Supplementary S4

1. Statistical Modeling and Model Checking

We modeled the effects of SfM parameterization and processing software on forest inventory metrics derived by UAS-SfM methods using a Bayesian model structured hierarchically to reflect how the data were collected [54]. We used the same model structure for each of the seven forest inventory metrics of interest which served as model dependent variables. However, for each of the seven dependent variables (y) we specified the probability density function of y to reflect the measurement scale of the data [37,39]. Individual tree detection quality as quantified by the F-score metric restricted within the range of [0,1] were modeled using a beta likelihood and a logit link function (equation S1). Basal area absolute percentage error values are strictly non-negative continuous data and were modeled using a Weibull likelihood and a log link function (equation S2). The remaining five dependent variables – basal area percentage error and the mean error and the RMSE for height and DBH – were modeled with a Gaussian likelihood and prior distributions made wide relative to scale and variance of the respective dependent variable (y) data (equation S3). Model fit was checked visually (Figure S7; Figure S8; Figure S9) and quantified using the Bayesian P-value [37] explained in section 4 below.

1.1. Overall Tree Detection (F-Score) Model

We modeled effects of SfM parameterization and processing software on individual tree detection quality as quantified by the F-score metric using a Bayesian hierarchical model shown mathematically in equation S1.

$$y_i \sim Beta(\mu_i, \phi) \tag{S1}$$

$$\begin{aligned} log(\mu_i) = {} & \beta_0 \\ & + \sum_{j=1}^{J=5} \beta_{1[j]} x_{1[j]} + \sum_{k=1}^{K=4} \beta_{2[k]} x_{2[k]} + \sum_{f=1}^{F=3} \beta_{3[f]} x_{3[f]} + \sum_{s=1}^{S=5} \beta_{4[s]} x_{4[s]} \\ & + \sum_{j,k} \beta_{1\times2[j,k]} x_{1\times2[j,k]} + \sum_{j,f} \beta_{1\times3[j,f]} x_{1\times3[j,f]} + \sum_{k,f} \beta_{2\times3[k,f]} x_{2\times3[k,f]} \\ & + \sum_{j,k,f} \beta_{1\times2\times3[j,k,f]} x_{1\times2\times3[j,k,f]} \end{aligned}$$

$$\begin{gathered} \beta_0 \sim Normal(0,1) \\ \beta_{1[j]} \sim Normal\left(0, \sigma_{\beta_1}\right) \\ \beta_{2[k]} \sim Normal\left(0, \sigma_{\beta_2}\right) \\ \beta_{3[f]} \sim Normal\left(0, \sigma_{\beta_3}\right) \\ \beta_{4[s]} \sim Normal\left(0, \sigma_{\beta_4}\right) \\ \beta_{1\times2[j,k]} \sim Normal\left(0, \sigma_{\beta_{1\times2}}\right) \\ \beta_{1\times3[j,f]} \sim Normal\left(0, \sigma_{\beta_{1\times3}}\right) \\ \beta_{2\times3[k,f]} \sim Normal\left(0, \sigma_{\beta_{2\times3}}\right) \\ \sigma_{\beta_1,\ldots,}\sigma_{\beta_{2\times3}} \sim Student(3,0,2.5) \\ \phi \sim Gamma(0.1,0.1) \end{gathered}$$

where y_i is the F-score of the i^{th} individual observation on study site s (five study sites) corresponding to the j^{th} depth map generation quality setting (with five possible levels) and the k^{th} depth map filtering mode setting (with four possible levels) within a processing software f (three software platforms). Observed F-score values (y_i), which were restricted within the range of [0,1], were modeled as beta distributed random variables using a beta likelihood and a logit link function with the centrality parameter μ and dispersion parameter φ. In our study, each study site location s (i.e. "subject") contributes data to multiple levels of the predictors and provides only a single datum per condition. This type of study design is commonly referred to as a "within-subject" design with repeated measures only across cells and not within cells. We account for this design structure in our model by assuming a main effect of study site, but no interaction of study site with other predictors such that the study site effect is constant across SfM

parameterization conditions and the SfM parameterization effects are constant across study sites [39,40]. All parameters have broad priors so that the priors have minimal influence on the posterior [39,40]. The baseline parameter, β_0, and the group deflection parameters, $\beta_{1[j]}$… $\beta_{2\times3[k,f]}$, are given a normal prior distribution that has a mean of zero. The dispersion of the distribution of deflection parameters (φ) and the standard deviation of the group deflection parameters, $\sigma_{\beta1}$… $\sigma_{\beta2\times3}$, were given only weakly informative priors based on default settings in the package brms in order to influence results as little as possible [36]. We present the prior distributions for this model in Figure S7.

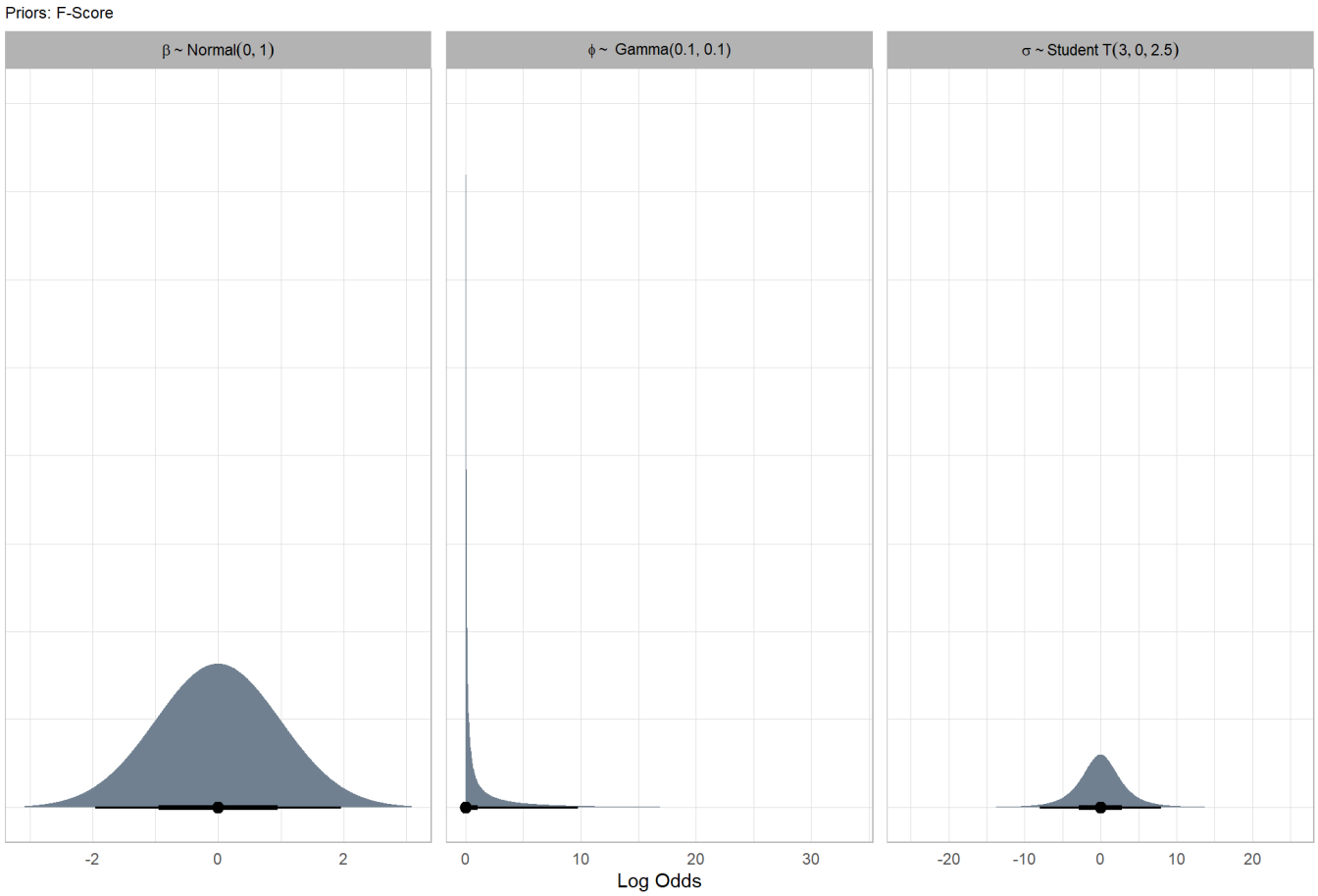


Figure S7. Prior distributions in our Bayesian model of the F-score metric. Broad priors were selected so that the priors have minimal influence on the posterior [39].

Trace-plots were utilized to visually assess convergence of the Markov chains with sufficient convergence supported by R-hat values near 1 [40] (Figure S8). To check the stability of the estimates from the MCMC chains for evidence that the chains have high resolution we used the effective sample size (ESS) indicator with reasonably stable estimates indicated by ESS values over or near 10,000 [41] (Figure S9). Posterior predictive checks were used to evaluate model goodness-of-fit by comparing data simulated from the model with the observed data used to estimate the model parameters [37] (Figure S10).

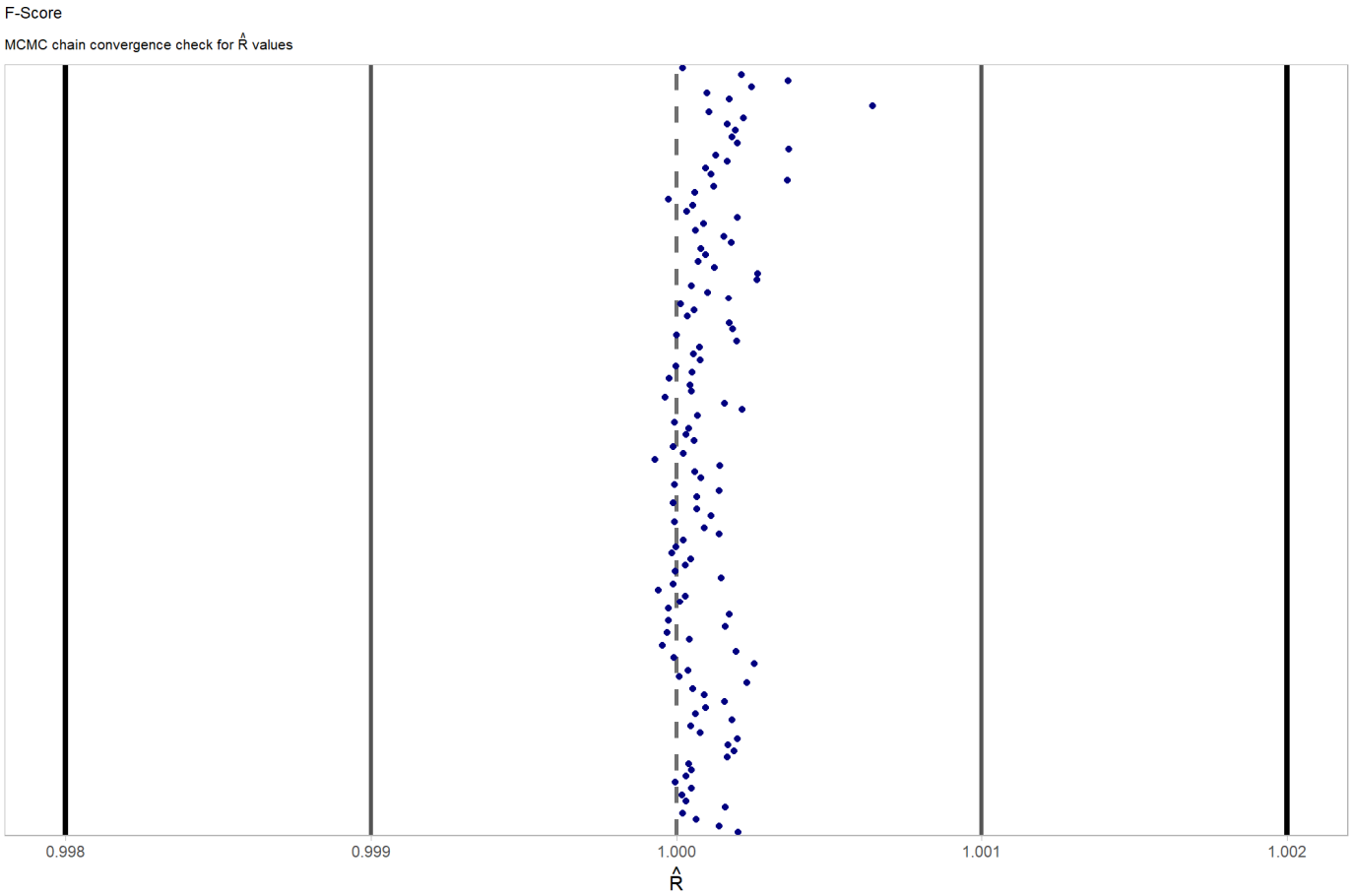


Figure S8. MCMC chain convergence evidence using the R-hat convergence statistic for every model parameter in our Bayesian model of the F-score metric. Values near 1 indicate sufficient convergence of the Markov chains [41].

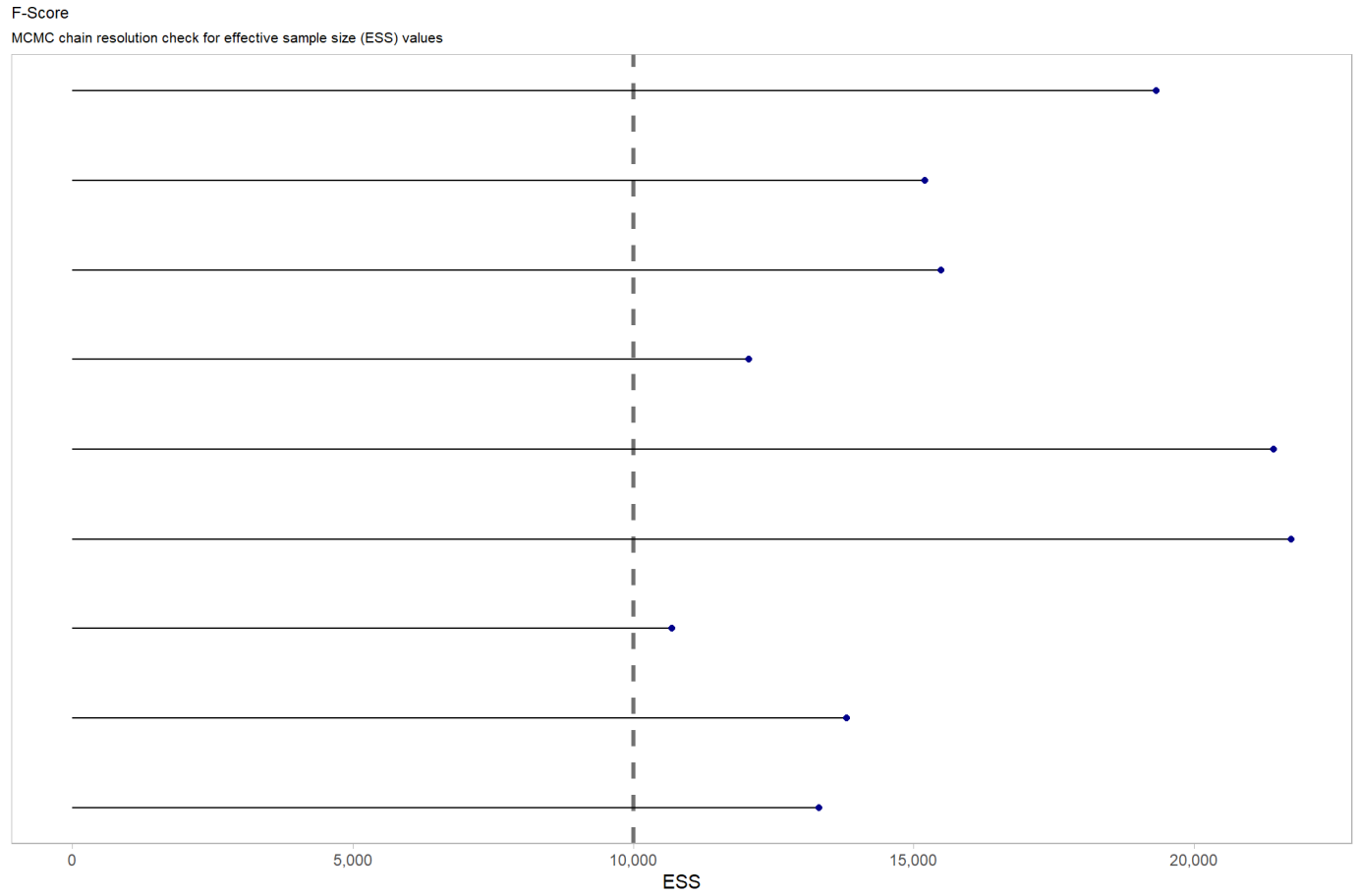


Figure S9. MCMC chain resolution evidence using the effective sample size (ESS) indicator for the baseline parameter and group level deflection parameters in our Bayesian model of the F-score metric where values over or near 10,000 indicate stable parameter estimates from the Markov chains [40].

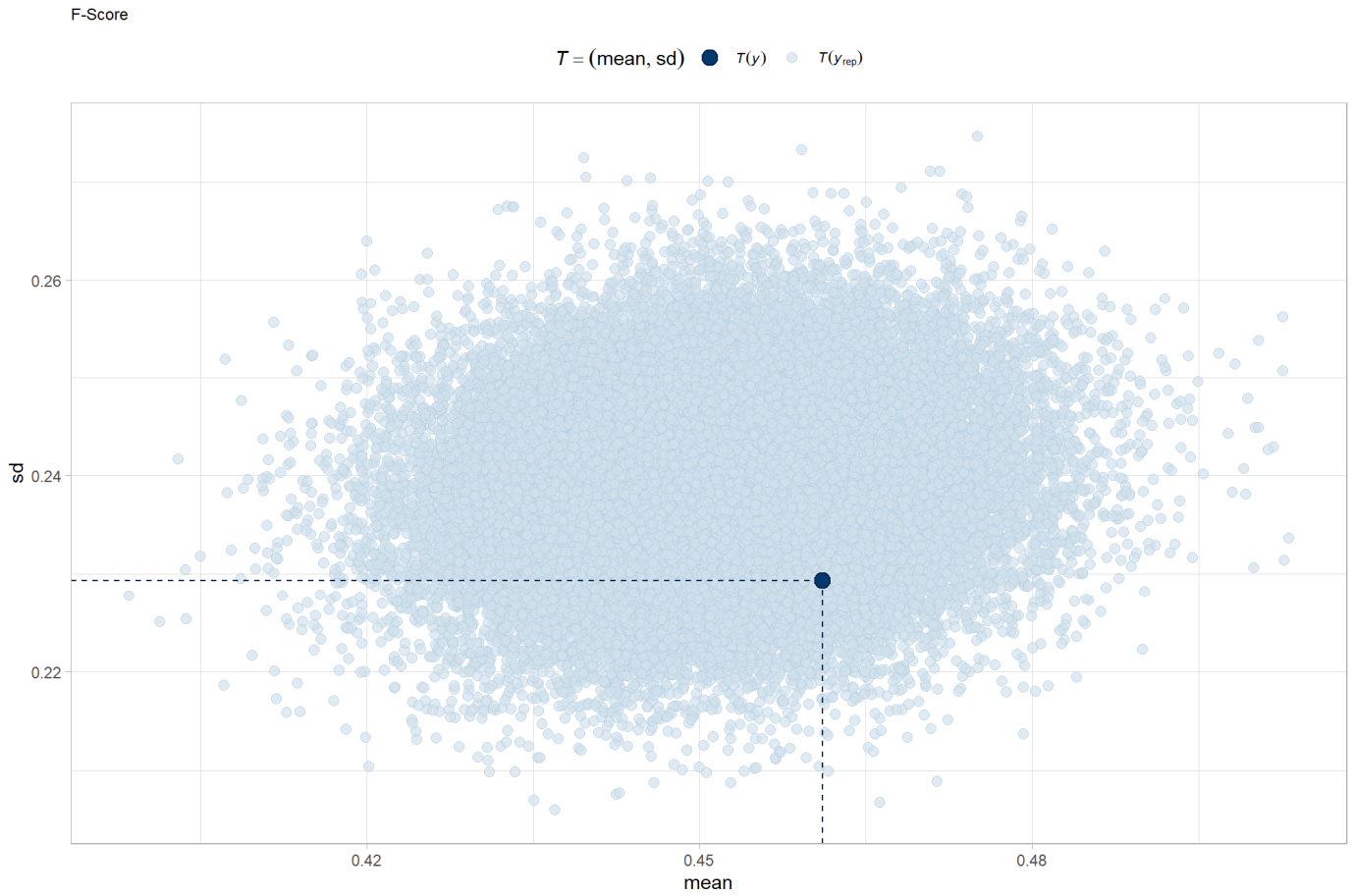


Figure S10. Graphical posterior predictive check providing evidence that our Bayesian model of the F-score metric usefully mimics the data. Where the test statistic of the observed data *T(y)* shown in dark blue is overlaid on the test statistic from the simulated data $T(y_{rep})$ shown in light blue.

2. Stand Basal Area

We modeled effects of SfM parameterization and processing software on stand basal area accuracy compared to field measured data using equation S2 for the absolute percentage error (precision) response variable and equation S3 for the percentage error (bias) response variable.

$$y_i \sim Weibull(\mu_i, \alpha) \tag{S2}$$

$$\begin{aligned} log(\mu_i) = \\ & \beta_0 \\ & + \sum_{j=1}^{J=5} \beta_{1[j]} x_{1[j]} + \sum_{k=1}^{K=4} \beta_{2[k]} x_{2[k]} + \sum_{f=1}^{F=3} \beta_{3[f]} x_{3[f]} + \sum_{s=1}^{S=5} \beta_{4[s]} x_{4[s]} \\ & + \sum_{j,k} \beta_{1\times2[j,k]} x_{1\times2[j,k]} + \sum_{j,f} \beta_{1\times3[j,f]} x_{1\times3[j,f]} + \sum_{k,f} \beta_{2\times3[k,f]} x_{2\times3[k,f]} \\ & + \sum_{j,k,f} \beta_{1\times2\times3[j,k,f]} x_{1\times2\times3[j,k,f]} \end{aligned}$$

$$\begin{aligned} \beta_0 &\sim Normal(-2.91, 1.77) \\ \beta_{1[j]} &\sim Normal(0, \sigma_{\beta_1}) \\ \beta_{2[k]} &\sim Normal(0, \sigma_{\beta_2}) \\ \beta_{3[f]} &\sim Normal(0, \sigma_{\beta_3}) \\ \beta_{4[s]} &\sim Normal(0, \sigma_{\beta_4}) \end{aligned}$$

$$\beta_{1\times2[j,k]} \sim Normal\left(0, \sigma_{\beta_{1\times2}}\right)$$
$$\beta_{1\times3[j,f]} \sim Normal\left(0, \sigma_{\beta_{1\times3}}\right)$$
$$\beta_{2\times3[k,f]} \sim Normal\left(0, \sigma_{\beta_{2\times3}}\right)$$
$$\sigma_{\beta_1,\ldots,}\sigma_{\beta_{2\times3}} \sim Student(3,0,2.5)$$
$$\alpha \sim Gamma(0.1,0.1)$$

where y_i is the basal area absolute percentage error of the i^{th} individual observation on study site s (five study sites) corresponding to the j^{th} depth map generation quality setting (with five possible levels) and the k^{th} depth map filtering mode setting (with four possible levels) within a processing software f (three software platforms). Observed basal area absolute percentage error values (equation S2), with continuous values that are all positive and a lower-bound of zero representing cases where UAS values perfectly match field measurements, were modeled using a Weibull likelihood and a log link function parameterized in terms of the mean (μ) and the shape (α). Basal area percentage error values (equation S3) can take on continuous positive values or negative values with a lower limit of -1 (in cases with zero UAS trees detected and >0 field identified trees) with values close to zero indicating low measurement bias were modeled using a Gaussian likelihood function parameterized in terms of the mean (μ) and the standard deviation (σ). The shape of the Weibull distribution (α) in equation S2, the dispersion of the Gaussian distribution (σ) in equation S3, and the standard deviation of the group deflection parameters ($\sigma_{\beta1}\ldots\ \sigma_{\beta2\times3}$) were given only weakly informative priors based on default settings in the package brms in order to influence results as little as possible [36].

In our study, each study site location s (i.e. "subject") contributes data to multiple levels of the predictors and provides only a single datum per condition. This type of study design is commonly referred to as a "within-subject" design with repeated measures only across cells and not within cells. We account for this design structure in our model by assuming a main effect of study site, but no interaction of study site with other predictors such that the study site effect is constant across SfM parameterization conditions and the SfM parameterization effects are constant across study sites [39,40]. All parameters have broad priors so that the priors have minimal influence on the posterior [39,40]. To define the baseline parameter, β_0, we follow Kruschke [39] and let the data serve as a proxy for setting prior distributions "with the goal to achieve scale invariance, so that whatever is the measurement scale of the data, the prior will be broad and noncommittal on that scale" (p. 560-561). The group deflection parameters, $\beta_{1[j]}\ldots\ \beta_{2\times3[k,f]}$, are given a normal prior distribution that has a mean of zero. We present the prior distributions for the stand basal area accuracy metric models in Figure S11.

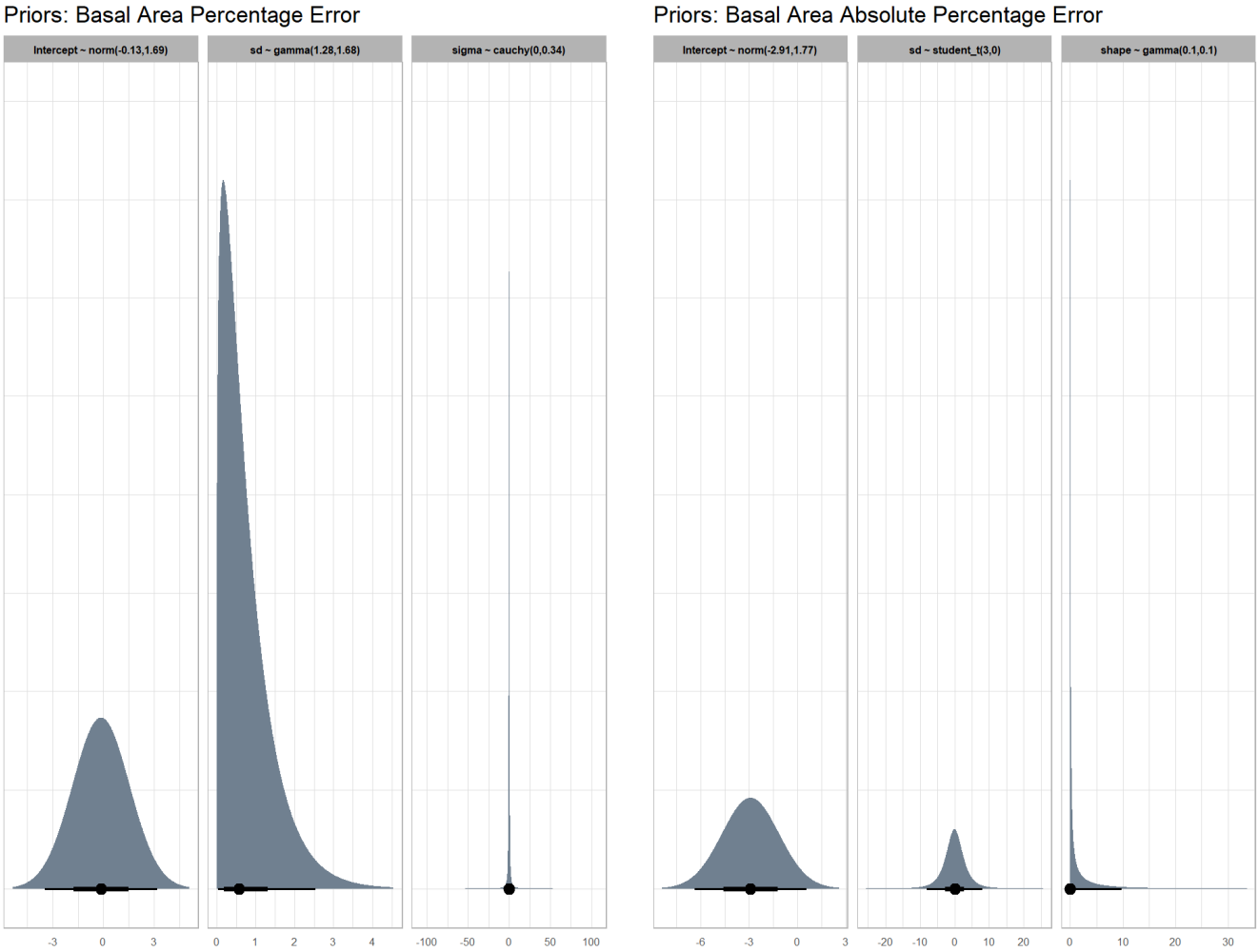


Figure S11. Prior distributions in our Bayesian models of the basal area percentage error metric and the basal area absolute percentage error metric. Broad priors were selected so that the priors have minimal influence on the posterior [39].

Trace-plots were utilized to visually assess convergence of the Markov chains with sufficient convergence supported by R-hat values near 1 [41] (Figure S12). To check the stability of the estimates from the MCMC chains for evidence that the chains have high resolution we used the effective sample size (ESS) indicator with reasonably stable estimates indicated by ESS values over or near 10,000 [39] (Figure S13). Posterior predictive checks were used to evaluate model goodness-of-fit by comparing data simulated from the model with the observed data used to estimate the model parameters [37] (Figure S14).

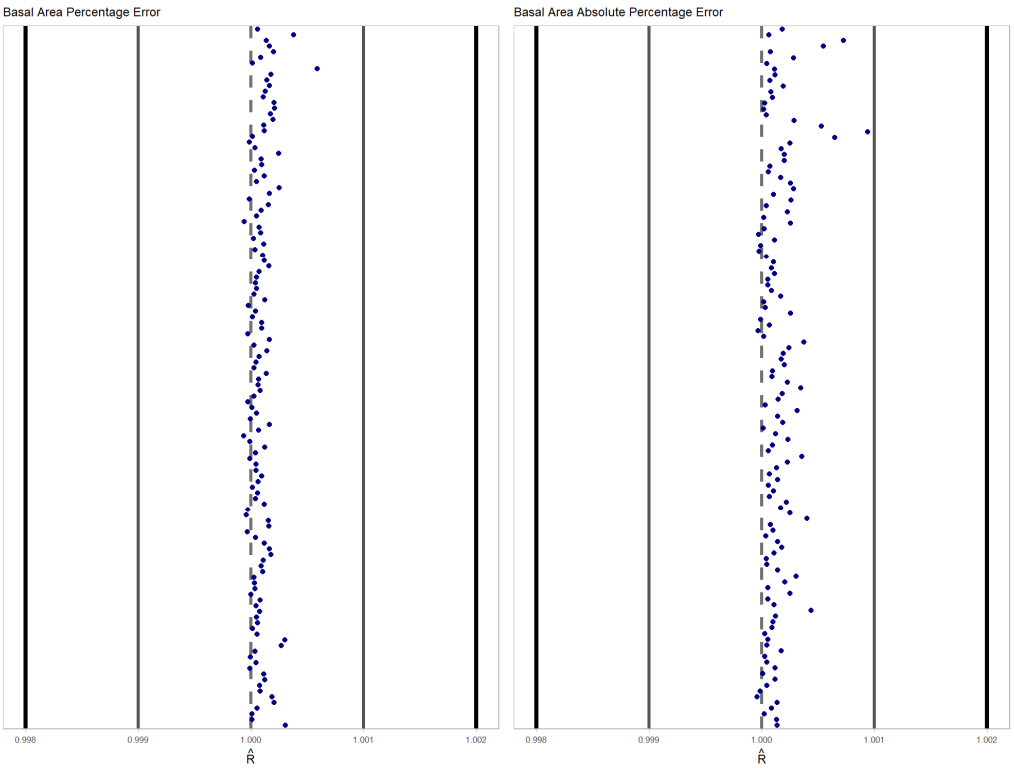


Figure S12. MCMC chain convergence evidence using the R-hat convergence statistic for every model parameter in our Bayesian models of the basal area percentage error metric and the basal area absolute percentage error metric. Values near 1 indicate sufficient convergence of the Markov chains [41].

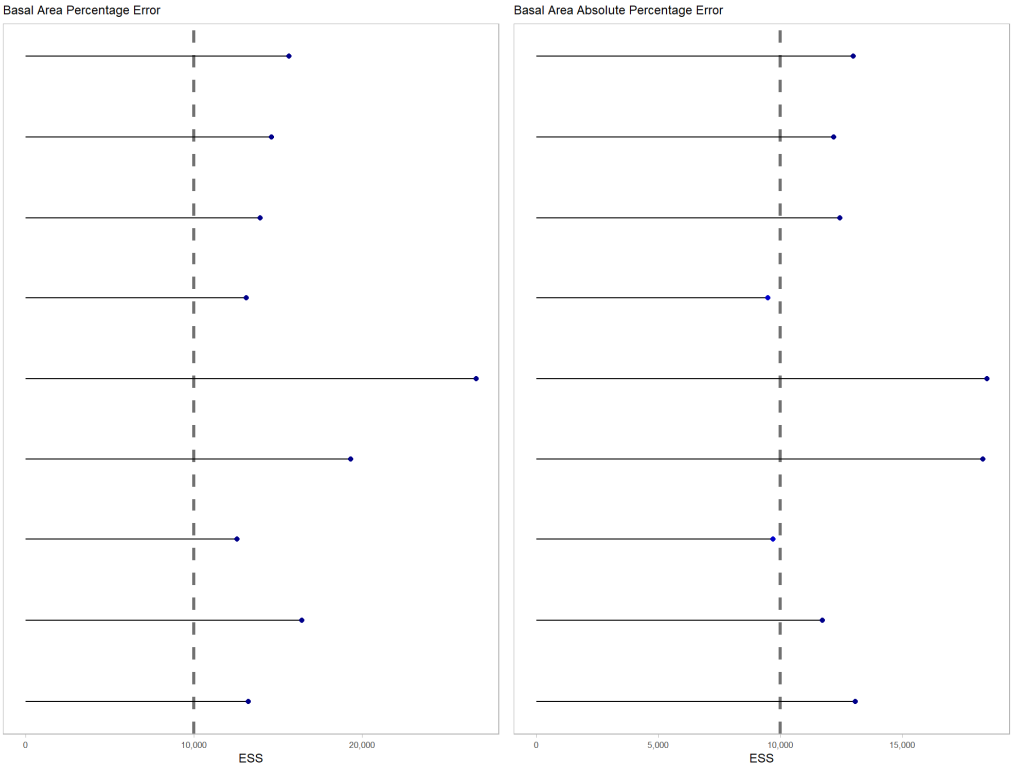


Figure S13. MCMC chain resolution evidence using the effective sample size (ESS) indicator for the baseline parameter and group level deflection parameters in our Bayesian models of the basal area percentage error metric and the basal area absolute percentage error metric. Values over or near 10,000 indicate stable parameter estimates from the Markov chains [39].

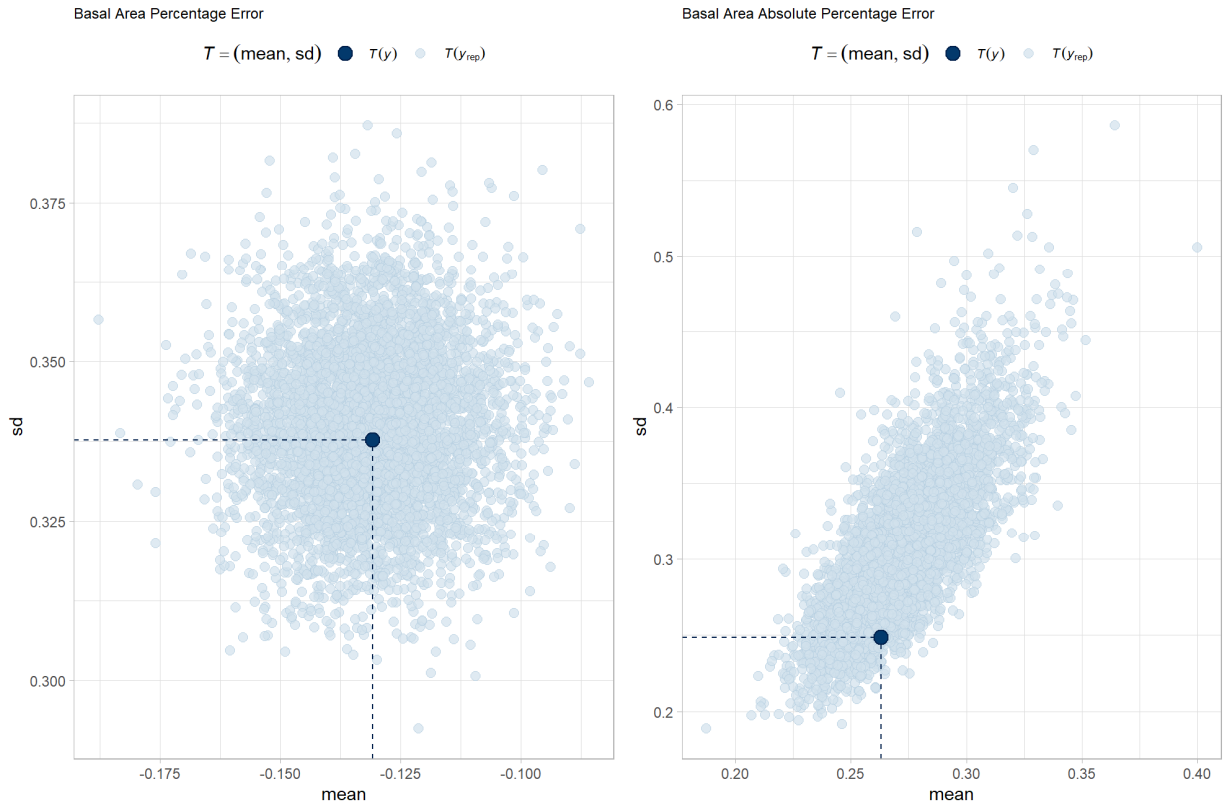


Figure S14. Graphical posterior predictive check providing evidence that our Bayesian models of the basal area percentage error metric and the basal area absolute percentage error metric usefully mimic the data. Where the test statistic of the observed data *T(y)* shown in dark blue is overlaid on the test statistic from the simulated data $T(y_{rep})$ shown in light blue.

3. Tree Height and DBH Accuracy

We modeled the basal area percentage error, and the mean error and RMSE for height and DBH of the True Positive trees with a Gaussian likelihood using five separate models with the same specification shown mathematically in equation S3 but modified prior distributions to be made very wide relative to scale and variance of the data.

$$y_i \sim Normal(\mu_i, \sigma_y) \tag{S3}$$

$$\begin{aligned}\mu_i = {} & \beta_0 \\ & + \sum_{j=1}^{J=5} \beta_{1[j]} x_{1[j]} + \sum_{k=1}^{K=4} \beta_{2[k]} x_{2[k]} + \sum_{f=1}^{F=3} \beta_{3[f]} x_{3[f]} + \sum_{s=1}^{S=5} \beta_{4[s]} x_{4[s]} \\ & + \sum_{j,k} \beta_{1\times2[j,k]} x_{1\times2[j,k]} + \sum_{j,f} \beta_{1\times3[j,f]} x_{1\times3[j,f]} + \sum_{k,f} \beta_{2\times3[k,f]} x_{2\times3[k,f]} \\ & + \sum_{j,k,f} \beta_{1\times2\times3[j,k,f]} x_{1\times2\times3[j,k,f]}\end{aligned}$$

$$\beta_0 \sim Normal(\bar{y}, y_{SD} \times 5)$$
$$\beta_{1[j]} \sim Normal(0, \sigma_{\beta_1})$$
$$\beta_{2[k]} \sim Normal(0, \sigma_{\beta_2})$$
$$\beta_{3[f]} \sim Normal(0, \sigma_{\beta_3})$$
$$\beta_{4[s]} \sim Normal(0, \sigma_{\beta_4})$$
$$\beta_{1\times2[j,k]} \sim Normal(0, \sigma_{\beta_{1\times2}})$$

$$\beta_{1\times3[j,f]} \sim Normal\left(0, \sigma_{\beta_{1\times3}}\right)$$

$$\beta_{2\times3[k,f]} \sim Normal\left(0, \sigma_{\beta_{2\times3}}\right)$$

$$\sigma_{\beta_1,\ldots,}\sigma_{\beta_{2\times3}} \sim Gamma(shape, rate)$$

$$\sigma_y \sim Cauchy(0, y_{SD})$$

where y_i is the basal area percentage error, tree height, or DBH accuracy metric of the i^{th} individual observation on study site s (five study sites) corresponding to the j^{th} depth map generation quality setting (with five possible levels) and the k^{th} depth map filtering mode setting (with four possible levels) within a processing software f (three software platforms). In our study, each study site location s (i.e. "subject") contributes data to multiple levels of the predictors and provides only a single datum per condition. This type of study design is commonly referred to as a "within-subject" design with repeated measures only across cells and not within cells. We account for this design structure in our model by assuming a main effect of study site, but no interaction of study site with other predictors such that the study site effect is constant across SfM parameterization conditions and the SfM parameterization effects are constant across study sites [39,40].

All parameters have broad priors so that the priors have minimal influence on the posterior [39,40]. For each model, we follow Kruschke [39] and let the data serve as a proxy for setting prior distributions "with the goal to achieve scale invariance, so that whatever is the measurement scale of the data, the prior will be broad and noncommittal on that scale" (p. 560-561). This methodology is shown mathematically in equation S3 in which the baseline parameter, β_0, is centered on the observed mean ($\bar{y}$) and made wide based on the standard deviation ($y_{SD} \times 5$) of the observed data. Similarly, the *shape* and *rate* parameters of the gamma distribution used to represent the standard deviation of the group deflection parameters ($\sigma_{\beta1}\ldots\ \sigma_{\beta2\times3}$) are set to be weakly informative using the functions explained by Kruschke [39] in Section 9.2.2 (p. 238) as implemented by Kurz [40]. Prior distributions for each of the four models for the mean error and RMSE for height and DBH are presented in Figure S15.

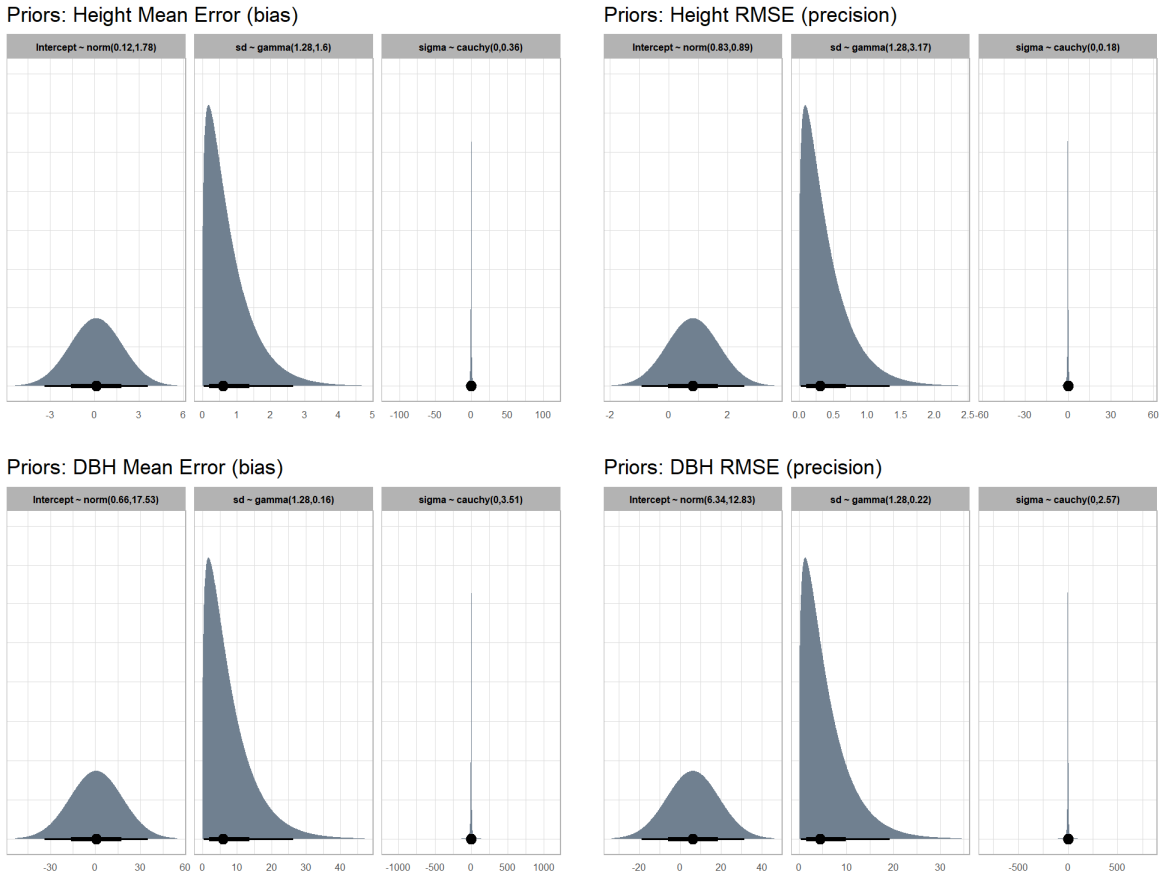


Figure S15. Parameter prior distributions set using the data as a proxy such that the priors have minimal influence on the posterior [39]. For each of the four models, delineated by the plot panel titles, we present the prior distribution for the intercept ($\boldsymbol{\beta}$), the parameter standard deviation ($\sigma_{\beta1}\ldots\ \sigma_{\beta2\times3}$), and the overall model dispersion (σ_y).

Trace-plots were utilized to visually assess convergence of the Markov chains with sufficient convergence supported by R-hat values near 1 [41] (Figure S16). To check the stability of the estimates from the MCMC chains for evidence that the chains have high resolution we used the effective sample size (ESS) indicator with reasonably stable estimates indicated by ESS values over or near 10,000 [39] (Figure S17). Posterior predictive checks were used to evaluate model goodness-of-fit by comparing data simulated from the model with the observed data used to estimate the model parameters [37] (Figure S18).

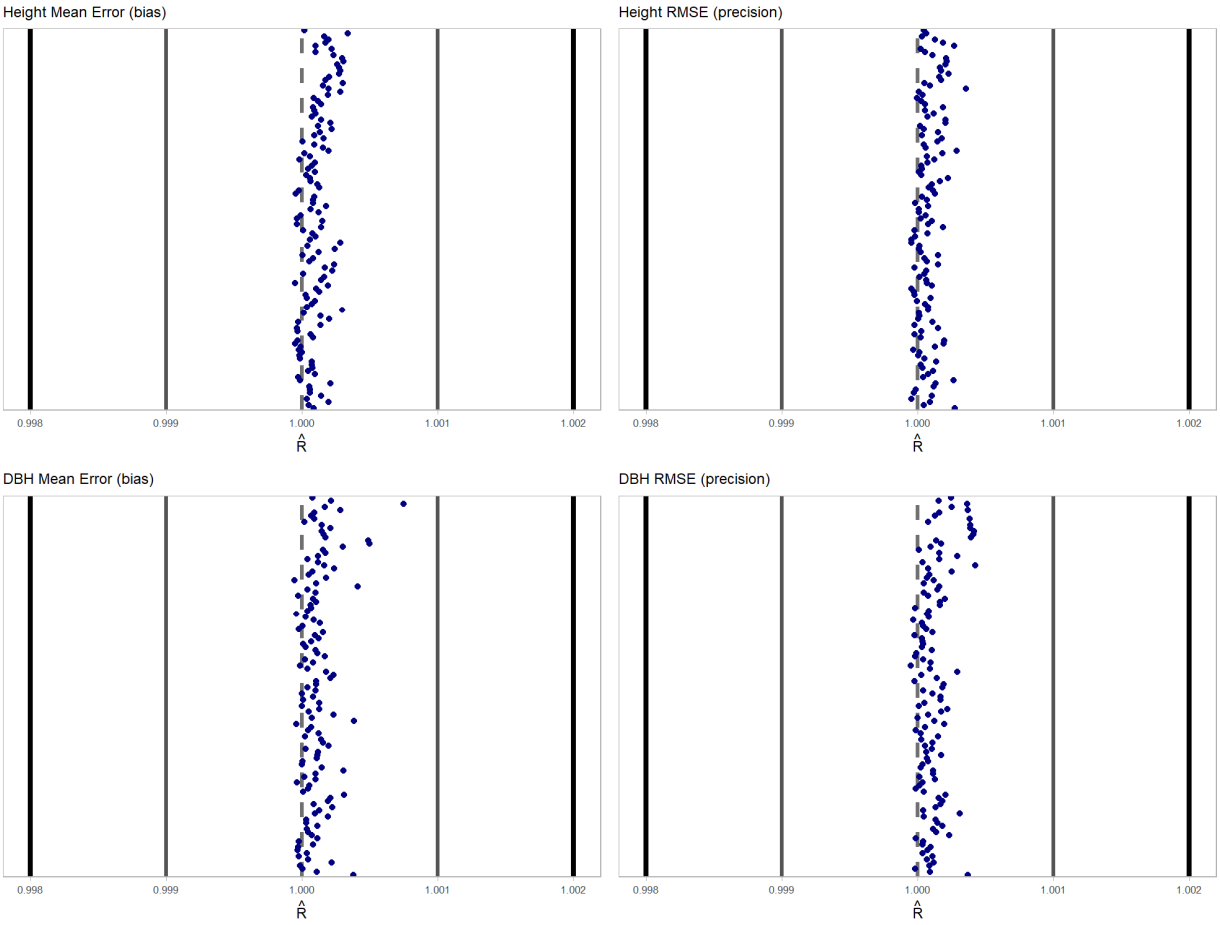


Figure S16. MCMC chain convergence evidence using the R-hat convergence statistic for every model parameter. Values near 1 indicate sufficient convergence of the Markov chains [41].

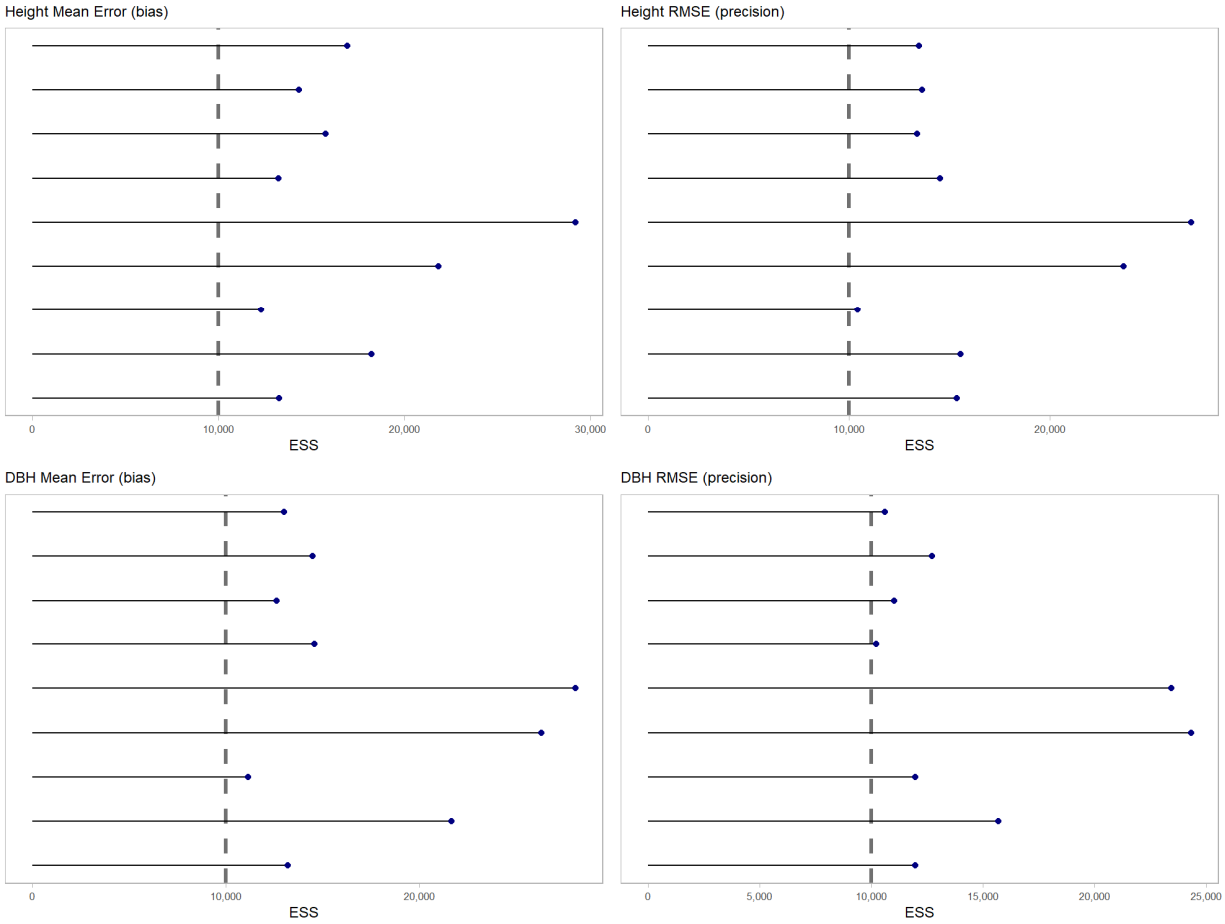


Figure S17. MCMC chain resolution evidence using the effective sample size (ESS) indicator for the baseline parameter and group level deflection parameters where values over or near 10,000 indicate stable parameter estimates from the Markov chains [39].

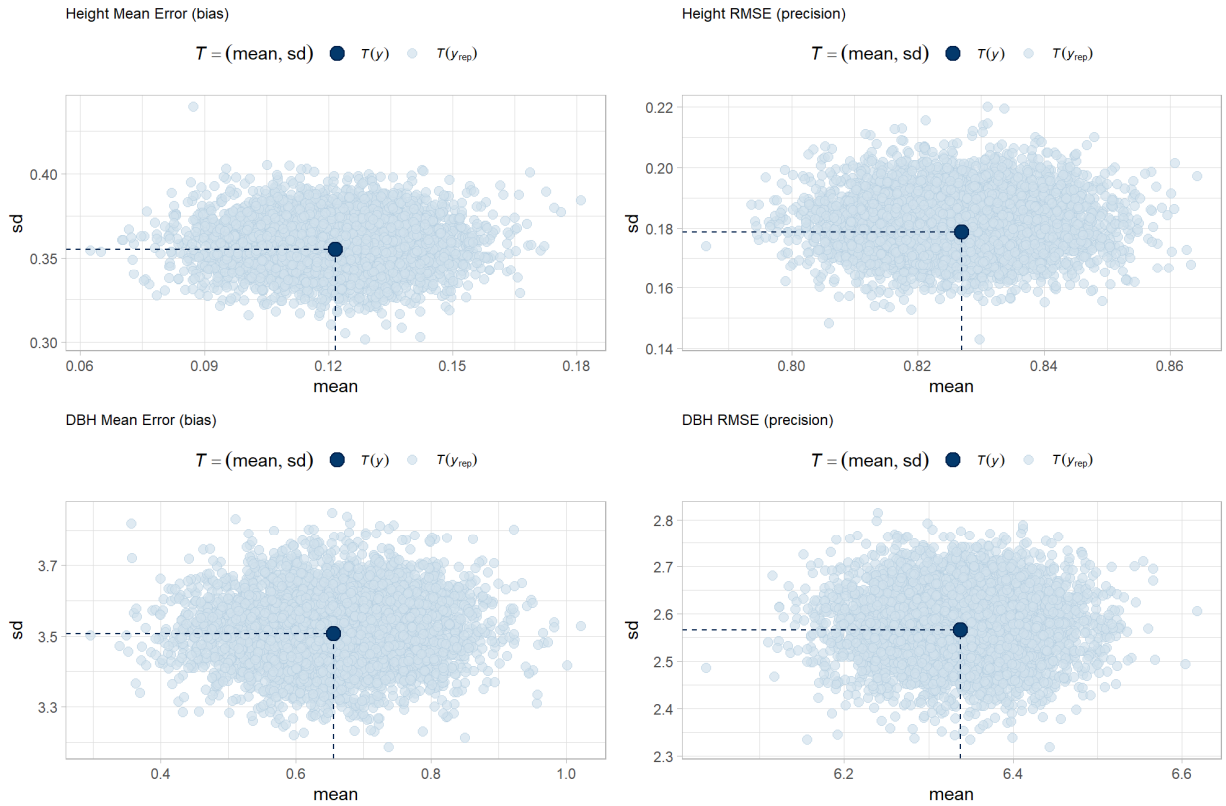


Figure S18. Graphical posterior predictive check providing evidence that the model usefully mimics the data. Where the test statistic of the observed data *T(y)* shown in dark blue is overlaid on the test statistic from the simulated data $T(y_{rep})$ shown in light blue. in light blue.

4. Model Bayesian P-values

Calculating the proportion of MCMC iterations in which the test statistic (we use the mean and standard deviation) from the simulated data and observed data are more extreme than one another provides the Bayesian P-value. Lack of fit is indicated by extreme values (i.e. < 0.10 or > 0.90) while a value of 0.5 indicates perfect fit ([37] pp. 188-190). For each model presented above, we estimated the probability that the test statistic calculated from the posterior predictive distribution arising from our model is more extreme than the test statistic calculated from the distribution of the observed data (Table S31).

Table S31. Bayesian P-values for mean and standard deviation test statistics for each of our seven models of forest inventory metrics derived by UAS-SfM methods. Lack of fit is indicated by a value close to 0 or 1 while a value of 0.5 indicates perfect fit [37].

Model	P mean	P SD
F-Score	0.224	0.881
Basal Area Percentage Error	0.493	0.564
Basal Area Abs. Percentage Error	0.671	0.894
Tree Height Mean Error	0.500	0.575
Tree Height RMSE	0.502	0.628
Tree DBH Mean Error	0.506	0.557
Tree DBH RMSE	0.491	0.557