 10a. Lisé-Pronovost et al. (2020) already have interpretations which separate these components in the original study. To account for the change in direction of the NRM, we subtract the high temperature component from the low temperature component, and then add the magnitude of these values to the magnitude of the low temperature component (see Figure 10 for a graphical explanation). The vector subtraction is necessary for the low-temperature component as we need a total TRM (pTRM$_{max}$) to scale by in order to penalize the result for shorter components. We then proceed to use the BiCEP method as previously described, using the original interpretations for the different components. For the sake of simplicity, we do not perform the magnetomineralogical change (MMC) correction (Valet et al., 1996). We also do not apply the corrections for anisotropy of TRM or cooling rate with these data, as they appeared to be negligible. Of course these could be applied in the usual fashion if necessary.

We display the results from multi-component remanences in Figure 11. We find that for the low temperature, Australian field, component (Figure 11a), our estimates for all firebricks contain the expected answer (61.17 μT) within the 95% credible interval. Our

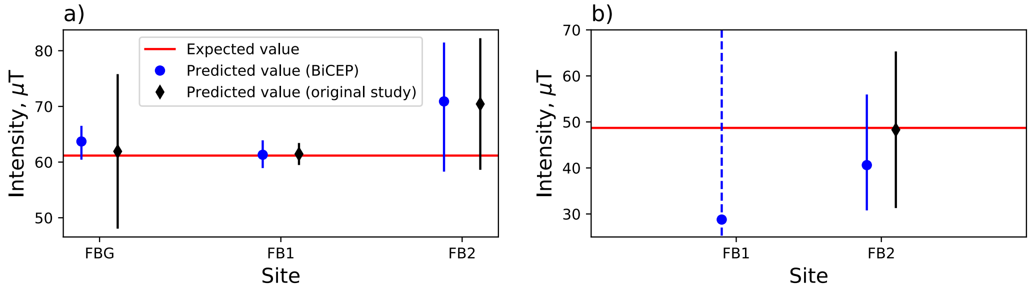

Figure 11. Expected and predicted intensities on the data of Lisé-Pronovost et al. (2020) using BiCEP (blue circles) and the method used in the original study (black diamonds). a) Results for the low temperature component (Australia, expected field value 61.17 μT) for each firebrick. b) Results for the high temperature component (Scotland, expected field value 48.3 μT), where this component was present. The dashed blue line indicates that the MCMC sampler failed to converge for site FB1.

interpretation for site FBG is slightly less accurate than the original analysis but with much higher precision. This difference is likely caused by not applying the MMC correction, as the specimens at this site were mostly of good quality, with none being excluded from the original analysis.

For the high temperature component (Figure 11b) our results behave differently. The sampler does not converge for site FB1, indicating too few specimens in the analysis. For site FB2, we have a result that is less accurate, but more precise than in the original study. The lack of MMC correction may contribute to the decreased accuracy in this example, whereas the reduced precision is likely caused by the smaller length of the interpretation on the Arai plot, leading to a higher uncertainty in the curvature for that specimen. Our results for this study demonstrate that BiCEP will obtain precise estimates for components which represent most of the magnetization, and be imprecise for components which have small NRM fraction.

4.7 Implications for bias in curved Arai plots

The success of our method demonstrates that Arai plot "curvature" or sagging does lead to a progressive bias in paleointensity estimation which increases as the amount of curvature increases as described by Tauxe et al. (2021) and strongly suggested by the data of Krása et al. (2003) (see Figure 1). Our estimates are made by using the tangent to a circle fit rather than fitting a line to part of the data, so one might expect them to be biased. However, it has been demonstrated by e.g. the data of (Krása et al., 2003) that fitting lines to the high temperature or low temperature slope of Arai plots yields even more biased results than using the total TRM, which is more similar to the tangent. The scaling used by our method incorporates the added uncertainty in the line slope and $\vec{k}$ associated with choosing one of these slopes, which allows for more consistent analysis between specimens with interpretations of varying quality. The bias seen generally underestimates paleointensity with higher (positive) curvature, but this is not the case for all sites, some of which exhibit the opposite trend.

The assumption of a quasi-linear dependence between the specimen level paleointensities and the curvature of the Arai plot does not have any theoretical basis. This does imply that the curvature is linearly related to the change in TRM susceptibility (or decay of the original magnetization) between the original and lab coolings, a relationship which should be further investigated. We stress that this relationship only needs to be loosely followed for our method to work. In cases where there does not appear to be a strong linear relationship between B_m and $\vec{k}_m$ (e.g. in Figure 8a), an accurate paleointensity estimate is still possible if there are enough specimens with low $|\vec{k}|$, as the intercept of the linear fit is still well constrained even if the slope is not. Conversely, if there are few specimens with low $|\vec{k}|$ and there is a poor linear relationship, then both the slope and intercept are poorly constrained, resulting in a huge uncertainty in B_{anc}, as is seen in Figure 8c.

5 Conclusions

- We present a new Bayesian method (BiCEP) which accounts for bias in paleointensity estimates in specimens.
- Instead of excluding specimens from the paleointensity analysis in the traditional (binary) selection criteria based approach, our method predicts an amount of bias for each specimen, using the curvature of the Arai plot as a metric of non-linearity and a predictor of bias. In this way, the BiCEP method is quite different from the recently published Bayesian approach of Kosareva et al. (2020).
- When tested on a compilation of sites for which an approximate paleointensity is known *a priori*, our method yields levels of accuracy and precision similar to, or better than restrictive paleointensity criteria, whilst accepting as many results as permissive criteria.
- Our method generates some slightly inaccurate paleointensity estimates with high levels of precision, but these can generally be explained with inaccuracies in the expected field (see Section 4.4).
- The BiCEP method handles uncertainties in a different way than using classical selection criteria, as the uncertainty in site level estimates decreases as the number of specimens increases, but this uncertainty remains high when the number of specimens is low due to inclusion of prior information. The Bayesian uncertainties are in this way more similar to the 'extended error bars' in the Thellier_GUI auto-interpreter of Shaar and Tauxe (2013).
- We propose a workflow in which sites are accepted and measurement of specimens can cease once a desired level of confidence in the site level estimate has been reached. Sites which do not reach this level of confidence after measuring several (> 5) specimens likely do not contain useful information and can be discarded.

Data Availability Statement

Data used in this paper may be found in the MagIC database at: `https://earthref.org/MagIC/17104/0326fdaa-4bcf-44f3-989d-0116b9a2fb75` for review and will be available to the public at `https://earthref.org/MagIC/17104` on publication.

6 Appendix

6.1 Change of variables

In Section 2.2.1 we mention that we need to use a change of variables to get from our original circle fitting parameters R, x_c, y_c to our new set of parameters $\vec{k}, D, \phi$. We can use the Jacobian of the parameter change to get the new formula for the posterior probability under our new parameters:

$$P(D, \phi, \vec{k}|x, y) = P(x_c, y_c, R|x, y)\left|\frac{\partial(x_c, y_c, R)}{\partial(D, \phi, \vec{k})}\right|. \tag{20}$$

We can evaluate this Jacobian as:

$$\left|\frac{\partial(x_c, y_c, R)}{\partial(D, \phi, \vec{k})}\right| = \left|\frac{\vec{k}}{|\vec{k}^3|}\left(D + \frac{1}{\vec{k}}\right)(\cos\phi + \sin\phi)\right|. \tag{21}$$

So our posterior looks like:

$$P(D, \phi, \vec{k}|x, y) \propto \left(\sum_{n=1}^{N} \sqrt{((D + \frac{1}{\vec{k}}\cos\theta) - x_n)^2 + ((D + \frac{1}{\vec{k}}\sin\theta) - y_n)^2} - \frac{1}{|\vec{k}|})^2\right)^{-N/2}$$
$$\left|\frac{\vec{k}}{|\vec{k}^3|}\left(D + \frac{1}{\vec{k}}\right)(\cos\phi + \sin\phi)\right| P(\vec{k}, \phi, D). \tag{22}$$

6.2 Markov chain Monte Carlo sampling

The Markov chain Monte Carlo (MCMC) sampling method generates a set of samples from the posterior probability distribution of B_{anc} which allows us to approximate it. We use the python bindings for the Stan software package (http://mc-stan.org) to generate these samples which provides diagnostic information and runs relatively quickly. For each site we run four Markov chains and generate 30,000 samples of B_{anc} in each chain. We discard the first half of the chain as 'burn in' for a total of 60,000 samples.

Stan provides several diagnostics that tell us whether we have successfully sampled the posterior distribution. These include the $\hat{R}$ score (Gelman & Rubin, 1992) which tells us about the convergence between chains, and is required to be between 1.1 and 0.9 which is necessary for convergence, the effective sample size, n_{eff} which should be large (> 1000) for a good sample and the number of divergent transitions (f_{div}) which should be zero in ideal cases. In most cases our results display high degrees of convergence with $\hat{R}$ close to 1 and high effective sample sizes. Some sites included divergent transitions in small numbers. These seem to occur at a specimen level for specimens where the posterior distribution of one of the circle parameters is long-tailed. In theory this can mean the posterior was inefficiently sampled, but because these specimens generally have large uncertainties on their $\vec{k}$ parameter, the final results do not change, even under a change of parameters. The sampler struggled to converge, with $\hat{R} > 1.1$ for several sites with very few specimens, where once again the distributions are extremely long tailed. The sampler also did not converge for site MSH, where the Arai plots were so non linear, with few points, that BiCEP struggled to fit circles to them. We consider these sites to have "failed" using our method (grade of 'D' in Figure 9) and use the prior distribution on B_{anc} (uniform between 0 and 250 μT) as an estimate of their intensity. We calculate the $\hat{R}$ furthest from unity, the n_{eff} for B_{anc} and the proportion of divergent samples f_{div} for our model.

6.3 Code and GUI

We present a simple GUI that can perform the BiCEP method on data in the MagIC format. The code uses Jupyter notebooks and can be found at (`http://github.com/bcych/BiCEP_GUI`) and contains a readme file detailing how to use the notebook. The GUI can also be accessed at the Earthref JupyterHub site (`http://jupyterhub.earthref.org`). To access the GUI this way:

- Sign up to Earthref at (`http://earthref.org`)
- Navigate to the Earthref JupyterHub site at (`http://jupyterhub.earthref.org`)
- Open and run all the cells in the "BiCEP GUI - Setup.ipynb" notebook.

- Upload MagIC formatted "sites", "samples", "specimens" and "measurements" files to the BiCEP_GUI directory in JupyterHub. These can be formatted using pmag_gui (Tauxe et al., 2016).
- Open the BiCEP GUI notebook and press the "App Mode" button.

For more detailed instructions, read the included readme file at the github site.

Acknowledgments

We are deeply grateful to Lennart de Groot and Greig Paterson for their very helpful reviews and for the advice and guidance given by Andrew Roberts, David Heslop and Joseph Wilson. This research was supported in part by NSF Grant EAR1827263 to LT. We are also grateful to Agnes Lisé-Pronovost for sharing her measurement level data for use in section 4.6.

References

Biggin, A. J. (2010). Paleointensity database updated and upgraded. *EOS*, *91*, 15.

Biggin, A. J., Perrin, M., & Dekkers, M. J. (2007). A reliable absolute palaeointensity determination obtained from a non-ideal recorder. *Earth and Planetary Science Letters*, *257*(3), 545-563. doi: https://doi.org/10.1016/j.epsl.2007.03.017

Bowles, J., Gee, J. S., Kent, D. V., Perfit, M. R., Soule, S. A., & Fornari, D. J. (2006). Paleointensity applications to timing and extent of eruptive activity, 9°–10°n east pacific rise. *Geochemistry, Geophysics, Geosystems*, *7*(6). doi: https://doi.org/10.1029/2005GC001141

Chernov, N., & Lesort, C. (2005). Least squares fitting of circles. *Journal of Mathematical Imaging and Vision*, *23*(3), 239–252. doi: 10.1007/s10851-005-0482-8

Cromwell, G., Tauxe, L., Staudigel, H., & Ron, H. (2015). Paleointensity estimates from historic and modern hawaiian lava flows using glassy basalt as a primary source material. *Phys. Earth Planet. Int.*, *241*, 44–56. doi: 10.1016/j.pepi.2014.12.007

Donadini, F., Kovacheva, M., Kostadinova, M., Casas, L., & Pesonen, L. (2007). New archaeointensity results from scandinavia and bulgaria: Rock-magnetic studies inference and geophysical application. *Physics of the Earth and Planetary Interiors*, *165*(3), 229-247. doi: https://doi.org/10.1016/j.pepi.2007.10.002

Dunlop, D., & Özdemir, O. (2001). Beyond Néel's theories: thermal demagnetization of narrow-band partial thermoremanent magnetization. *Phys. Earth Planet. Int.*, *126*, 43-57.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (Second ed.). Chapman & Hall/CRC, Boca Raton, FL.

Gelman, A., & Rubin, D. B. (1992, 11). Inference from iterative simulation using multiple sequences. *Statist. Sci.*, *7*(4), 457–472. doi: 10.1214/ss/1177011136

Hoffman, K. A., Constantine, V. L., & Morse, D. L. (1989). Determinaton of absolute palaeointensity using a multi-specimen procedure. *Nature*, *339*, 295-297.

Königsberger, J. G. (1938). Natural residual magnetism of eruptive rocks. *Terrestrial Magnetism and Atmospheric Electricity*, *43*(3), 299-320. doi: 10.1029/TE043i003p00299

Korte, M., Donadini, F., & Constable, C. G. (2009). Geomagnetic field for 0–3 ka: 2. a new series of time-varying global models. *Geochemistry, Geophysics, Geosystems*, *10*(6). doi: https://doi.org/10.1029/2008GC002297

Kosareva, L. R., Kuzina, D. M., Nurgaliev, D. K., Sitdikov, A. G., Luneva, O. V., Khasanov, D. I., ... Spassov, S. (2020). Archaeomagnetic investigations in Bolgar (Tatarstan). *Stud. Geophys. Geod.*, *64*(2), 255–292. doi: 10.1007/s11200-019-0493-3

Krása, D., Heunemann, C., Leonhardt, R., & Petersen, N. (2003). Experimental procedure to detect multidomain remanence during thellier–thellier experiments. *Phys. Chem Earth (A/B/C)*, *28*(16), 681 - 687. (Paleo, Rock and Environmental Magnetism 2002) doi: 10.1016/S1474-7065(03)00122-0

Lisé-Pronovost, A., Mallett, T., & Herries, A. I. R. (2020). Archaeointensity of nineteenth-century scottish firebricks from a foundry in melbourne, australia: comparisons with field models and magnetic observatory data. *Geological Society, London, Special Publications*, *497*(1), 27–45. doi: 10.1144/SP497-2019-72

Muxworthy, A. R., Heslop, D., Paterson, G. A., & Michalk, D. (2011). A preisach method for estimating absolute paleofield intensity under the constraint of using only isothermal measurements: 2. experimental testing. *Journal of Geophysical Research: Solid Earth*, *116*(B4). doi: https://doi.org/10.1029/2010JB007844

Nagata, T., Arai, Y., & Momose, K. (1963). Secular variation of the geomagnetic total force during the last 5000 years. *J. Geophys. Res.*, *68*(18), 5277-5281. doi: 0.1029/j.2156-2202.1963.tb00005.x

Nagy, L., Williams, W., Muxworthy, A. R., Fabian, K., Almeida, T. P., Conbhuí, P. Ó., & Shcherbakov, V. P. (2017). Stability of equidimensional pseudo–single-domain magnetite over billion-year timescales. *Proc. Natl. Acad. Sci. U.S.A.*, *114*(39), 10356–10360. doi: 10.1073/pnas.1708344114

Néel, L. (1949). Théorie du traînage magnétique des ferromagnétiques en grains fins avec applications aux terres cuites. *Ann. géophys.*, *5*, 99–136.

Paterson, G. A. (2011). A simple test for the presence of multidomain behavior during paleointensity experiments. *J. Geophys. Res.*, *116*. doi: 10.1029/2011JB008369

Paterson, G. A., Biggin, A. J., Yamamoto, Y., & Pan, Y. (2012). Towards the robust selection of Thellier-type paleointensity data: The influence of experimental noise. *Geochem. Geophys. Geosyst.*, *13*(5). doi: 10.1029/2012GC004046

Paterson, G. A., Muxworthy, A. R., Roberts, A. P., & Mac Niocaill, C. (2010). Assessment of the usefulness of lithic clasts from pyroclastic deposits for paleointensity determination. *Journal of Geophysical Research: Solid Earth*, *115*(B3). doi: https://doi.org/10.1029/2009JB006475

Paterson, G. A., Tauxe, L., Biggin, A. J., Shaar, R., & Jonestrask, L. C. (2014). On improving the selection of thellier-type paleointensity data. *Geochem. Geophys. Geosyst.*, *15*(4), 1180–1192. doi: 10.1002/2013GC005135

Pick, T., & Tauxe, L. (1993). Holocene paleointensities: Thellier experiments on submarine basaltic glass from the east pacific rise. *Journal of Geophysical Research: Solid Earth*, *98*(B10), 17949-17964. doi: https://doi.org/10.1029/93JB01160

Santos, C. N., & Tauxe, L. (2019). Investigating the accuracy, precision, and cooling rate dependence of laboratory-acquired thermal remanences during paleointensity experiments. *Geochem., Geophys., Geosyst.*, *20*(1), 383–397. doi: 10.1029/2018GC007946

Shaar, R., Ron, H., Tauxe, L., Kessel, R., & Agnon, A. (2011). Paleomagnetic field intensity derived from non-sd: Testing the thellier izzi technique on md slag and a new bootstrap procedure. *Earth and Planetary Science Letters*, *310*(3), 213-224. doi: https://doi.org/10.1016/j.epsl.2011.08.024

Shaar, R., Ron, H., Tauxe, L., Kessel, R., Agnon, A., Ben-Yosef, E., & Feinberg, J. M. (2010). Testing the accuracy of absolute intensity estimates of the ancient geomagnetic field using copper slag material. *Earth and Planetary Science Letters*, *290*(1), 201-213. doi: https://doi.org/10.1016/j.epsl.2009.12.022

Shaar, R., & Tauxe, L. (2013). Thellier_gui: An integrated tool for analyzing paleointensity data from thellier-type experiments. *Geochem. Geophys. Geosys.*, *14*, 677–692. doi: doi:10.1002/ggge.20062

Shaw, J. (1974). A new method of determining the magnitude of the paleomanetic field application to 5 historic lavas and five archeological samples. *Geophys. J. R. astr. Soc.*, *39*, 133-141.

Tanaka, H., Hashimoto, Y., & Morita, N. (2012, 05). Palaeointensity determinations from historical and Holocene basalt lavas in Iceland. *Geophysical Journal International*, *189*(2), 833-845. doi: 10.1111/j.1365-246X.2012.05412.x

Tauxe, L., Santos, C., Cych, B., Zhao, X., Roberts, A., Nagy, L., & Williams, W. (2021). Understanding non-ideal paleointensity recording in igneous rocks: Insights from aging experiments on lava samples and the causes and consequenes of 'fragile' curvature in arai plots. *Geochem. Geophys. Geosyst.*, *22*, e2020GC009423. doi: 10.1029/2020GC009423

Tauxe, L., Shaar, R., Jonestrask, L., Swanson-Hysell, N. L., Minnett, R., Koppers, A. a. P., … Fairchild, L. (2016). PmagPy: Software package for paleomagnetic data analysis and a bridge to the magnetics information consortium (MagIC) database. *Geochem., Geophys., Geosyst.*, *17*(6), 2450–2463. doi: 10.1002/2016GC006307

Tauxe, L., & Yamazaki, T. (2015). Paleointensities. In M. Kono (Ed.), *Geomagnetism* (2nd Edition ed., Vol. 5, p. 461-509). Elsevier.

Thébault, E., Finlay, C. C., Beggan, C. D., Alken, P., Aubert, J., Barrois, O., … Zvereva, T. (2015). International Geomagnetic Reference Field: the 12th generation. *Earth Planets Space*, *67*(1), 79. doi: 10.1186/s40623-015-0228-9

Thellier, E., & Thellier, O. (1959). Sur l'intensité du champ magnétique terrestre dans le passé historique et géologique. *Ann. Geophys.*, *15*, 285.

Valet, J.-P., Brassart, J., Le Meur, I., Soler, V., Quidelleur, X., Tric, E., & Gillot, P.-Y. (1996). Absolute paleointensity and magnetomineralogical changes. *Journal of Geophysical Research: Solid Earth*, *101*(B11), 25029-25044. doi: https://doi.org/10.1029/96JB02115

Williams, W., & Dunlop, D. J. (1989). Three-dimensional micromagnetic modelling of ferromagnetic domain structure. *Nature*, *337*, 634–637.

Yamamoto, Y., & Hoshi, H. (2008). Paleomagnetic and rock magnetic studies of the sakurajima 1914 and 1946 andesitic lavas from japan: A comparison of the ltd-dht shaw and thellier paleointensity methods. *Phys. Earth and Planet. Inter.*, *167*, 118-143.

Yamamoto, Y., Tsunakawa, H., & Shibuya, H. (2003). Palaeointensity study of the hawaiian 1960 lava: implications for possible causes of erroneously high intensities. *Geophys J Int*, *153*(1), 263-276.

Yamamoto, Y., & Yamaoka, R. (2018). Paleointensity study on the Holocene surface lavas on the Island of Hawaii using the Tsunakawa-Shaw method. *Front. Earth Sci.*, *6*. doi: 10.3389/feart.2018.00048

Yu, Y., Tauxe, L., & Genevey, A. (2004). Toward an optimal geomagnetic field intensity determination technique. *Geochem., Geophys., Geosyst.*, *5*(2). doi: 10.1029/2003GC000630

Supporting Information for "Bias Corrected Estimation of Paleointensity (BiCEP): An improved methodology for obtaining paleointensity estimates"

Brendan Cych[1], Matthias Morzfeld[1], Lisa Tauxe[1]

[1]University of California, San Diego

[1]9500 Gilman Drive, La Jolla, CA, 92093, USA

Contents of this file

1. Figure S1- plots of B_{exp} vs B_{anc} results using the BiCEP method and other sets of selection criteria used in this study.

2. Figure S2- plots of the credible intervals of BiCEP results for subsamples of 5 specimens from sites hw108, VM and hw123.

3. Figures S3 to S32- plots of unpooled B_{anc} and $\vec{k}$ fits and the BiCEP method applied to all sites used in this study.

Additional Supporting Information (Files uploaded separately)

1. Caption for Data Set S1 - This dataset contains the full set of results shown in Figure 7 and described in Section 3.

Data Set S1

This data set is a csv file containing the full set of results for each version of the BiCEP method (Linear, Quadratic and Cubic models with 5, 10 and 20 μT standard deviations

---

on the prior for σ_{site}, vs classic selection criteria. Columns are named by the model class (Linear, Quadratic, Cubic, or Selection Criteria, followed by the criteria name/prior standard deviation in μT) and then the parameter in question, either the percentile (2.5%, 50%, 97.5%) or f_prob for the f_{prob} value. The first three columns of the csv file contain the site name, B_{exp} and number of specimens (M).

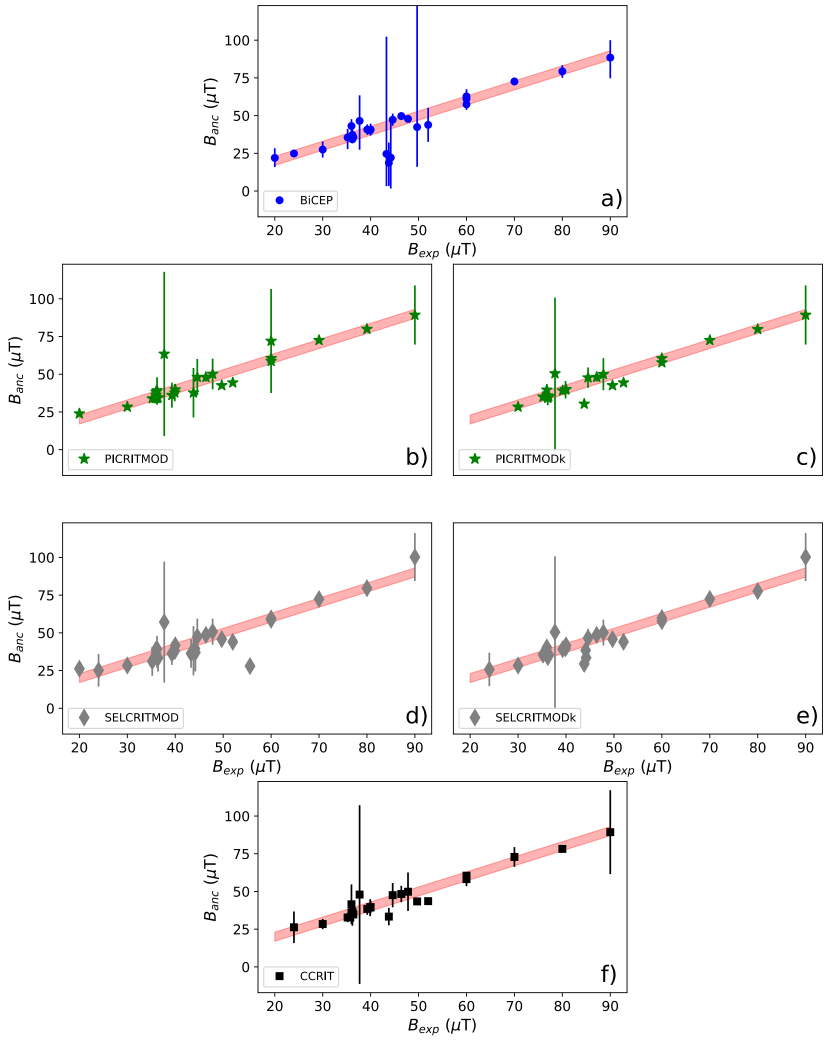

Figure S1. Plot of B_{anc} vs B_{exp} for the passing sites for BiCEP and each of our criteria. a) BiCEP results, b) PICCRITMOD results, c) PICRITMODk results, d) SELCRIT results, e) SELCRITMODk results, f) CCRIT results. Red shaded area represents B_{exp} ± 3 μT. Note that for BiCEP, most of the severely underestimated and highly inaccurate results are for sites with low numbers of specimens ($M < 5$) which did not pass the other criteria.

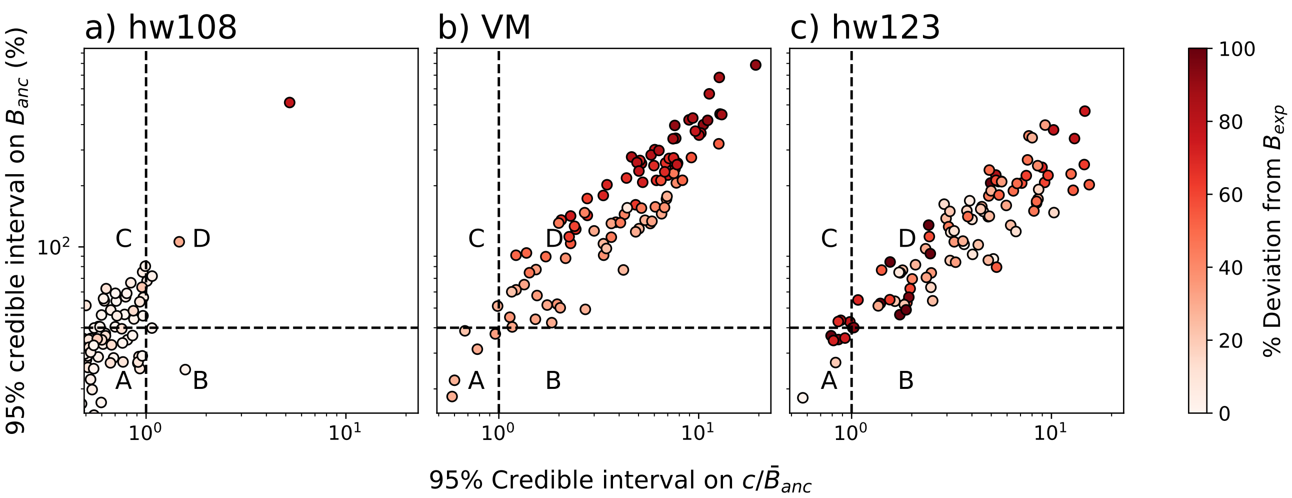

Figure S2. Figure in the style of Figure 9 from the main manuscript. A set of 100 random subsamples of 5 specimens were taken from a) site hw108, b) site VM and c) site hw123. We applied the BiCEP method to these subsamples and plot the 95% credible interval on B_{anc} as a percentage against the 95% credible interval on $c/\tilde{B}_{anc}$. Dashed lines represent the boundaries between regions A, B, C and D, defined in Section 4.3, and point colors represent the percentage deviation of the median from B_{exp}. It is apparent that for our sites where we obtain an inaccurate answer for a large number of specimens (VM and $hw123$), our subsamples fall in region D the majority of the time (94/100 and 90/100 respectively). Conversely, our site hw108, which returns an accurate result with higher numbers of specimens only has 5 results in region D. This indicates that we should either continue measuring specimens (region C, 24/100 results) or that we have a relatively accurate and precise result already (regions A and B, 66/100 results). Although few sites/subsamples were used for this analysis, this test indicates that our labelling scheme has a predictive accuracy of 90-95%

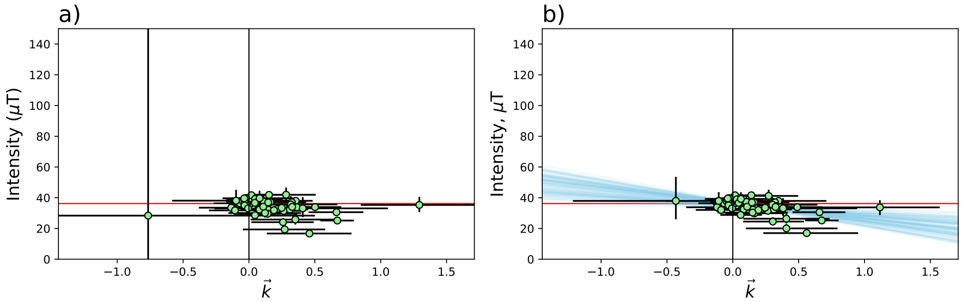

Figure S3. Unpooled and BiCEP models applied to site 1991-1992 Eruption Site a). Independent estimates of B_{anc} vs $\vec{k}$ are plotted as green circles. These estimates are made without assuming a linear relationship between B_{anc} and $\vec{k}$, similar to the analysis in Figure 9a) Black error bars represent 95% credible intervals. b) The BiCEP method applied to this site. Green circles and error bars represent the 95% credible interval, blue lines represent draws from the posterior distribution. The $\vec{k}$=0 axis is plotted as a black vertical line, and the value of B_{exp} is plotted as a red horizontal line. For an accurate estimate, the blue lines should cross the point where these two lines intersect.

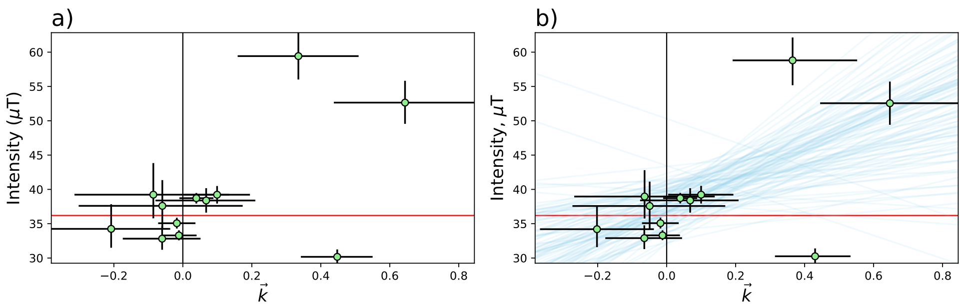

Figure S4. The same methodology described in the caption for Figure S3 applied to site BBQ

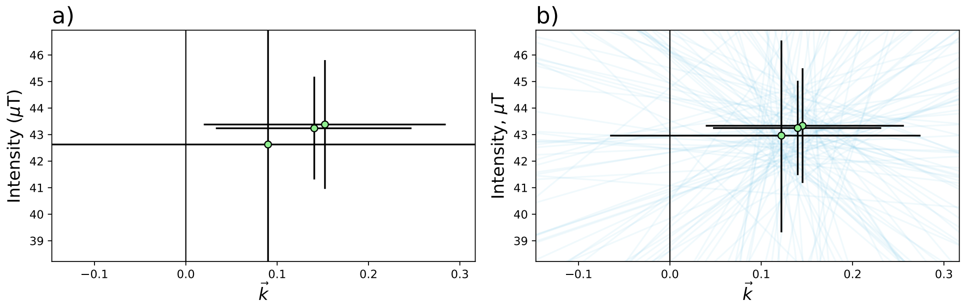

Figure S5. The same methodology described in the caption for Figure S3 applied to site BR06

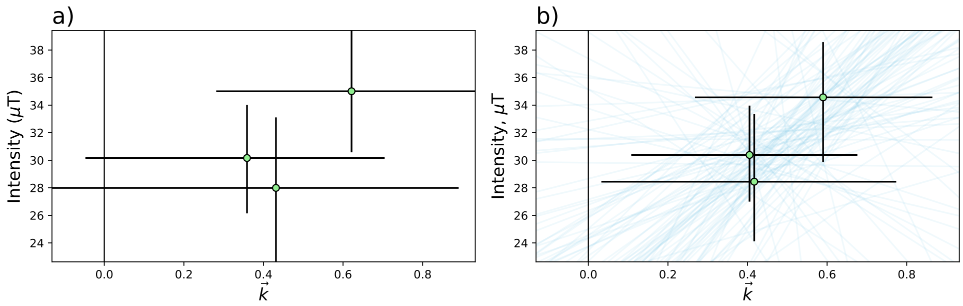

Figure S6. The same methodology described in the caption for Figure S3 applied to site ET1

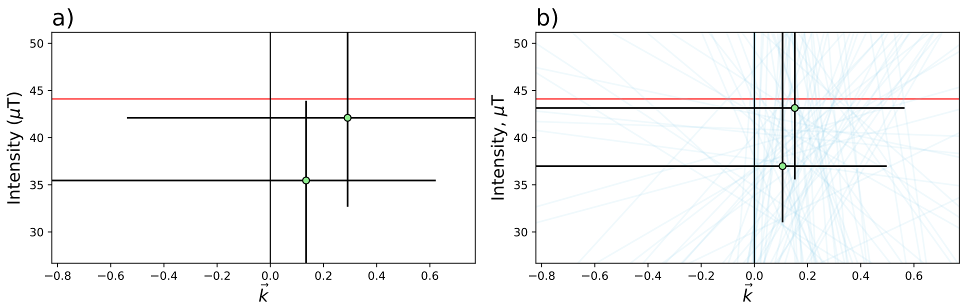

Figure S7. The same methodology described in the caption for Figure S3 applied to site ET2

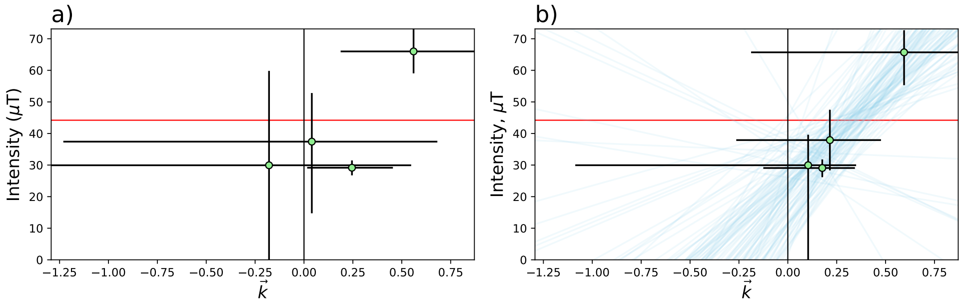

Figure S8. The same methodology described in the caption for Figure S3 applied to site ET3

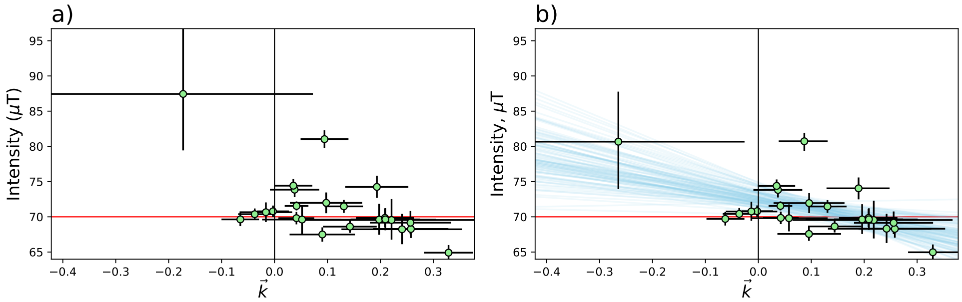

Figure S9. The same methodology described in the caption for Figure S3 applied to site FreshTRM

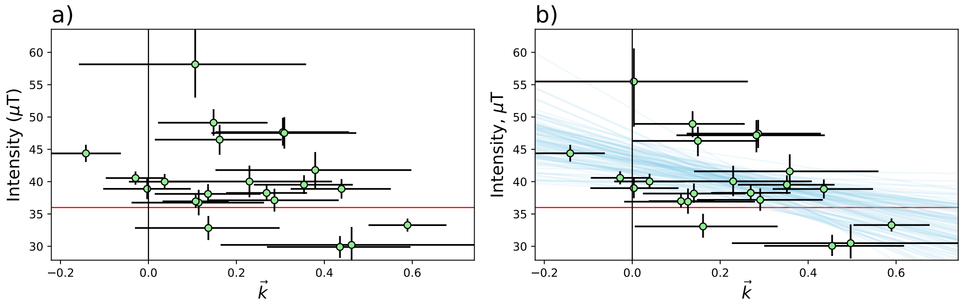

Figure S10. The same methodology described in the caption for Figure S3 applied to site Hawaii 1960 Flow

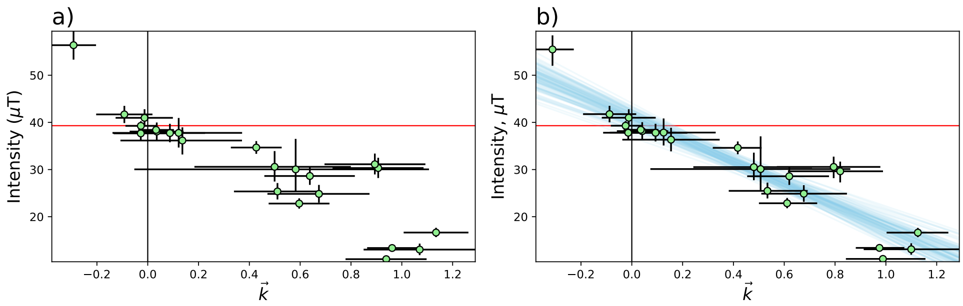

Figure S11. The same methodology described in the caption for Figure S3 applied to site hw108

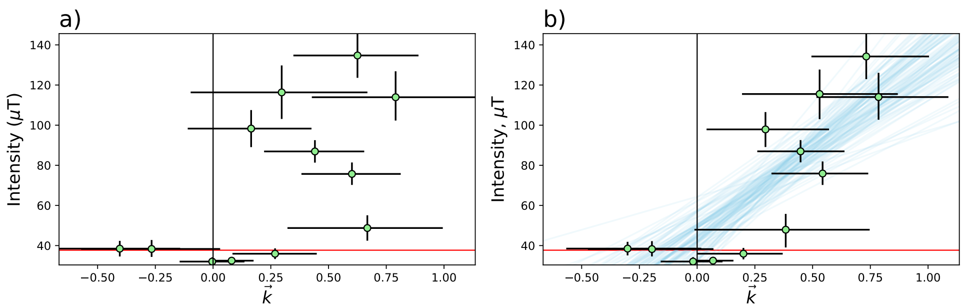

Figure S12. The same methodology described in the caption for Figure S3 applied to site hw123

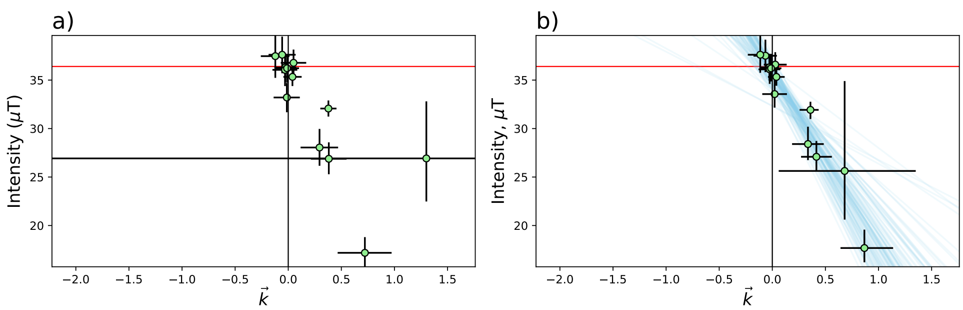

Figure S13. The same methodology described in the caption for Figure S3 applied to site hw126

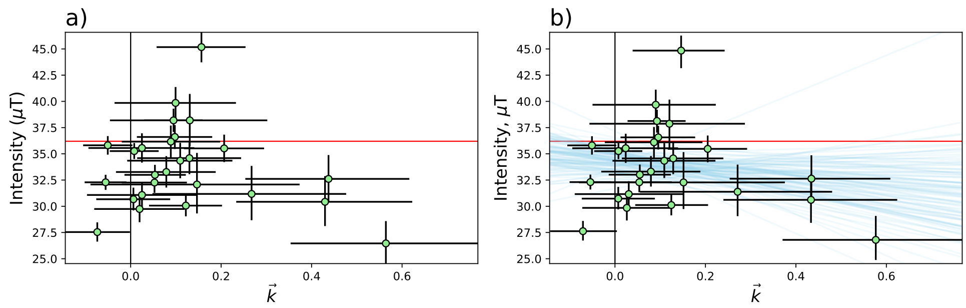

Figure S14. The same methodology described in the caption for Figure S3 applied to site hw128

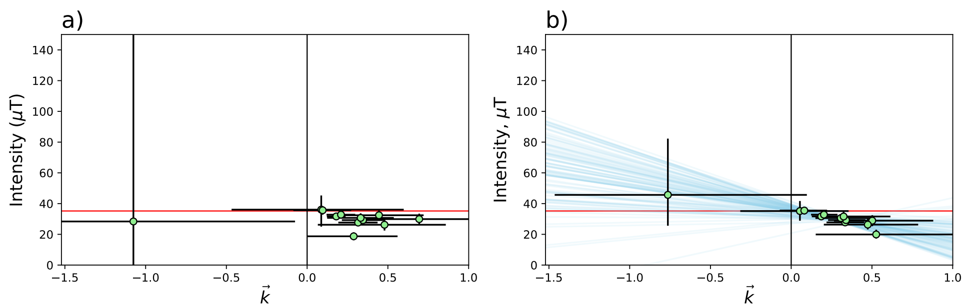

Figure S15. The same methodology described in the caption for Figure S3 applied to site hw201

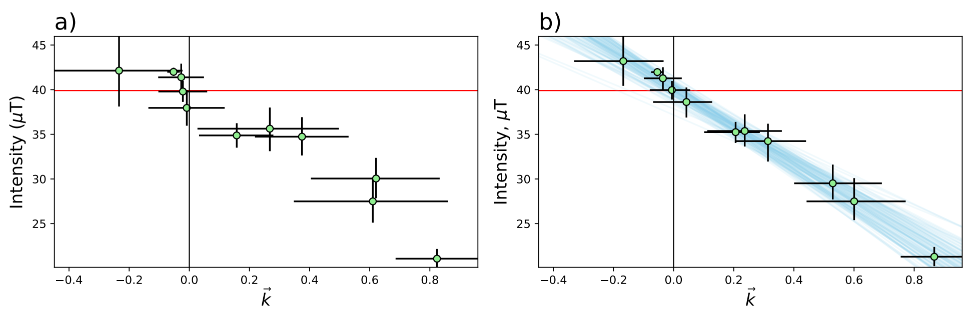

Figure S16. The same methodology described in the caption for Figure S3 applied to site hw226

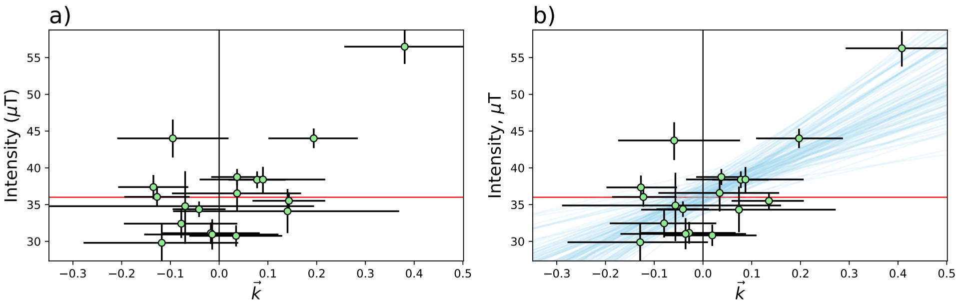

Figure S17. The same methodology described in the caption for Figure S3 applied to site hw241

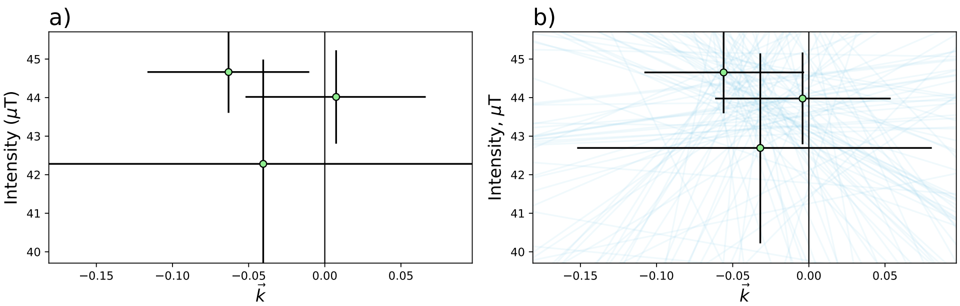

Figure S18. The same methodology described in the caption for Figure S3 applied to site kf

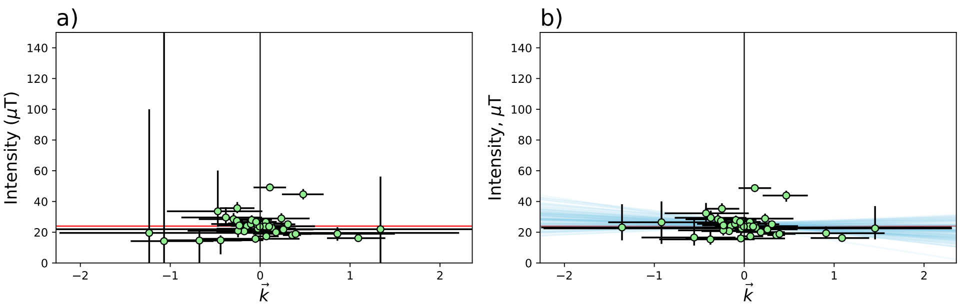

Figure S19. The same methodology described in the caption for Figure S3 applied to site LV

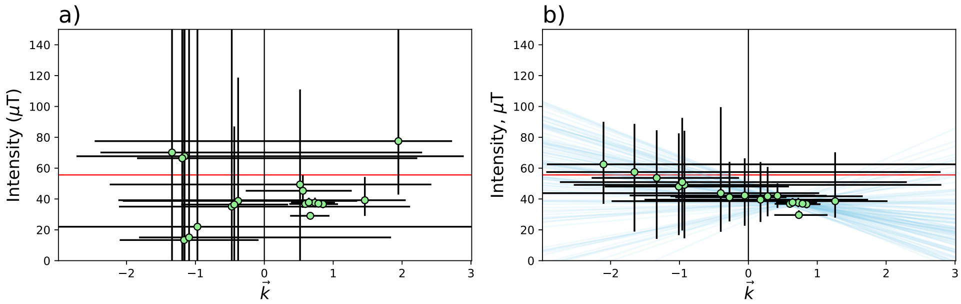

Figure S20. The same methodology described in the caption for Figure S3 applied to site MSH

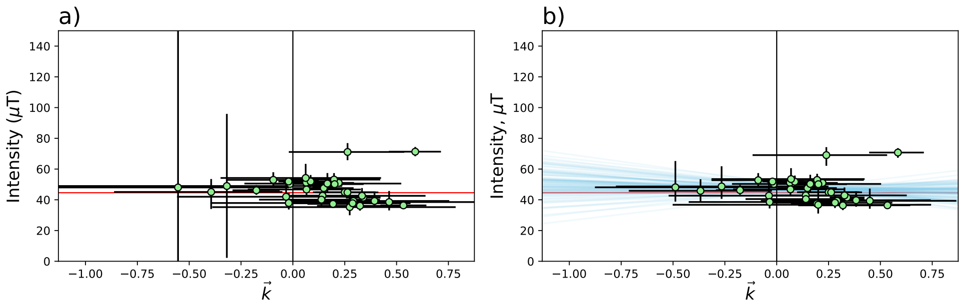

Figure S21. The same methodology described in the caption for Figure S3 applied to site P

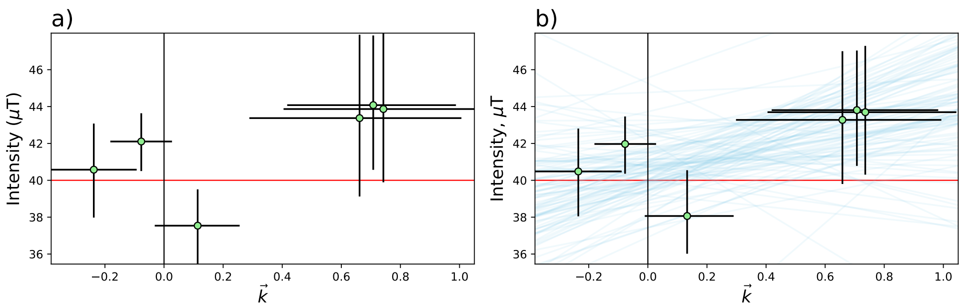

Figure S22. The same methodology described in the caption for Figure S3 applied to site remag-rs61

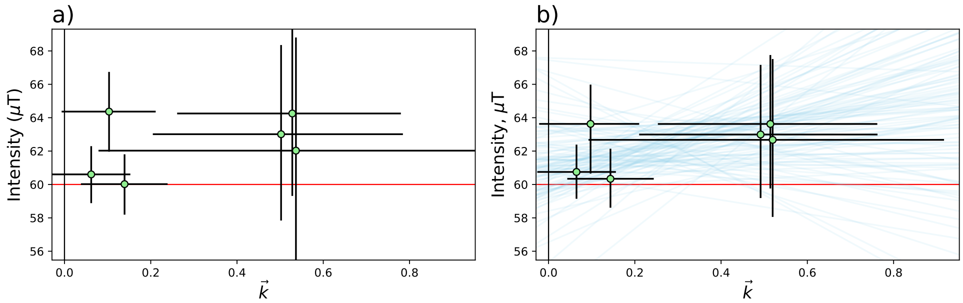

Figure S23. The same methodology described in the caption for Figure S3 applied to site remag-rs62

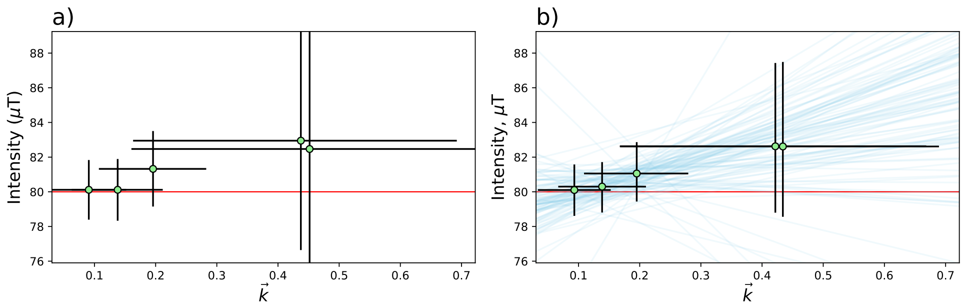

Figure S24. The same methodology described in the caption for Figure S3 applied to site remag-rs63

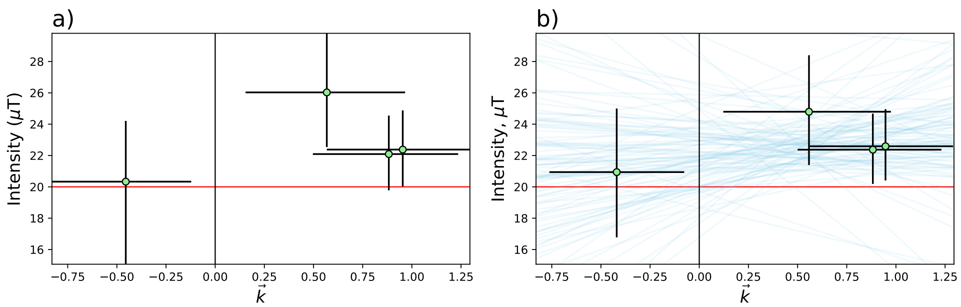

Figure S25. The same methodology described in the caption for Figure S3 applied to site remag-rs78

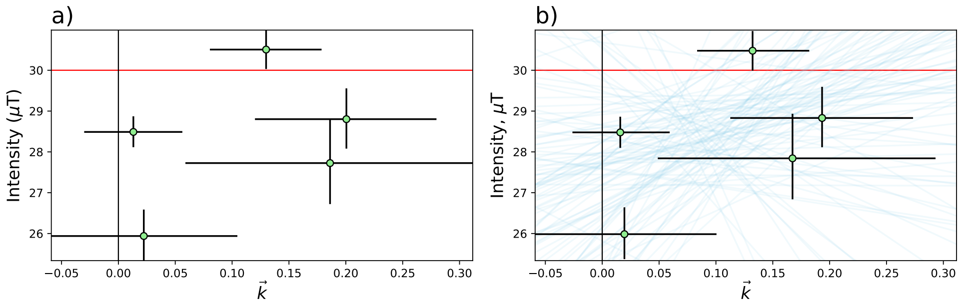

Figure S26. The same methodology described in the caption for Figure S3 applied to site rs25

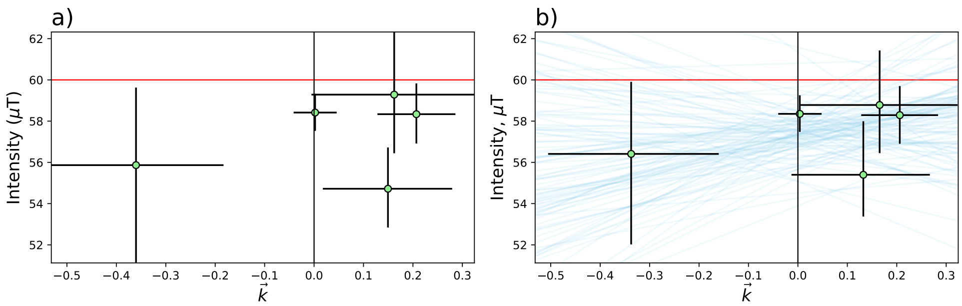

Figure S27. The same methodology described in the caption for Figure S3 applied to site rs26

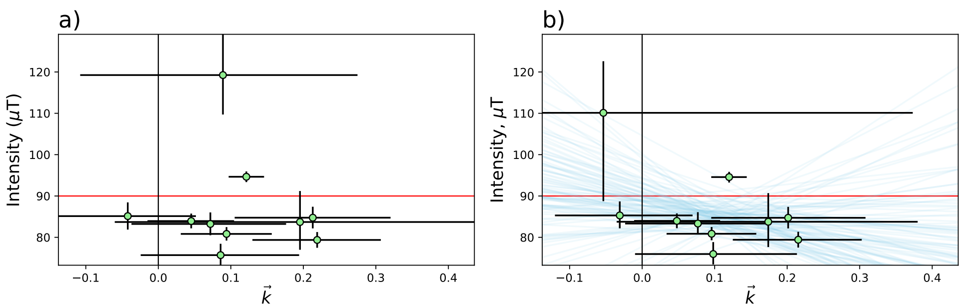

Figure S28. The same methodology described in the caption for Figure S3 applied to site rs27

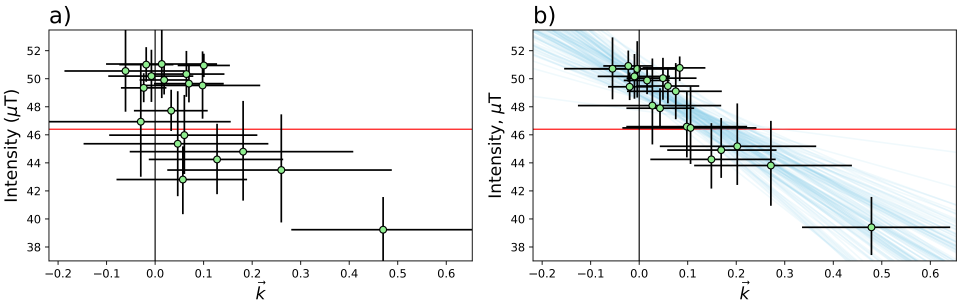

Figure S29. The same methodology described in the caption for Figure S3 applied to site SW

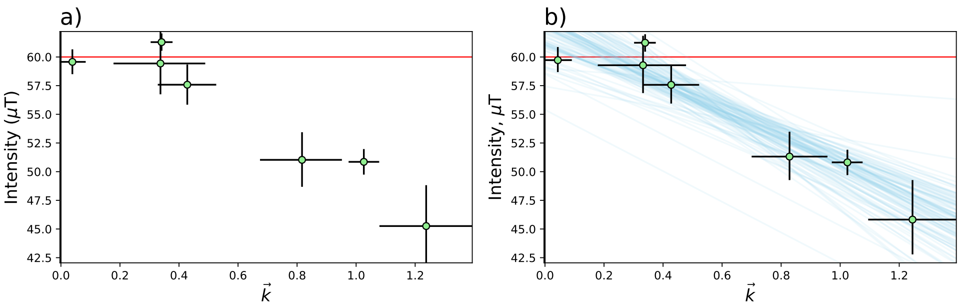

Figure S30. The same methodology described in the caption for Figure S3 applied to site Synthetic60

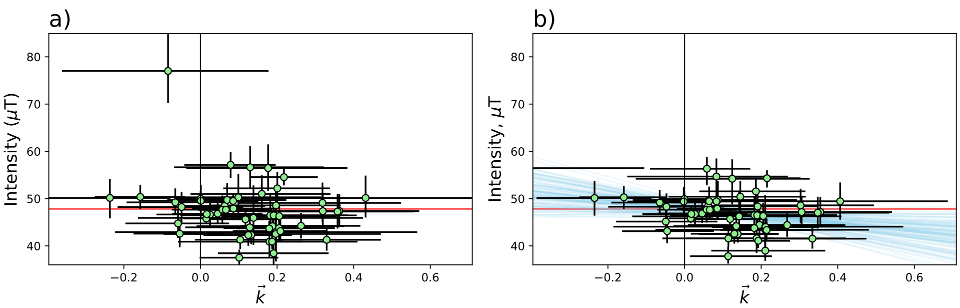

Figure S31. The same methodology described in the caption for Figure S3 applied to site TS

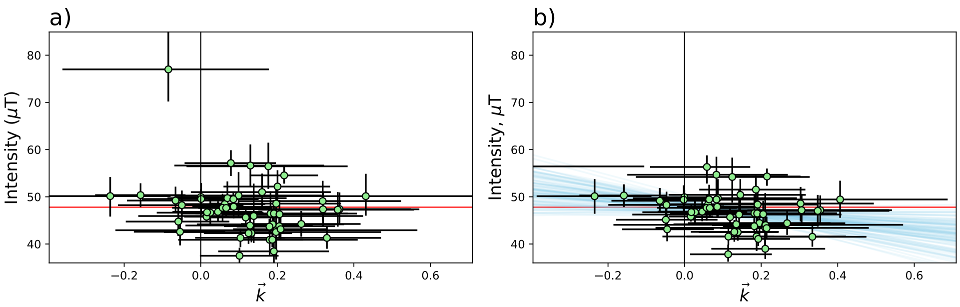

Figure S32. The same methodology described in the caption for Figure S3 applied to site VM