

<sup>1</sup> Filtering ground noise from LiDAR returns produces  
<sup>2</sup> inferior models of forest aboveground biomass

<sup>3</sup> Michael J Mahoney<sup>a,\*</sup>, Lucas K Johnson<sup>a</sup>, Eddie Bevilacqua<sup>b</sup>, Colin M Beier<sup>b</sup>

<sup>4</sup> *<sup>a</sup>Graduate Program in Environmental Science, State University of New York College of  
Environmental Science and Forestry, 1 Forestry Drive, Syracuse, New York, 13210*

<sup>5</sup> *<sup>b</sup>Department of Sustainable Resources Management, State University of New York College of  
Environmental Science and Forestry, 1 Forestry Drive, Syracuse, New York, 13210*

---

\*Corresponding Author

Email addresses: [mjmahone@esf.edu](mailto:mjmahone@esf.edu) (Michael J Mahoney), [ljohns11@esf.edu](mailto:ljohns11@esf.edu) (Lucas K Johnson), [ebevilacqua@esf.edu](mailto:ebevilacqua@esf.edu) (Eddie Bevilacqua), [cbeier@esf.edu](mailto:cbeier@esf.edu) (Colin M Beier)

---

**8 Abstract**

9 Airborne LiDAR has become an essential data source for large-scale, high-  
10 resolution modeling of forest biomass and carbon stocks, enabling predictions  
11 with much higher resolution and accuracy than can be achieved using optical  
12 imagery alone. Ground noise filtering – that is, excluding returns from LiDAR  
13 point clouds based on simple height thresholds – is a common practice meant to  
14 improve the ‘signal’ content of LiDAR returns by preventing ground returns from  
15 masking useful information about tree size and condition contained within canopy  
16 returns. Although this procedure originated in LiDAR-based estimation of mean  
17 tree and canopy height, ground noise filtering has remained prevalent in LiDAR  
18 pre-processing, even as modelers have shifted focus to forest aboveground biomass  
19 (AGB) and related characteristics for which ground returns may actually contain  
20 useful information about stand density and openness. In particular, ground  
21 returns may be helpful for making accurate biomass predictions in heterogeneous  
22 landscapes that include a patchy mosaic of vegetation heights and land cover  
23 types.

24 In this paper, we applied several ground noise filtering thresholds while  
25 mapping two study areas in New York (USA), one a forest-dominated area and  
26 the other a mixed-use landscape. We observed that removing ground noise via  
27 any height threshold systematically biases many of the LiDAR-derived variables  
28 used in AGB modeling. By fitting random forest models to each of these predictor  
29 sets, we found that that ground noise filtering yields models of forest AGB with  
30 lower accuracy than models trained using predictors derived from unfiltered point  
31 clouds. The relative inferiority of AGB models based on filtered LiDAR returns  
32 was much greater for the mixed land-cover study area than for the contiguously  
33 forested study area. Our results suggest that ground filtering should be avoided

<sup>34</sup> when mapping biomass, particularly when mapping heterogeneous and highly  
<sup>35</sup> patchy landscapes, as ground returns are more likely to represent useful ‘signal’  
<sup>36</sup> than extraneous ‘noise’ in these cases.

---

<sup>37</sup> **Keywords**

- <sup>38</sup> • random forest  
<sup>39</sup> • LiDAR  
<sup>40</sup> • aboveground biomass  
<sup>41</sup> • ground noise  
<sup>42</sup> • machine learning

<sup>43</sup> **1. Introduction**

<sup>44</sup> Accurate assessment of forest carbon stocks for the purposes of greenhouse  
<sup>45</sup> gas accounting and climate change mitigation requires high-resolution maps  
<sup>46</sup> of above-ground biomass (AGB) across large spatial extents. The production  
<sup>47</sup> of these maps has been aided in recent years by the proliferation of publicly  
<sup>48</sup> available airborne LiDAR data, allowing researchers access to granular data  
<sup>49</sup> on land cover heights at fine grained resolutions (Dubayah and Drake, 2000).  
<sup>50</sup> By aggregating returns to a pixel or object level and computing descriptive  
<sup>51</sup> statistics characterizing the distributions of heights of returns, modelers are able  
<sup>52</sup> to convert these point clouds into tabular data formats which may then be used  
<sup>53</sup> to fit regression models for predicting AGB (Hawbaker et al., 2010).

<sup>54</sup> However, there exists some disagreement about precisely which returns to ag-  
<sup>55</sup> gregate when computing these statistics. While some LiDAR-based AGB models  
<sup>56</sup> include all returns when calculating summary statistics (Hudak et al., 2020),  
<sup>57</sup> others first filter out returns below various height thresholds when calculating

percentile heights (Ma et al., 2018), density percentiles (Huang et al., 2019), or their entire suite of predictors (García et al., 2010). Filtering is typically described as being done to remove ground noise from return data, in order to avoid having “ground” returns mask any signal contained in the remaining “canopy” returns. The height threshold used in this process varies across studies, with examples ranging from 0.3 m (García et al., 2010) to 1.3 m (Deo et al., 2017; Ma et al., 2018) to 2 m (Anderson and Bolstad, 2013) to 2.5 m (Huang et al., 2019).

This diversity of approaches demonstrates a lack of consensus about a data processing technique that results in systematically greater estimates of percentile heights and other computed predictors. The practice itself appears to have originated with Nilsson (1996), whose early work with airborne LiDAR focused on calculating tree heights based on the maximum heights of returns, as well as stand volume as a function of the mean height of all returns. Nilsson does not appear to filter returns based on a height threshold; rather, they set the height values of all points below 2 m to 0, in effect reducing the resulting mean height values. The following year, Næsset (1997) published what may be the earliest rationale for ground noise filtering in a study calculating mean stand height from LiDAR returns, excluding returns below 2m in order to avoid interference from shrubs, rocks, and other understory features. In concert, these two studies have provided the justification for filtering out ground returns in a multitude of forest modeling studies (Anderson and Bolstad, 2013; Magnussen and Boudewyn, 1998; Wasser et al., 2013), to the extent that it appears to now be such a commonly accepted practice as to not merit discussion or citation at all (e.g. White et al., 2015; Hawbaker et al., 2010).

Yet this practice, initially justified so as to not include the height of stones in calculating the mean heights of trees (Næsset, 1997), may not be necessary or

desirable as modelers turn their attention to stand characteristics such as AGB. Increased density of ground returns may be associated with sparser stands, and as a result, the left-censoring of variables derived from LiDAR pulses by omitting ground noise may remove useful information about stand structure available for predictive models. This common practice may therefore result in inferior estimates of forest AGB. Filtering may particularly harm predictive accuracy in less contiguously forested and mixed-use landscapes, as we might expect filtering to exclude more returns in areas without tree canopies intercepting and reflecting pulses. As a result, these filtering procedures may adjust LiDAR-derived variables by greater amounts in these settings compared to contiguously forested regions, given their increased proportion of ground returns. It is likely that modeling such heterogeneous landscapes will be an increasing concern over time, as larger data sets and improved computing power enables modelers to map AGB over larger spatial scales; however, there has not been much discussion in the literature concerning any effects filtering may have on forest AGB predictions either in these landscapes or in more homogeneous settings.

In this paper, we use LiDAR data sets representing both continuously forested and mixed-use landscapes to investigate the impacts of ground noise filtering on predictive models of forest AGB. We set out to first identify how filtering ground noise impacts the distribution of commonly used LiDAR-derived predictors, using multiple height thresholds as found throughout the literature. We then fit models to each of these predictor sets using the random forest algorithm (Breiman, 2001), a popular tool used in modeling AGB, to assess how the different predictor distributions may impact model performance. Our results suggest ground filtering is actively detrimental to predictions of AGB, particularly in models that incorporate mixed-use landscapes and areas with only marginal forest cover. These results may help inform future work looking to accurately

<sup>112</sup> predict forest AGB using models incorporating predictors derived from airborne  
<sup>113</sup> LiDAR data products.

<sup>114</sup> **2. Methods**

<sup>115</sup> *2.1. LiDAR Data Sets and Site Characteristics*

<sup>116</sup> In order to identify the impacts of ground filtering on predictive models of  
<sup>117</sup> AGB, we obtained leaf-off LiDAR data sets flown for two regions within New York  
<sup>118</sup> State (Figure 1). The first of these data sets represents the majority of Cayuga  
<sup>119</sup> and Oswego counties in Central New York (New York Office of Information  
<sup>120</sup> Technology Services, 2018), a mixed agricultural and developed landscape with  
<sup>121</sup> small regions of continuous forest and a large amount of marginal forestland  
<sup>122</sup> composed of many small fragments of tree cover (Figure 2). LiDAR data for this  
<sup>123</sup> region was acquired from flights between April and May of 2018 and spans an  
<sup>124</sup> area of 4,455 square kilometers with a nominal pulse spacing of 0.7 meters.

<sup>125</sup> The second data set covers the northern sections of Warren and Washington  
<sup>126</sup> counties and the southern section of Essex County, with smaller inclusions of  
<sup>127</sup> Hamilton and Franklin Counties (New York Office of Information Technology  
<sup>128</sup> Services, 2015). This region (which we refer to as the “Warren, Washington and  
<sup>129</sup> Essex” region) in the northeastern part of the state is largely situated within New  
<sup>130</sup> York’s Adirondack State Park, the largest protected area within the contiguous  
<sup>131</sup> United States (Thorndike, 1999). As a result, this area is predominantly forest  
<sup>132</sup> land, with less developed and agricultural land than Cayuga and Oswego counties  
<sup>133</sup> (Figure 2). LiDAR data was acquired from flights between April and May of  
<sup>134</sup> 2015 and spans an area of 6,278 square kilometers with a nominal pulse spacing  
<sup>135</sup> of 0.556 meters.

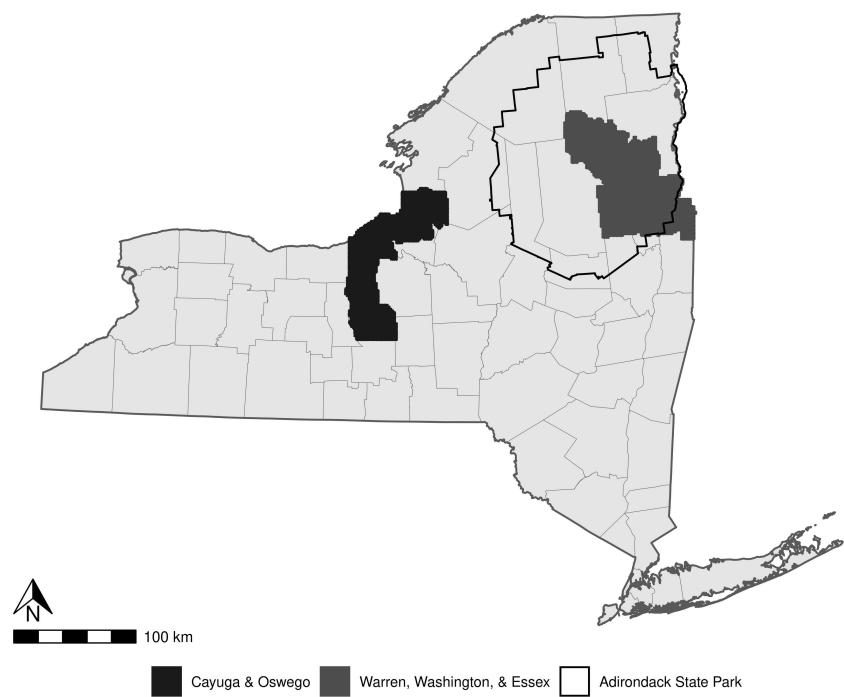


Figure 1: Location of the "Cayuga and Oswego" and "Warren, Washington, and Essex" LiDAR data sets within New York State. The border of the Adirondack State Park is included to show the portion of the Warren, Washington, and Essex data set located within protected lands.

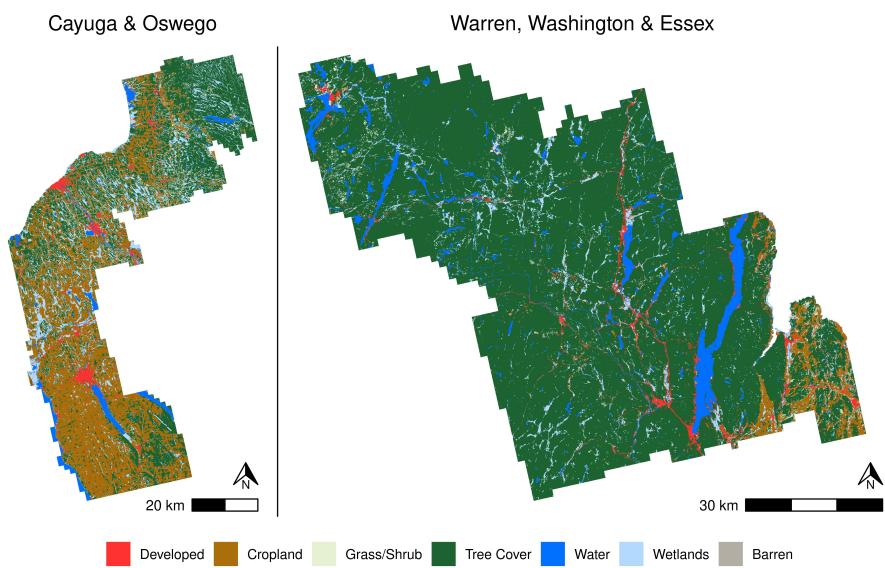


Figure 2: A comparison of land cover across the "Cayuga and Oswego" (left) and "Warren, Washington, and Essex" (right) regions, using land cover classifications from LCMAP (Brown et al., 2020). Colors represent the same land cover categories across both regions, while scale bars differ between regional maps.

136      *2.2. Field Data*

137      Field measurements of AGB for all trees measuring  $\geq 12.7$  cm diameter at  
138      breast height were taken as part of the United States Department of Agriculture  
139      (USDA) Forest Inventory and Analysis (FIA) program (Gray et al., 2012),  
140      with true plot centroid locations obtained under agreement with the USDA.  
141      Measurements were recorded in pounds, then converted and area-normalized  
142      to units of megagrams per hectare ( $Mg\ ha^{-1}$ ). Only FIA plots sampled the  
143      same year as LiDAR flights, or FIA plots with samples from before and after  
144      the LiDAR acquisition year with a difference in AGB within  $[-5\%, \infty)$  were  
145      used for training and evaluating models. In situations where FIA year did not  
146      match LiDAR acquisition year, AGB was calculated by linearly interpolating  
147      between the values measured in the temporally closest FIA samples. Plots  
148      were additionally excluded if any subplots were marked as nonsampled, if FIA  
149      measurements indicated  $0\ Mg\ ha^{-1}$  of AGB but maximum LiDAR return heights  
150      at the plot exceeded 10 meters, or if the convex hull of all LiDAR returns for a  
151      subplot contained less than 90% of the subplot's area. In total, 33 suitable FIA  
152      plots were identified within the Cayuga and Oswego region and 129 within the  
153      Warren, Washington and Essex region, for a total of 162 plots in the combined  
154      data set.

155      *2.3. LiDAR Pre-Processing*

156      A digital terrain model (DTM) was calculated for both sites using a k-  
157      nearest-neighbors inverse-distance weighting imputation algorithm (using  $k =$   
158      5) as implemented in the lidR R package (Roussel et al., 2020), fit using the  
159      points classified as “ground” within the raw LiDAR point cloud data set. The  
160      calculated terrain at each point was then subtracted from the point's z value to  
161      create a height-normalized point cloud. Ground noise filtering rules were then  
162      applied to create five separate points clouds for each site, each representing a

<sup>163</sup> different ground noise filtering approach: one point cloud containing all points  
<sup>164</sup> in the original file (hereafter referred to as “unfiltered”), one removing all points  
<sup>165</sup> classified as “ground” in the original metadata (“ground”), and three removing  
<sup>166</sup> all points with normalized z values below a 0.1, 1, or 2 meter threshold (“0.1m”,  
<sup>167</sup> “1m”, and “2m”, respectively).

<sup>168</sup> Separate sets of 40 predictors, chosen due to their prevalence in published  
<sup>169</sup> models of AGB and forest structure, were derived from each of these point  
<sup>170</sup> clouds using the lidR R package (Table 1) (Hawbaker et al., 2010; Huang et al.,  
<sup>171</sup> 2019; Pflugmacher et al., 2012, 2014; Roussel et al., 2020). Predictors computed  
<sup>172</sup> for FIA plot locations were derived from only the pooled returns coincident  
<sup>173</sup> with the sampled subplot locations, so as to not include any returns from the  
<sup>174</sup> unsampled regions of the macroplot. For plots where ground noise filtering  
<sup>175</sup> resulted in the removal of all points, variables were set to a default value of 0. As  
<sup>176</sup> highly correlated predictor variables may provide the random forest model less  
<sup>177</sup> information for AGB predictions, relationships between predictors were assessed  
<sup>178</sup> using Pearson’s linear correlation coefficient. Changes in predictor distributions  
<sup>179</sup> under different filtering methodologies were assessed using Kolmogorov-Smirnov  
<sup>180</sup> statistics (Massey, 1951).

<sup>181</sup> *2.4. Model Fitting*

<sup>182</sup> AGB models were fit using the ranger R package’s implementation of the  
<sup>183</sup> random forest algorithm (Breiman, 2001; Wright and Ziegler, 2017), a popular  
<sup>184</sup> machine learning technique for predicting forest biomass across landscapes (see  
<sup>185</sup> for instance Huang et al., 2019; Hudak et al., 2020). Separate models were fit  
<sup>186</sup> on predictors calculated using each level of ground noise filtering (“unfiltered”,  
<sup>187</sup> “ground”, “0.1m”, “1m”, and “2m” thresholds) for each LiDAR data set (Cayuga  
<sup>188</sup> and Oswego, Warren, Washington and Essex, and a combination of the two  
<sup>189</sup> regions), for a total of fifteen separate models. Models were fit solely on LiDAR

Table 1: Definitions of LiDAR-derived predictors used for model fitting.

Predictor	Definition
H0, H10, ... H100, H95, H99	Decile heights of returns, in meters, as well as 95th and 99th percentile return heights.
D10, D20... D90	Density of returns above a certain height, as a proportion. After return height is divided into 10 equal bins ranging from 0 to the maximum height of returns, this value reflects the proportion of returns at or above each breakpoint.
N	Number of returns at a given plot or pixel
ZMEAN, ZMEAN_C	Mean height of all returns (ZMEAN) and all returns above 2.5m (ZMEAN_C)
Z_KURT, Z_SKEW	Kurtosis and skewness of height of all returns
QUAD_MEAN, QUAD_MEAN_C	Quadratic mean height of all returns (QUAD_MEAN) and all returns above 2.5m (QUAD_MEAN_C)
CV, CV_C	Coefficient of variation for heights of all returns (CV) and all returns above 2.5m (CV_C)
L2, L3, L4, L_CV, L_SKEW, L_KURT	L-moments and their ratios as defined by Hosking (1990), calculated for heights of all returns
CANCOV	Ratio of returns above 2.5m to all returns (Pflugmacher et al. 2012)
HVOL	CANCOV * ZMEAN (Pflugmacher et al. 2012)
RPC1	Ratio of first returns to all returns (Pflugmacher et al. 2012)

190 derived predictors to ensure differences in model performance resulting from  
191 ground noise filtering were not mediated by the introduction of variables which  
192 might be highly correlated with the unfiltered predictors.

193 Each of these were tuned separately using a standard uniform grid search, with  
194 each model evaluated using the same 17,784 combinations of hyperparameters  
195 detailed in Supplementary Materials S1. The top 100 sets of hyperparameters  
196 for each model, as determined via mean root-mean-squared error (RMSE) from  
197 5-fold cross validation (Stone, 1974) (Equation (1)), were then evaluated using  
198 leave-one-out cross validation (Lachenbruch and Mickey, 1968), with the set of  
199 hyperparameters associated with the lowest RMSE used to fit the final model  
200 reported in the text. This method ensured that each random forest compared  
201 is the best version of the model that could be fit to these predictors, with the  
202 result that any difference in model performance will be due to ground noise  
203 filtering and not stochastic differences between models or effort spent in tuning  
204 hyperparameters. Recent work has suggested cross validation assessments of  
205 model accuracy are likely overoptimistic compared to actual predictive accuracy  
206 (Bates et al., 2021), which does not impact our aim of comparing ground noise  
207 filtering approaches within a single study, but should be kept in mind when  
208 assessing these models as AGB estimators in their own right.

209 All modeling work was done using R version 4.0.5 (R Core Team, 2021).

210 *2.5. Model Assessment*

211 Models were evaluated using multiple metrics calculated via leave-one-out  
212 cross validation (Lachenbruch and Mickey, 1968). Performance metrics calculated  
213 included root-mean-squared error both as a value in Mg ha<sup>-1</sup> (RMSE, equation  
214 (1)) and as a percentage of mean plot AGB (RMSE %, equation (2)), mean  
215 absolute error (MAE, equation (3)), and the coefficient of determination ( $R^2$ ,  
216 equation (4)).

$$\text{RMSE} = \sqrt{\left(\frac{1}{n}\right) \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (1)$$

$$\text{RMSE \%} = 100 \cdot \frac{\text{RMSE}}{\bar{y}} \quad (2)$$

$$\text{MAE} = \left(\frac{1}{n}\right) \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4)$$

217 Where  $n$  is the number of FIA plots included in the data set,  $\hat{y}_i$  is the  
 218 predicted value of AGB,  $y_i$  the AGB value measured at the corresponding  
 219 location, and  $\bar{y}$  the mean AGB value from FIA field measurements.

220 **3. Results**

221 *3.1. Variable Distribution*

222 Filtering out ground noise resulted in notable shifts in predictor distribu-  
 223 tions (Figure 3). Mean predictor values for each ground noise filtering method,  
 224 alongside Kolmogorov-Smirnov test statistic values comparing the distributions  
 225 of filtered predictors to that of the unfiltered predictors, are presented in Sup-  
 226 plementary Materials S2. Filtering returns based upon z-thresholds or ground  
 227 classifications resulted in systematically elevated height percentile and return  
 228 density predictors (the H and D prefixed predictors in Table 1; Figure 3), with  
 229 differences persisting into the highest percentiles calculated. Notable differences  
 230 in distributions also existed for all L-moment based predictors, with increasing  
 231 height thresholds associated with increased magnitude of difference.

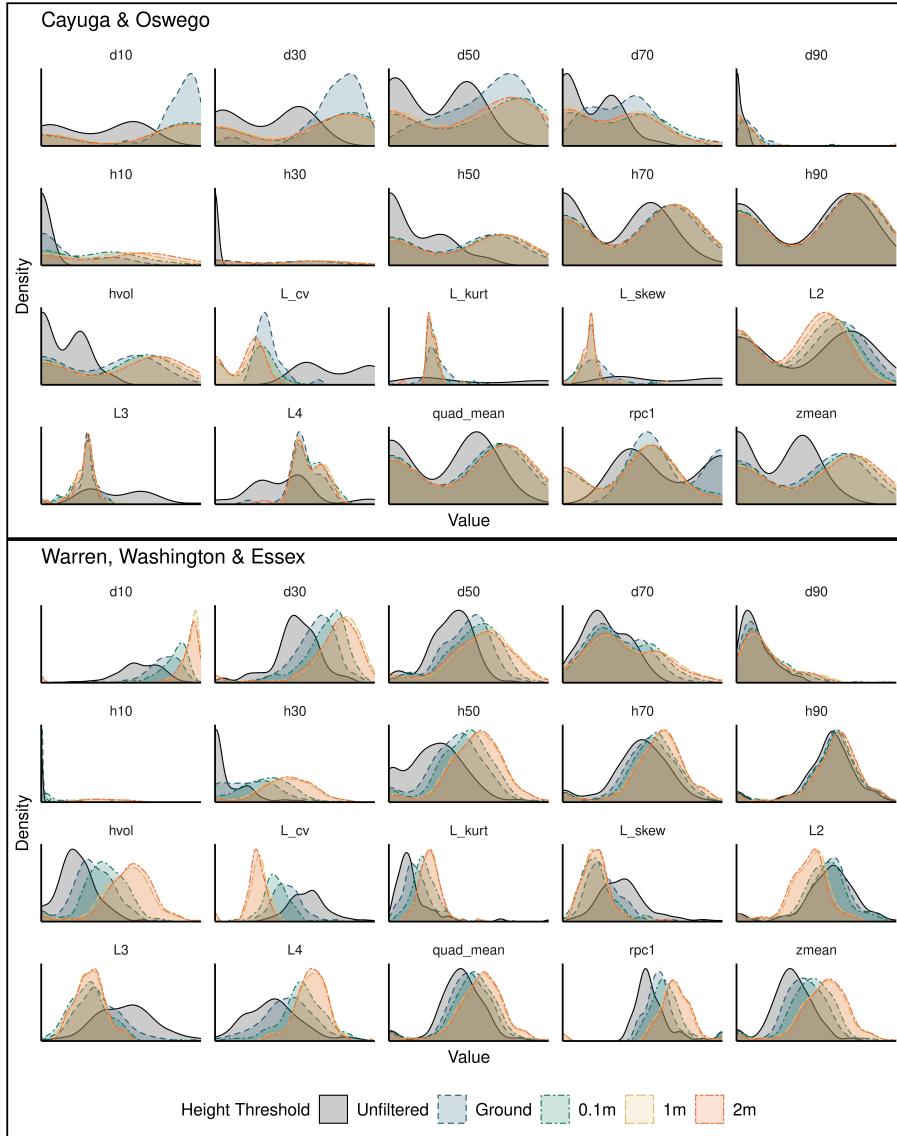


Figure 3: Selected LiDAR-derived predictor distributions for five ground noise filtering approaches. Each subplot is scaled independently so that the X-axis represents the full range of that predictor and the Y-axis represents the full range of the kernel density estimate of that predictor.

Table 2: Mean (with standard deviation in parentheses) Pearson correlation coefficients of LiDAR-derived variables calculated from point clouds processed with five different ground noise filtering methodologies across two separate regions and a combined data set. Variables with standard deviations of 0 after filtering (such as when minimum return height at all plots became 0 due to ground noise filtering) were excluded from calculations.

	Cayuga & Oswego	Warren, Washington & Essex	Combined
Unfiltered	0.262 (0.768)	0.191 (0.571)	0.212 (0.618)
Ground	0.236 (0.590)	0.192 (0.541)	0.200 (0.538)
0.1m	0.553 (0.479)	0.257 (0.538)	0.397 (0.472)
1m	0.611 (0.497)	0.418 (0.463)	0.507 (0.465)
2m	0.574 (0.534)	0.430 (0.463)	0.510 (0.464)

<sup>232</sup> Changing variable distributions resulted in changes to correlation between  
<sup>233</sup> variables, as measured via Pearson correlation coefficients. More aggressive  
<sup>234</sup> filtering approaches were generally associated with stronger positive correlations  
<sup>235</sup> between all variables (Figure 4; Table 2).

<sup>236</sup> *3.2. Model Performance*

<sup>237</sup> Models fit on the unfiltered set of predictors were consistently more accurate  
<sup>238</sup> than those fit to predictors derived from ground noise filtered point clouds,  
<sup>239</sup> both for each region separately as well as in the combined data set (Table 3).  
<sup>240</sup> While differences in model accuracy between the ground noise filtered sets were  
<sup>241</sup> slight, treatments with less aggressive filtering (the ground point removal and 0.1  
<sup>242</sup> meter threshold groups) were generally more accurate than the more aggressively  
<sup>243</sup> filtered predictor sets (1 meter and 2 meter thresholds; Table 3).

<sup>244</sup> Models trained on unfiltered predictors tended to perform better at predicting  
<sup>245</sup> all but the highest AGB plots when compared against those trained on predictors  
<sup>246</sup> calculated after ground noise filtering (Figure 5). Most predictor sets were  
<sup>247</sup> similarly inaccurate when predicting plots with the highest AGB values, a known  
<sup>248</sup> limitation of AGB models built using solely LiDAR-derived predictors (St-Onge

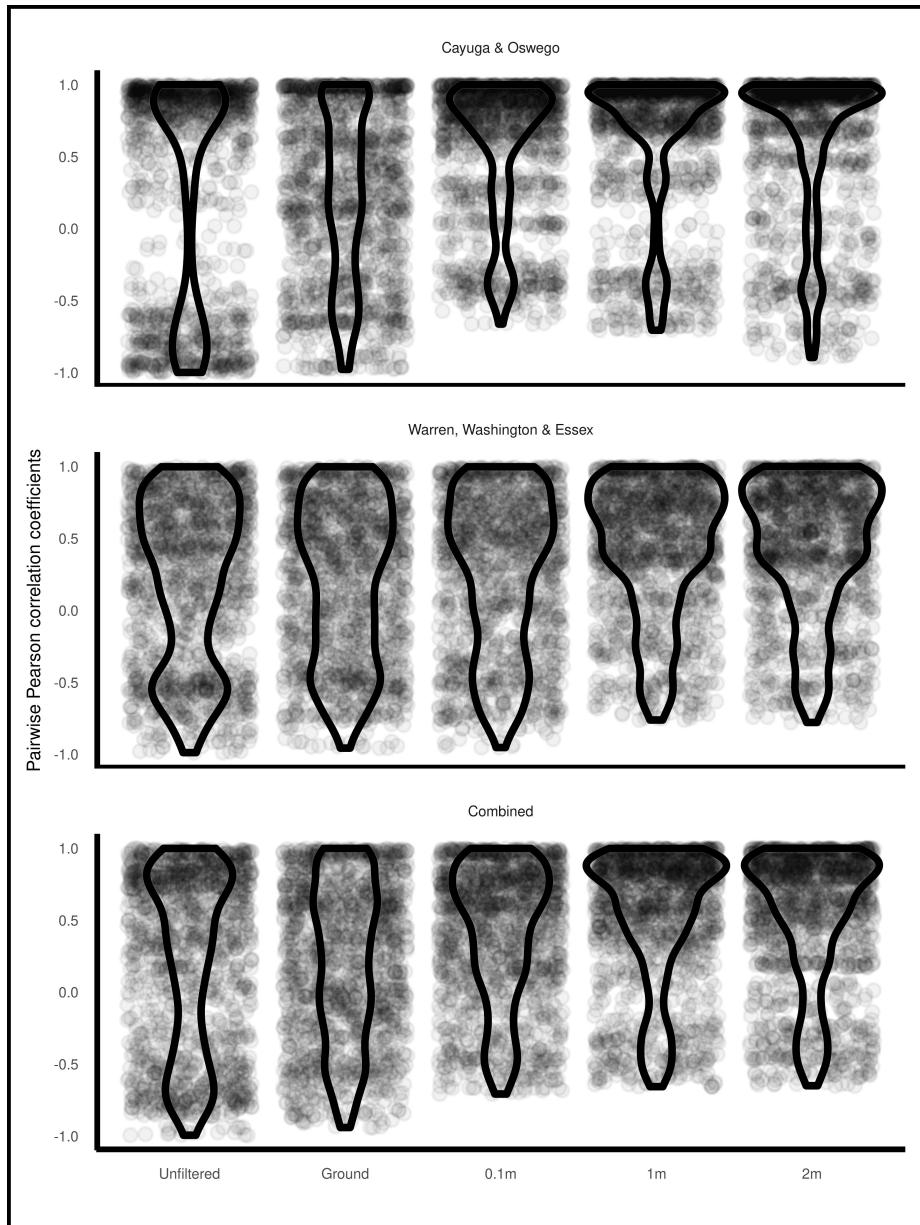


Figure 4: A comparison of the distributions of pairwise Pearson correlation coefficients for each permutation of LiDAR-derived variables examined in this study. Variables with standard deviations of 0 after filtering (such as when minimum return height at all plots became 0 due to filtering) were excluded. Points represent individual correlation coefficients and are slightly transparent, such that darker regions correspond larger densities of correlation coefficients.

Table 3: Model accuracy and agreement metrics as assessed by leave-one-out cross validation.

	Unfiltered	Ground	0.1m	1m	2m
<b>Cayuga &amp; Oswego</b>					
RMSE	23.195	27.243	28.530	28.214	28.853
RMSE (%)	29.934	35.159	36.820	36.412	37.236
MAE	14.698	18.454	20.493	18.566	20.019
R2	0.899	0.863	0.849	0.858	0.846
<b>Warren, Washington &amp; Essex</b>					
RMSE	40.605	41.406	41.882	42.026	43.005
RMSE (%)	31.741	32.367	32.739	32.852	33.617
MAE	30.396	31.749	32.049	31.890	32.587
R2	0.576	0.559	0.544	0.541	0.517
<b>Combined Data</b>					
RMSE	38.140	39.952	40.890	40.412	41.103
RMSE (%)	32.418	33.958	34.755	34.349	34.936
MAE	28.037	29.188	29.567	29.203	30.098
R2	0.681	0.653	0.632	0.641	0.629

<sup>249</sup> et al., 2008). An exception to this pattern was in the Cayuga and Oswego region,  
<sup>250</sup> where models fit using ground noise filtered predictors were particularly poor  
<sup>251</sup> at predicting plots with more than 100 Mg ha<sup>-1</sup> AGB, instead predicting values  
<sup>252</sup> near the mean AGB value for the higher-AGB subgroup (Figure 6). The model  
<sup>253</sup> fit on unfiltered predictors for this region did not exhibit this behavior.

#### <sup>254</sup> 4. Discussion

<sup>255</sup> This study set out to identify an empirical justification for threshold-based  
<sup>256</sup> ground noise filtering for models of forest AGB, given that there exists no clear  
<sup>257</sup> inductive justification for the practice. Instead we found that this common  
<sup>258</sup> practice results in worse models of AGB, with lower predictive accuracy and  
<sup>259</sup> agreement when fit on multiple sets of measured AGB values derived from  
<sup>260</sup> separate LiDAR projects representing both continually forested and mixed-use

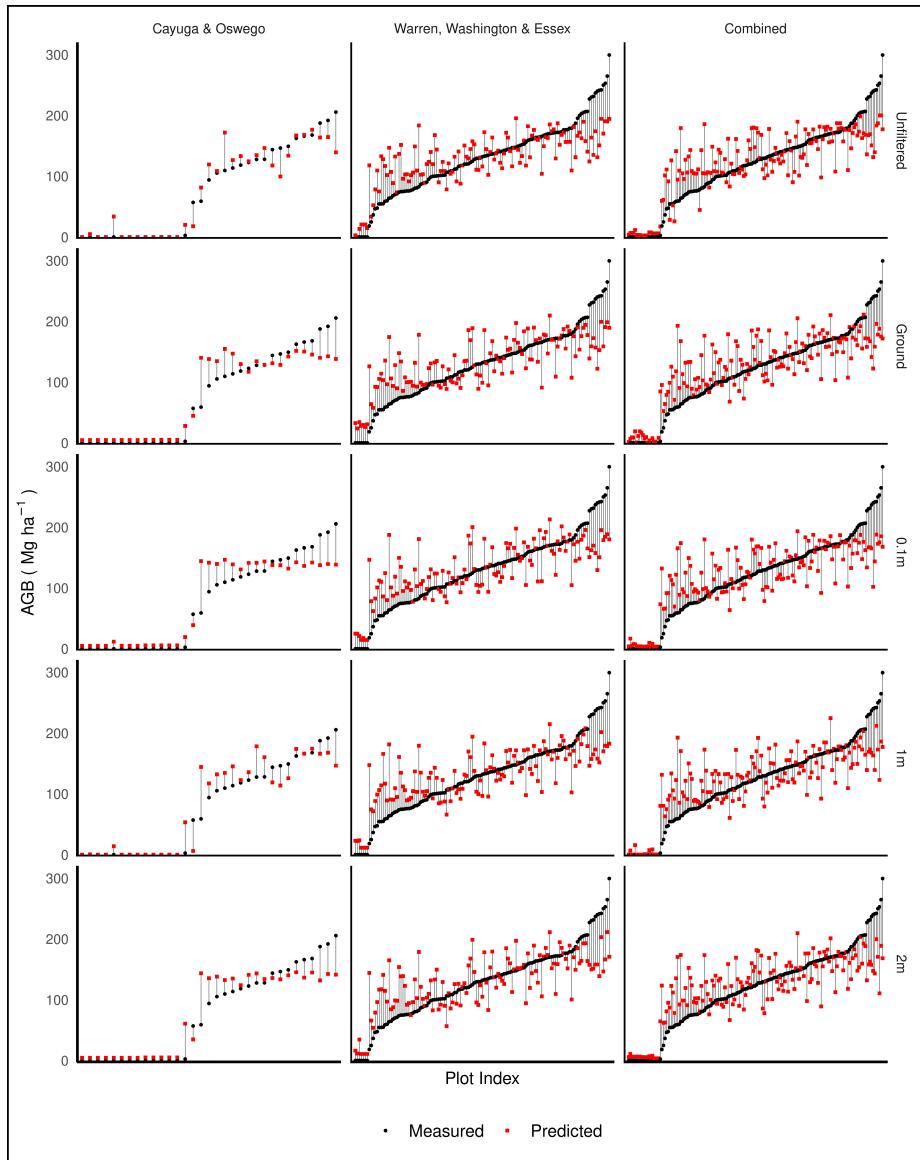


Figure 5: Measured and predicted AGB at each FIA plot for each combination of ground noise filtering approach (rows) and regional data set (columns). Plots are arranged along the X axis by AGB, so that plots with the lowest AGB value in a data set are on the left extreme and those with the highest AGB are on the right, with each plot evenly spaced from its neighbors. The distance between measured AGB (black circles) and predicted AGB (red squares) represents prediction error (black line).

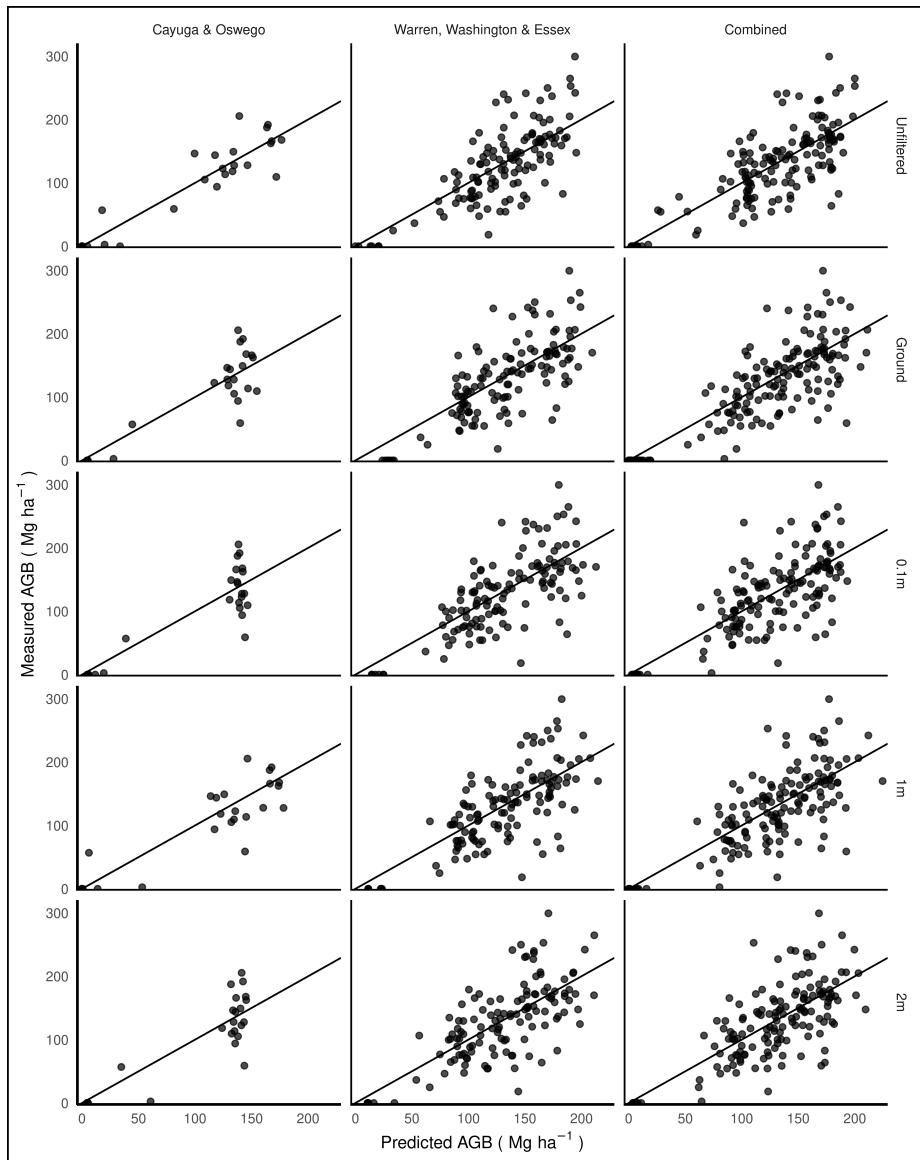


Figure 6: Scatter plots of predicted and measured AGB for five different ground noise filtering approaches (rows) across two regions and a combined data set (columns). A 1:1 relationship is included on each panel as a solid black line.

261 landscapes. These results should encourage future modeling studies to use  
262 unfiltered point clouds when deriving variables for AGB models.

263 *4.1. Ground noise filtering produces inferior predictive models*

264 Our study demonstrates that the ground noise filtering approaches commonly  
265 used in preprocessing data for models of AGB systematically biases LiDAR-  
266 derived variables, with an end result being inferior models that produce less  
267 accurate predictions than models fit on unfiltered data sets (Figure 3, Table 3).  
268 These models fit on filtered data are generally inferior at predicting all but the  
269 highest AGB values relative to their unfiltered counterparts while exhibiting  
270 similar inaccuracy on the higher end of AGB values (Figure 5), likely due to signal  
271 saturation (St-Onge et al., 2008). Increasing intensity of ground noise filtering  
272 was generally, but not universally, associated with worse model performance  
273 (Table 3). Overall, these patterns were strongest in models fit using only data  
274 from a single region.

275 These results are intuitive when thinking about the actual stand character-  
276 istics that may lead to an abundance or lack of ground returns. Dense forest  
277 stands making full use of the available light should be expected to have fewer  
278 returns reaching below the uppermost branches, while landscapes with many  
279 gaps in the canopy will have more such returns. If we conceive of our returns as  
280 providing information about the height structure of the stand as a whole, rather  
281 than about individual trees, it stands to reason that variables calculated using  
282 all returns are more informative about stand metrics such as AGB than those  
283 using filtered point clouds which may sacrifice information about stand openness.  
284 This could explain the impact of ground noise filtering seen in this study using  
285 leaf-off LiDAR; we might expect this impact to be even more pronounced were  
286 we to use leaf-on LiDAR in its place.

287 Our results also make sense mechanistically given the properties of the

random forest algorithm used to construct AGB models in this study. Random forests excel at predicting outcomes based upon the consensus of weak learners (Breiman, 2001), individual decision trees which themselves rely upon small and ephemeral correlations between predictor variables and the outcome of interest. As shown (in Figure 4 and Table 2), ground noise filtering approaches increase positive correlations between predictor variables, with the resulting increased collinearities shrinking the number and magnitude of possible weak correlations between individual variables and AGB (Langford et al., 2001). While the decision trees comprising the random forest may be able to take advantage of the correlations between predictor variables and the outcome to achieve similar accuracy as when trained on unfiltered data sets, we would not expect that a process that uniformly increases the positive linear correlation between variables would be associated with improved predictions.

Insights drawn from these results may not be limited to only machine learning based models. Anderson and Bolstad (2013) note that, when fitting linear models to predict AGB, models based on unfiltered point clouds always provided better results than those fit to predictors calculated using only returns above 2 meters. However, few other AGB modeling studies have performed similar investigations, necessitating our current study. Our conclusions may not apply to AGB models of non-forest systems; investigations of ground noise filtering as a preprocessing step for models of corn AGB found improvements in predictive accuracy with relatively low height thresholds (Luo et al., 2016), emphasizing that commonly accepted data processing practices cannot be assumed to transfer across systems or domains to new questions of interest.

#### *4.2. Differences between regional models*

Although we found that models fit using predictors derived from unfiltered point clouds to be the most accurate across both regions and the combined

315 data set, the degree to which ground noise filtering damaged predictive accuracy  
316 and the relationship between filtering intensity and accuracy varied between  
317 regions. Of particular interest is the degree to which models performed worse  
318 when fit using predictors derived from filtered point clouds within the Cayuga  
319 and Oswego region (Figure 6, Table 3). This region is characterized by large  
320 amounts of marginal forestland spread across a mixed-use landscape, resulting in  
321 a notably higher proportion of plots with no or low AGB and much lower mean  
322 AGB values compared to the Warren, Washington and Essex region. As a result,  
323 the models appear to not have sufficient numbers of observations about these  
324 relatively higher AGB plots to reliably differentiate them once any information  
325 on stand structure conveyed by ground returns is removed. As a result, the  
326 random forest algorithm produces relatively few nodes dedicated to separating  
327 out these observations and instead predicts near the subgroup mean for all plots  
328 with more than 100 Mg ha<sup>-1</sup> AGB.

329 Therefore, ground noise filtering may be more detrimental to models trained  
330 in regions dominated by low-AGB forestlands. While models fit using predic-  
331 tors derived from filtered point clouds were consistently inferior to those using  
332 unfiltered data, the filtering procedures were less detrimental within the contigu-  
333 ously forested Warren, Washington and Essex region than within the mixed-use  
334 Cayuga and Oswego region. We therefore suggest that our results in Cayuga and  
335 Oswego, where ground noise filtering produced a model with an RMSE up to  
336 24% greater than that of the unfiltered model, represent close to the maximum  
337 impact ground noise filtering may have on model performance. Our combined  
338 data set – fit on many more points representing much more contiguous forestland  
339 – is likely more similar to a typical AGB mapping project, and as such we believe  
340 the approximately 5-7% increase in RMSE introduced via ground noise filtering  
341 is closer to the impact that would be seen in multi-region models of AGB.

342 *4.3. Limitations as AGB models*

343 The models discussed in this study were purposefully designed so as to  
344 maximize the potential effect of ground noise filtering on model performance. For  
345 this reason, models were fit using only LiDAR-derived predictors, as predictors  
346 obtained from additional data sources may be correlated with unfiltered predictors  
347 and as such used in their place by the random forest algorithm (Efron, 2020), thus  
348 mediating the impact of the filtering approaches. Additionally, these models were  
349 fit using relatively few field measurements (a total of 162 FIA plots) located across  
350 two spatially disparate regions with varying cover types, with hyperparameter  
351 tuning performed via an automated process so as to avoid unintentionally biasing  
352 results by giving different models differing levels of attention or time in tuning.  
353 Further, model assessment was done using leave-one-out cross validation, which  
354 is sufficient for comparison between individual models but lacking as a way  
355 to characterize model AGB predictions spatially and across multiple scales  
356 (Riemann et al., 2010). While none of these limitations impact the comparison  
357 of ground noise filtering approaches at the center of this study, in combination  
358 they prevent us from using these models to make fine-scale estimates about AGB  
359 stocks across these regions and how model predictions compare to regional FIA  
360 estimates.

361 *4.4. Recommendations for future models*

362 Our results and examination of the literature suggest that ground noise  
363 filtering procedures are not well justified for studies modeling AGB, given  
364 both the potential information lost about stand density and structure, and the  
365 empirical inferiority of models fit using predictors derived from filtered point  
366 clouds. We make no such claim about researchers modeling other variables using  
367 LiDAR-derived predictors; for instance, when modeling mean tree heights similar  
368 to Næsset's (1997) study which originated the practice of ground noise filtering.

369 The best data preprocessing procedure will necessarily depend on the purpose of  
370 the model (Sambasivan et al., 2021).

371 More generally, we recommend our approach to any researcher considering  
372 a new (or reviewing an old) data preprocessing step to include in their model.  
373 While tracing methodological details to their origins in the literature may not  
374 always be fruitful, researchers should ideally have the ability to separate out  
375 small sections of their data to evaluate model performance with and without the  
376 proposed procedure. The results of these small tests may justify including the  
377 procedure in the data preprocessing workflow for the full data set, or alternately  
378 lead a team to remove a processing step to save data cleaning time without  
379 damaging predictive accuracy. In these early days of big data in environmental  
380 science, we remain wanting for a cohesive theory of optimal prediction (Efron,  
381 2020); as a result, beliefs about methodological improvements are still best tested  
382 by experiment.

383 **5. Conclusion**

384 Our study demonstrates that preprocessing LiDAR point clouds to filter  
385 out ground noise may be detrimental when making predictions of above-ground  
386 biomass using machine learning methods. By removing signal of stand density  
387 and structure from LiDAR-derived predictors, ground noise filtering produces  
388 models that are systematically worse at predicting low AGB plots while impairing  
389 the ability of models to accurately capture the variance present in higher AGB  
390 regions. This impact is particularly notable within mixed-use and otherwise  
391 heterogeneous landscapes, given the increased proportion of ground returns  
392 recorded when mapping these areas compared to contiguously forested regions.  
393 Although well-justified in its original context of modeling mean stand heights,  
394 the persistence of ground noise filtering in LiDAR-based AGB modeling appears

395 to produce less accurate predictions than could be achieved using currently  
396 available data.

397 More broadly, this study serves as a reminder that commonly accepted data  
398 preprocessing workflows do not necessarily transcend domains and methodologies.  
399 Noise which may mask information in one modeling application may provide  
400 useful signal when modeling other outcomes, requiring modelers to reevaluate  
401 data transformations when moving between problems and contexts. Whether  
402 such an evaluation is done empirically through comparisons of model performance  
403 or by examining the logical basis for the manipulation (or as presented here,  
404 both), critically assessing data preprocessing pipelines remains an essential task  
405 in the production of accurate and useful models from remotely sensed data  
406 sources.

407 **6. Acknowledgements**

408 We would like to thank the US Forest Service Forest Inventory & Analysis  
409 program for their data sharing and cooperation, the New York State GIS  
410 Program Office for compiling LIDAR data, and the New York State Department  
411 of Environmental Conservation, Office of Climate Change, for funding support.

412      **References**

- 413      Anderson, R.S., Bolstad, P.V., 2013. Estimating Aboveground Biomass and  
414      Average Annual Wood Biomass Increment with Airborne Leaf-on and Leaf-off  
415      LiDAR in Great Lakes Forest Types. Northern Journal of Applied Forestry 30,  
416      16–22. <https://doi.org/10.5849/njaf.12-015>
- 417      Bates, S., Hastie, T., Tibshirani, R., 2021. Cross-validation: What does it  
418      estimate and how well does it do it? ArXiv:2104.00673v2 [stat.ME].
- 419      Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- 420      Brown, J.F., Tollerud, H.J., Barber, C.P., Zhou, Q., Dwyer, J.L., Vogelmann,  
421      J.E., Loveland, T.R., Woodcock, C.E., Stehman, S.V., Zhu, Z., Pengra, B.W.,  
422      Smith, K., Horton, J.A., Xian, G., Auch, R.F., Sohl, T.L., Sayler, K.L., Gallant,  
423      A.L., Zelenak, D., Reker, R.R., Rover, J., 2020. Lessons learned implementing an  
424      operational continuous united states national land change monitoring capability:  
425      The land change monitoring, assessment, and projection (lcmap) approach.  
426      Remote Sensing of Environment 238, 111356. <https://doi.org/10.1016/j.rse.2019.111356>
- 427      Deo, R.K., Russell, M.B., Domke, G.M., Andersen, H.-E., Cohen, W.B.,  
428      Woodall, C.W., 2017. Evaluating site-specific and generic spatial models of  
429      aboveground forest biomass based on landsat time-series and lidar strip samples  
430      in the eastern usa. Remote Sensing 9. <https://doi.org/10.3390/rs9060598>
- 431      Dubayah, R.O., Drake, J.B., 2000. Lidar Remote Sensing for Forestry.  
432      Journal of Forestry 98, 44–46. <https://doi.org/10.1093/jof/98.6.44>
- 433      Efron, B., 2020. Prediction, estimation, and attribution. Journal of the Amer-  
434      ican Statistical Association 115, 636–655. <https://doi.org/10.1080/01621459.2020.1762613>
- 435      García, M., Riaño, D., Chuvieco, E., Danson, F.M., 2010. Estimating

439 biomass carbon stocks for a mediterranean forest in central spain using lidar  
440 height and intensity data. *Remote Sensing of Environment* 114, 816–830. <https://doi.org/10.1016/j.rse.2009.11.021>

442 Gray, A.N., Brandeis, T.J., Shaw, J.D., McWilliams, W.H., Miles, P., 2012.  
443 Forest inventory and analysis database of the united states of america (fia).  
444 *Biodiversity and Ecology* 4, 225–231. <https://doi.org/10.7809/b-e.00079>

445 Hawbaker, T.J., Gobakken, T., Lesak, A., Trømborg, E., Contrucci, K.,  
446 Radeloff, V., 2010. Light Detection and Ranging-Based Measures of Mixed  
447 Hardwood Forest Structure. *Forest Science* 56, 313–326. <https://doi.org/10.1093/forestscience/56.3.313>

449 Hosking, J.R.M., 1990. L-moments: Analysis and estimation of distributions  
450 using linear combinations of order statistics. *Journal of the Royal Statistical  
451 Society. Series B (Methodological)* 52, 105–124.

452 Huang, W., Dolan, K., Swatantran, A., Johnson, K., Tang, H., O’Neil-Dunne,  
453 J., Dubayah, R., Hurt, G., 2019. High-resolution mapping of aboveground  
454 biomass for forest carbon monitoring system in the tri-state region of maryland,  
455 pennsylvania and delaware, USA. *Environmental Research Letters* 14, 095002.  
456 <https://doi.org/10.1088/1748-9326/ab2917>

457 Hudak, A.T., Fekety, P.A., Kane, V.R., Kennedy, R.E., Filippelli, S.K.,  
458 Falkowski, M.J., Tinkham, W.T., Smith, A.M.S., Crookston, N.L., Domke, G.M.,  
459 Corrao, M.V., Bright, B.C., Churchill, D.J., Gould, P.J., McGaughey, R.J., Kane,  
460 J.T., Dong, J., 2020. A carbon monitoring system for mapping regional, annual  
461 aboveground biomass across the northwestern USA. *Environmental Research  
Letters* 15, 095003. <https://doi.org/10.1088/1748-9326/ab93f9>

463 Lachenbruch, P.A., Mickey, M.R., 1968. Estimation of error rates in discrim-  
464 inant analysis. *Technometrics* 10, 1–11. <https://doi.org/10.2307/1266219>

465 Langford, E., Schwertman, N., Owens, M., 2001. Is the property of being

- <sup>466</sup> positively correlated transitive? *The American Statistician* 55, 322–325. <https://doi.org/10.1198/000313001753272286>
- <sup>467</sup> //doi.org/10.1198/000313001753272286
- <sup>468</sup> Luo, S., Chen, J.M., Wang, C., Xi, X., Zeng, H., Peng, D., Li, D., 2016.
- <sup>469</sup> Effects of lidar point density, sampling size and height threshold on estimation
- <sup>470</sup> accuracy of crop biophysical parameters. *Opt. Express* 24, 11578–11593. <https://doi.org/10.1364/OE.24.011578>
- <sup>471</sup> //doi.org/10.1364/OE.24.011578
- <sup>472</sup> Ma, W., Domke, G.M., D'Amato, A.W., Woodall, C.W., Walters, B.F.,
- <sup>473</sup> Deo, R.K., 2018. Using matrix models to estimate aboveground forest biomass
- <sup>474</sup> dynamics in the eastern USA through various combinations of LiDAR, landsat,
- <sup>475</sup> and forest inventory data. *Environmental Research Letters* 13, 125004. <https://doi.org/10.1088/1748-9326/aaeaa3>
- <sup>476</sup> //doi.org/10.1088/1748-9326/aaeaa3
- <sup>477</sup> Magnussen, S., Boudewyn, P., 1998. Derivations of stand heights from
- <sup>478</sup> airborne laser scanner data with canopy-based quantile estimators. *Canadian*
- <sup>479</sup> *Journal of Forest Research* 28, 1016–1031. <https://doi.org/10.1139/x98-078>
- <sup>480</sup> Massey, F.J., 1951. The kolmogorov-smirnov test for goodness of fit. *Journal*
- <sup>481</sup> *of the American Statistical Association* 46, 68–78. <https://doi.org/10.1080/01621459.1951.10500769>
- <sup>482</sup>
- <sup>483</sup> Næsset, E., 1997. Determination of mean tree height of forest stands using
- <sup>484</sup> airborne laser scanner data. *ISPRS Journal of Photogrammetry and Remote*
- <sup>485</sup> *Sensing* 52, 49–56. [https://doi.org/10.1016/S0924-2716\(97\)83000-6](https://doi.org/10.1016/S0924-2716(97)83000-6)
- <sup>486</sup> New York Office of Information Technology Services, 2018. LIDAR collection
- <sup>487</sup> (QL2) for Cayuga County and most of Oswego County, New York Lidar; Classified
- <sup>488</sup> Point Cloud [Data file].
- <sup>489</sup> New York Office of Information Technology Services, 2015. NY WarrenWash-
- <sup>490</sup> ingtonEssex Spring 2015; Classified Point Cloud [Data file].
- <sup>491</sup> Nilsson, M., 1996. Estimation of tree heights and stand volume using
- <sup>492</sup> an airborne lidar system. *Remote Sensing of Environment* 56, 1–7. [https://doi.org/10.1016/S0034-4293\(96\)80001-1](https://doi.org/10.1016/S0034-4293(96)80001-1)

- 493        /[doi.org/10.1016/0034-4257\(95\)00224-3](https://doi.org/10.1016/0034-4257(95)00224-3)
- 494        Pflugmacher, D., Cohen, W.B., Kennedy, R.E., 2012. Using landsat-derived  
495        disturbance history (1972–2010) to predict current forest structure. *Remote*  
496        *Sensing of Environment* 122, 146–165. <https://doi.org/10.1016/j.rse.2011.09.025>
- 497        Pflugmacher, D., Cohen, W.B., Kennedy, R.E., Yang, Z., 2014. Using  
498        landsat-derived disturbance and recovery history and lidar to map forest biomass  
499        dynamics. *Remote Sensing of Environment* 151, 124–137. <https://doi.org/10.1016/j.rse.2013.05.033>
- 501        R Core Team, 2021. R: A language and environment for statistical computing.  
502        R Foundation for Statistical Computing, Vienna, Austria.
- 503        Riemann, R., Wilson, B.T., Lister, A., Parks, S., 2010. An effective assessment  
504        protocol for continuous geospatial datasets of forest characteristics using usfs  
505        forest inventory and analysis (fia) data. *Remote Sensing of Environment* 114,  
506        2337–2352. <https://doi.org/10.1016/j.rse.2010.05.010>
- 507        Roussel, J., Auty, D., Coops, N.C., Tompalski, P., Goodbody, T.R., Meador,  
508        A.S., Bourdon, J.-F., de Boissieu, F., Achim, A., 2020. LidR: An r package for  
509        analysis of airborne laser scanning (als) data. *Remote Sensing of Environment*  
510        251, 112061. <https://doi.org/10.1016/j.rse.2020.112061>
- 511        Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., Aroyo,  
512        L., 2021. “Everyone wants to do the model work, not the data work”: Data  
513        cascades in high-stakes ai, in: *Proceedings of Chi 2021*.
- 514        Stone, M., 1974. Cross-validatory choice and assessment of statistical pre-  
515        dictions. *Journal of the Royal Statistical Society. Series B (Methodological)* 36,  
516        111–147.
- 517        St-Onge, B., Hu, Y., Vega, C., 2008. Mapping the height and above-  
518        ground biomass of a mixed forest using lidar and stereo ikonos images. *In-*  
519        *ternational Journal of Remote Sensing* 29, 1277–1294. <https://doi.org/10.1080/>

520 01431160701736505

521 Thorndike, E., 1999. New York's Adirondack Park: Where U.S. wilderness  
522 preservation began. *International Journal of Wilderness* 5, 9–14.

523 Wassер, L., Day, R., Chasmer, L., Taylor, A., 2013. Influence of vegetation  
524 structure on lidar-derived canopy height and fractional cover in forested riparian  
525 buffers during leaf-off and leaf-on conditions. *PLOS ONE* 8, 1–13. <https://doi.org/10.1371/journal.pone.0054776>

526

527 White, J.C., Arnett, J.T., Wulder, M.A., Tompalski, P., Coops, N.C., 2015.  
528 Evaluating the impact of leaf-on and leaf-off airborne laser scanning data on the  
529 estimation of forest inventory attributes with the area-based approach. *Canadian  
530 Journal of Forest Research* 45, 1498–1513. <https://doi.org/10.1139/cjfr-2015-0192>

531

532 Wright, M.N., Ziegler, A., 2017. ranger: A fast implementation of random  
533 forests for high dimensional data in C++ and R. *Journal of Statistical Software*  
534 77, 1–17. <https://doi.org/10.18637/jss.v077.i01>