Earth Science Informatics
https://doi.org/10.1007/s12145-024-01268-9

RESEARCH

How does extreme point sampling affect non-extreme simulation in geographical random forest?

Hui Wang[1,2] · Meixu Chen[3] · Zhe Wang[4] · Li Huang[2] · Christopher C. Caudill[5] · Shijin Qu[6] · Xiang Que[4,7]

Received: 24 November 2023 / Accepted: 2 March 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract

Spatial heterogeneity brings numerous uncertainties to training datasets in the modeling process. An arbitrary selection of training samples can result in a biased simulation. Although previous research provides a chance of reducing the degree of spatial variance through homogeneous divisions, detailed information regarding the impact of the configuration of divisions for training remains unknown. Moreover, few studies investigate the cross impact of extreme sampling on non-extreme simulation. Therefore, we extend previous research to investigate the cross impact and further examine whether the divisions of extremely high (EXH) and low (EXL) quantiles contribute equally to the simulation bias when employing the spatial stratified sampling. Statistical assessment demonstrates that the selection of extreme training sample does affect the non-extreme simulation. The model has the best performance (RMSE: 2.735, VE: 7.481, Bias: -0.033) when the least proportion (25%) of EXH and EXL was selected for training. Further analysis also indicated that the EXH and EXL divisions contribute unequally to the process. Particularly, the non-extreme simulation is more sensitive to the EXH training data with a steeper change rate of 0.043. This research provides a critical insight into the extreme point sampling for a machine learning process. Different sensitivity of division calls upon that extreme training sample should be adjusted on a basis of percentage rather than their amounts when applying stratified sampling in Geographical Random Forest.

Keywords Extreme point sampling · Geographical random forest · LiDAR · Canopy height

Introduction

Sampling refers to a process of extracting information over an entire population by using limited number of observations that are as representative as possible (Berry and Marble 1968). Over the past decades, sampling approaches are divided into two categories based on their conceptualizations: the design-based (e.g. random sampling, systematic sampling, stratified random sampling) (Gómez Puente et al. 2013; Gregoire 1998), and the model-based (e.g. universal kriging, co-kriging) (Brus and De Gruijter 1997). Previous studies suggest that systematic sampling is the most efficient method, but its efficiency is highly mutable in a clustered population where stratified random sampling is considered as a preference (Dunn and Harrison 1993; Sayed and Ibrahim 2018). Moreover, a non-stationary distribution, such as vegetation height over a complex landscape, was found to have a severe influence on systematic sampling (J.-F. Wang et al. 2012; Wang et al. 2010). This would probably because a neglect of this clustered spatial distribution could lead to the fact which breaks a tenet of the

Communicated by H. Babaie.

✉ Shijin Qu
qusj@cug.edu.cn

1 Department of Geosciences, Mississippi State University, Mississippi State, Mississippi 39762, USA

2 Institute for Modeling Collaboration and Innovation, University of Idaho, Moscow, Idaho 83844, USA

3 Department of Geography and Planning, University of Liverpool, Liverpool L69 7ZT, UK

4 Department of Computer Science, University of Idaho, Moscow, Idaho 83844, USA

5 Department of Fish and Wildlife Sciences, University of Idaho, Moscow, Idaho 83844, USA

6 School of Public Administration, China University of Geosciences, Wuhan 430074, China

7 College of Computer and Information Sciences, Fujian Agriculture and Forestry University, Fuzhou 350002, China

word "representativeness" (Dixon 1950). To address this concern, a spatial stratified sampling (SSS) idea similar to the stratified random sampling was developed (J.-F. Wang et al. 2012; Wang et al. 1997). With this method, a heterogeneous area is divided into several subareas or zones that are as homogeneous as possible prior to stratified sampling to reduce the degree of spatial variance (Wang et al. 1997).

Although the division by choosing optimal distribution of sample can mitigate the issue of spatial heterogeneity, further questions regarding "how much" should be taken as sample within each subarea remain unclear. The criteria of selection for subarea attracts more attention when population are partitioned by the value of target variable (e.g., quantile) with extreme observations involved. An arbitrary selection of these extreme samples as training data may bring unexpected bias to statistical estimates, resulting in under or overestimated results (Kwak and Kim 2017).

A biased estimation is recently detected when the random forest (RF) is applied (Belgiu and Drăguţ 2016). It is a widely-used decision tree-based ensemble learning method but is sensitive to training data characteristics, such as sample size, range of training data and spatial autocorrelation (Millard and Richardson 2015; Wang et al. 2021). This leads to the fact that an inappropriate configuration of training divisions will affect the efficiency of prediction of the model. For example, Millard and Richardson (2015) concluded that the predicted proportion of division shares a positive correlation with the proportion of the division in the training dataset. That is to say, after applying the SSS, less training data of a division may cause lower accuracy of prediction for the same division due to the insufficient learning of the model. Thus, optimizing this tradeoff is especially important when attempting to simulate both extreme and central conditions.

Following the study of Wang et al. (2021), we learned that the extremely high (EXH) and low (EXL) tree canopy heights are under and over-estimated, respectively, with random sampling. Although the divisions (i.e., EXH, non-extreme, EXL) are predetermined for the research, there is no method dealing with the spatial heterogeneity issues in the sampling process. For example, the divisions of training data (e.g., EXH and EXL) may contribute unequally to the sensitivity of model performance, and sample size can also be an essential factor leading to less convincing simulated results (Boukerche et al. 2020; Byrd et al. 2012; Uçar et al. 2020). In this study, therefore, we proposed to address the issue by applying SSS based on the predetermined divisions. As aforementioned, we understand the rule of how each division can be affected by the selection of its training data, but the cross impact of extreme division sampling on non-extreme simulation remains unknown. A neglect of this understanding would affect the modeling results when a significant number of extreme observations are involved. In the present study, we follow up the partial results of Wang et al. (2021) and aim to investigate two research questions when elimination of spatial heterogeneity is considered:

1) It is known that extreme point sampling could affect simulation of non-extreme values. But what is the sensitivity of non-extreme simulation to the EXH and EXL divisions? Will these divisions contribute equally to the process?
2) If not, can we quantitatively assess these uneven contributions? Are there any strategies to follow when choosing extreme samples for the training data?

Method and data

Method description

This study uses the concept of SSS and the geographical random forest (GRF) (Georganos et al. 2021), a local machine learning model, to examine the two research questions. Various configurations of training data are generated from the sampling and thus the rest of population are used for validation in the GRF.

The SSS is employed to mitigate the issue of spatial heterogeneity based on the predetermined divisions of tree canopy height over a mixed landscape. The mechanism of this method shares a similarity with the idea of zoning stratified sampling (Wang et al. 2010) while the criteria of division are slightly different. The rules of division can be on a basis of prior knowledge about the area of interest, standardized criteria, or the spatial distribution of other known influencing variables. Tree canopy height is the dependent and only variable of target in this study, and the value has been calculated for the corresponding location. Therefore, the judgement of whether the heights can be defined as extreme depends on their statistical distribution. The height with a predefined high (low) rank order is recognized as one of the EXH (EXL) canopies. In the present research, as aforementioned, we continue to use the divisions (i.e., EXH, non-extreme, EXL) determined in the previous study, but the scope is redefined to better reflect the concept of extreme. After the divisions are plotted on a map, random sampling is eventually applied to select the training data. Although the major process of design-based sampling focuses on choosing training data through some randomization mechanism, Fig. 1 illustrates how the SSS works and its difference from the other design-based sampling methods (Dumelle et al. 2022; James and Knaub 1999). One of the major differences

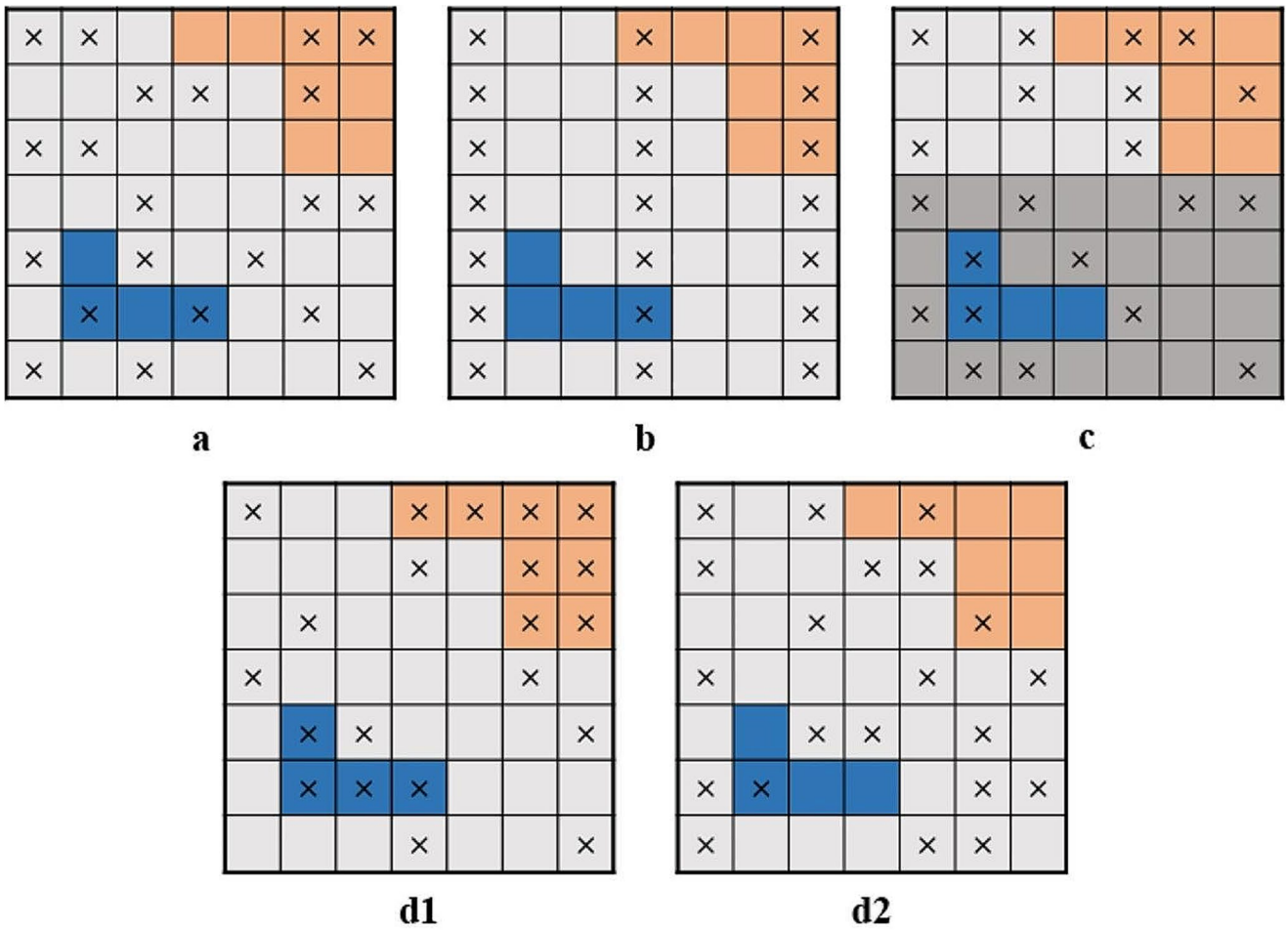

Fig. 1 Examples of design-based sampling methods. An "x" represents the selected pixel for training dataset. The blue and orange squares indicate the pixels with extremely low and high values, respectively. Figures a, b and c represent a random sampling, a systematic sampling, and a stratified sampling (weighted by area, light and gray boxes), respectively. Figure d depicts the spatial stratified sampling; d1 (100% & 100%) and d2 (25% & 25%) show the configurations of EXH and EXL for training

takes ideas of weight and space into consideration when applying the SSS.

GRF is a modified random forest integrated with the geographically weighted regression (GWR) model (Brunsdon et al. 1998) to ease the issue of spatial variance for the selected covariates (Wadoux et al. 2020). Similarly, this method applies an adjustable kernel to choose a bandwidth with maximum radius, capturing as much information as possible to offset the negative effect brought by an unequal distribution of tree canopy over the study area. The major discrepancy between GRF and traditional random forest (TRF) is the dimensionality over space, with an attribute of location embedded in GRF. The formulas can be expressed as:

$$H_i = \alpha_i u_i + \cdots + e_i \tag{1}$$

$$H_{i(x,y)} = \alpha_{i(x,y)} u_i + \cdots + e_{i(x,y)} \tag{2}$$

Where H_i denotes the simulated canopy height at pixel i, contributions of selected features (e.g., $\alpha_i u_i$) are at the right side of the equation, e_i represents the simulation error at the pixel i and (x, y) provides GRF with the locational information at the site. As a decision tree-based method, GRF uses the same calibrated results of the two most important parameters (Zafari et al. 2019), number of decision trees and selected features, as TRF in the subsequent validation stage. It is worth mentioning that the validated pixel only uses the closest calibrated GRF to predict the canopy height in its location. Previous research indicates that a TRF-GRF fusion model provided the least biased prediction, and a trade-off approach is suggested for an optimal configuration (e.g., 50% of TRF-50% of GRF, hereafter, "50% GRF") based on actual needs of a given project. Following the study of Wang et al. (2021), we decided to apply a configuration of 50% GRF in this study because the model accuracy is maximized and the spatial autocorrelation is addressed moderately at

this level. The GRF algorithm was developed as the package 'SpatialML 0.1.3' in R by Kalogirou and Georganos (2018).

Data acquisition and predefinition

As we proposed the questions on a basis of conclusions made by the study of Wang et al. (2021), this research shares the same data acquisition and the set of calibrated parameters for both GRF and TRF over the study area within the Mann Creek Watershed in the state of Idaho, USA (Fig. 2). This section only provides some key results of data cleaning, while detailed introduction of data background can be found in their research (Wang et al. 2021).

Our study is primarily based on two major types of data, the Light Detection and Ranging (LiDAR) point cloud dataset and Landsat imagery. Tree canopy height analysis is conducted through a set of pixel-based datasets derived from the Canopy Height Model (CHM) created by the LiDAR (Fig. 2). The LiDAR data were collected from September 9th ~ October 14th, 2017. For a compliance with

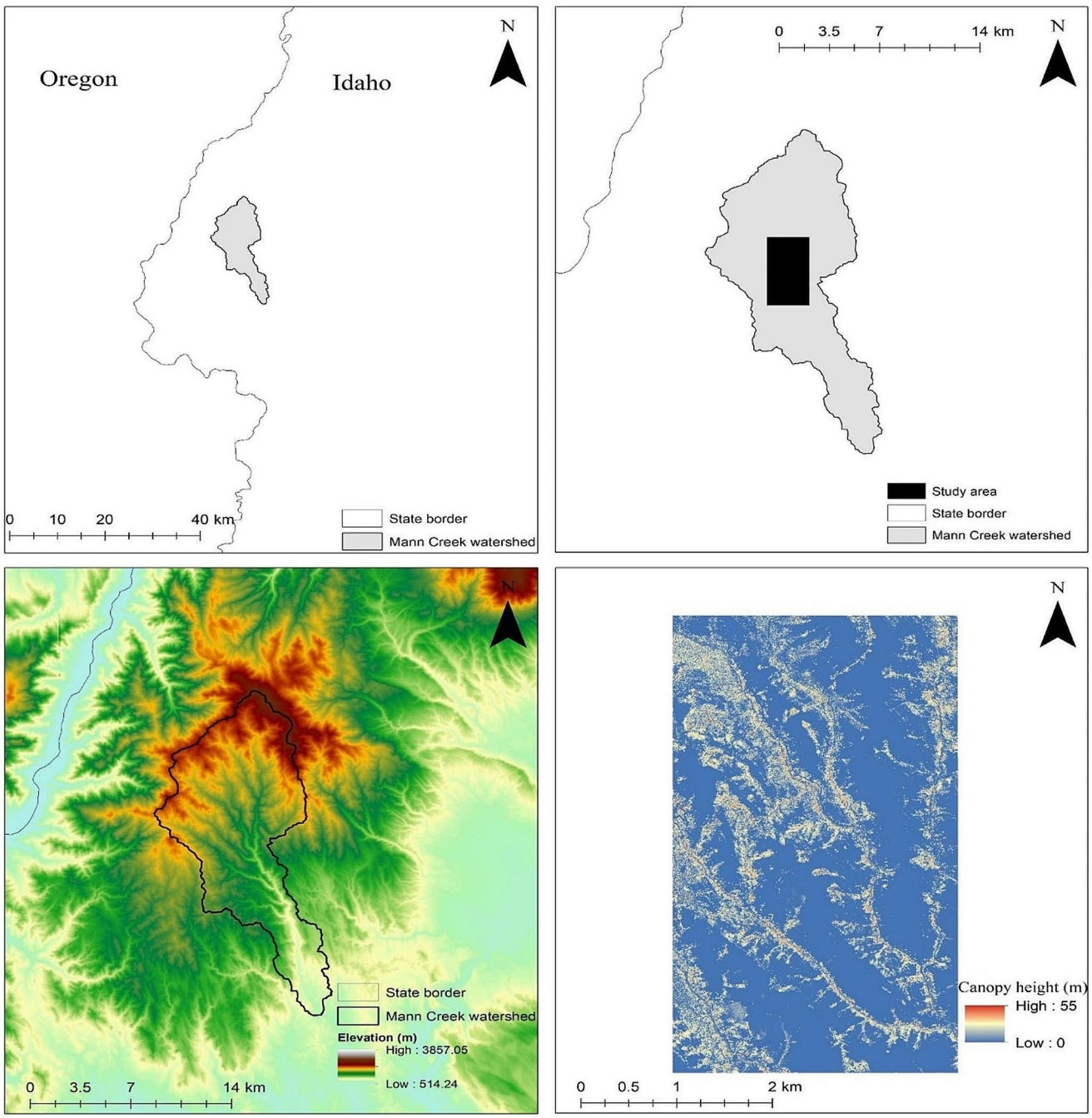

Fig. 2 Study area of this research

other source data, mean value of the LiDAR CHM (0.25 m × 0.25 m) was calculated for each 30 m × 30 m pixel as a reference of true canopy height. The outermost pixels were eliminated because of the edge effect caused by the loss of neighbor. To avoid the intervention brought by other vegetation, we excluded those pixels, primarily dominated by homogenous sagebrush in uplands, less than or equal to 1 m in further analysis. The Landsat images were acquired on October 6th, 2017, in order to be synchronous with the date of the LiDAR data acquisition. Previous research shows that the reflectance of vegetated surfaces depends on its structure, optical property, and underlying soil (Myneni et al. 1995a, b; Myneni, Hall, Sellers, & MarshaMyneni et al. 1995a, b; Zeng et al. 2021). However, we omitted blue band due to its sensitivity to the aerosol conditions in the atmosphere (Flood et al. 2013). Under this case, to prepare for the calibration of the model, we retained 5 basic land surface reflectance (LSF) bands, green, red, near infrared (NIR), short-wave infrared 1 (SWIR1), and short-wave infrared 2 (SWIR2), to generate vegetation indices and band rations. Therefore, a total of 27 parameters, including 5 LSF bands, 12 vegetation indices and 10 band ratios are prepared to run the model (Table 1). A feature selection was then implemented to rule out those with high collinearity (Pearson coefficient) and low importance (Gini importance), leading to a total of 12 parameters remaining. Predefining extreme division of tree canopy height plays a decisive role in model design. Following the previous study, we selected half number of pixels for respective training and validation datasets, resulting in 8868 for each to ensure a balance and a full coverage of the extreme. To better reflect the concept of extreme, the scope of division is redefined based on the quantile statistics of true canopy height. To maintain a reasonable amount of data within extreme division, we eventually identified the canopy heights lower than the 1st quantile (< 3.72 m) and higher than the 95% quantile (> 19.53 m) as the EXL and EXH. Under this case, the divisions of extreme and non-extreme contain 2913 and 5775 pixels, respectively. In order to investigate the impact of extreme sampling on non-extreme simulation, we decided to adjust the SSS weighting of extreme values in the training dataset, leading to a range from 25%~100% at a 25% interval. Based on previous studies (Masud et al. 2008; Vabalas et al. 2019), this limited training sample will only result in a minor change of the simulation with a small increment of percentage. Therefore, we only retained those same percentages and the entire amount (i.e., 100%) of extreme training divisions. Under this case, there are 10 configurations of extreme divisions applied in this study, such as 25% of EXL and 25% of EXH (Table 2). Model optimization was initiated through determining the number of decision trees employed in RF. The highest cross validation score (0.846) was achieved when 1500 decision trees were used. An adaptive kernel of 51 neighboring points was selected for GRF calibration. To quantitatively assess the impact of extreme training sample on the non-extreme simulation, the root mean squared error (RMSE), variance of error (VE) and bias statistics were calculated for only non-extreme division at the validation stage. Equations were expressed as:

$$\mathrm{RSME} = \sqrt{\frac{1}{n}\sum (h_i - x_i)^2} \tag{1}$$

$$\mathrm{VE} = \frac{1}{n-1}\sum_{i=1}^{n}\left(e_i - \bar{e}\right)^2 \tag{2}$$

$$\mathrm{Bias} = \sum_{i=1}^{n}\frac{h_i - x_i}{n} \tag{3}$$

where n represents the number of sample points, h_i and x_i were the predicted and observed or true canopy height at point i, respectively. e_i denotes the difference between h_i and x_i at point i while $\bar{e}$ is the average of errors for all sample points. Additionally, we used residual to measure the deviation of a simulated value and its corresponding true value. The formula is indicated as follows:

$$x_i - h_i = \mathrm{Residual}$$

Therefore, a positive residual implies that the corresponding tree canopy is underestimated while a negative residual indicates an overestimation. To examine the quantitative impact of each configuration set (e.g., 25% of EXL and 25%

Table 1 Spectral, ratio and band features used in the study

Vegetation Index	Band Ratio	LSF bands
Normalized Difference Vegetation Index (NDVI); Green Soil Adjusted Vegetation Index (GSAVI); Green Normalized Vegetation Index (GNDVI); Chlorophyll Vegetation Index (CVI); Normalized Difference Greenness Index (NDGI); Normalized Burn Ratio SWIR2 (NBR); Normalized Burn Ratio SWIR1 (NDII); Green Difference Vegetation Index (GDVI); Modified Soil Adjusted Vegetation Index (MSAVI); Difference Vegetation Index (DVI); Soil adjusted Vegetation index (SAVI); Modified Simple Ratio (MSR)	Red/Green; SWIR1/NIR NIR/Green; SWIR2/Green NIR/Red; SWIR2/Red SWIR1/Green; SWIR2/NIR SWIR1/Red; SWIR2/SWIR1	Green; Red; NIR; SWIR1; SWIR2

Table 2 SSS weighting configurations for extreme value training datasets and statistical results

Case #	% of EXL	% of EXH	# of EXL	# of EXH	RMSE	VE	Bias
1	25%	25%	604	124	2.735	7.481	-0.033
2	50%	50%	1208	249	2.752	7.571	-0.067
3	75%	75%	1812	373	2.863	8.190	-0.079
4	100%	100%	2416	497	2.903	8.427	-0.058
5	25%	100%	604	497	2.818	7.874	-0.266
6	50%	100%	1208	497	2.834	7.976	-0.237
7	75%	100%	1812	497	2.898	8.384	-0.118
8	100%	100%	2416	497	2.903	8.427	-0.058
9	100%	25%	2416	124	2.776	7.675	0.180
10	100%	50%	2416	249	2.801	7.836	0.121
11	100%	75%	2416	373	2.851	8.130	0.048
12	100%	100%	2416	497	2.903	8.427	-0.058

Note Cases are divided into three groups for analysis. Group 1 (G1) spans from Case #1–4, representing the condition when the proportion of EXH and EXL increase synchronously. Group 2 (G2) spans from Case #5–8, representing the condition when the EXH division is fixed as 100%. Group 3 (G3) spans from Case #9–12, representing the condition when the EXL division is fixed as 100%

of EXH) on the non-extreme simulated data, a series of figures of canopy height – residual is compiled in this study.

Results and discussion

Table 2 shows that the impact of extreme training sample on the non-extreme simulation can be varying depending on the different configuration of EXH and EXL. In general, the RMSE and VE manifest an increasing trend when the percentage of either EXH or EXL is growing in the training dataset. This reconfirmed that extreme point sampling could affect simulation of non-extreme values. The Bias indicator exhibits that the negative and positive values are associated with a fixed (100%) EXL or EXH, indicating that various configurations of extreme training sample can lead to an overestimated or underestimated non-extreme simulation. Although discrepancies between cases of each assessment index are considered as references to investigate the research questions, the accuracy of each case of the non-extreme simulation is still relatively high and acceptable for subsequent analysis. Integrating the configurations of extreme training sample with the statistics, we can conclude that both EXH and EXL affected the simulation of non-extreme division while their contributions were different.

When the proportion of EXH and EXL increase synchronously, the values of RMSE and VE rise by 0.168 and 0.946, respectively. This finding shows that the predictive ability of GRF to the non-extreme simulation is weakened slightly with more extreme training sample involved. This inference also applies to the circumstances (i.e., Case #5–12) when proportion of an extreme division is fixed as 100%. The increasing trend of VE implies that a greater percentage of either EXH or EXL can lead to a larger variance of error of simulation, providing that more extreme training data can cause the model overfitting due to a larger random noise. Moreover, comparisons between G2 and G3 indicate that these two statistics are more sensitive (higher change rates) to the EXH even though the number of EXH is less than that of EXL in the training data. Evidence shows that the errors are more intense under the cases that the entire EXH division is selected for training (G2). This conclusion is made by the fact that the rate of variation of the errors is faster when the EXL is fixed as 100%, in spite of the significantly lower increment of the EXH than that of the EXL in G2 (Fig. 3).

The Bias statistic reflects an instability with both negative and positive values, showing that various proportions of extreme training sample can lead to over and underestimates, respectively. Comparisons between G1, G2 and G3 manifest that the 1st and 5th case own the smallest and greatest bias, respectively. The negative value indicates the overestimated simulation in the cases of G1 and G2, regardless of the proportion of the EXL applied to the training dataset. Based on this evidence, it is appropriate to infer that the model is more sensitive to the EXH division, with the weakest predictive ability occurred when the smallest (25%) and largest (100%) percentages of EXL and EXH are used, respectively. This speculation is validated again by the cases of G3 where the entire division of EXL is employed. The simulated canopy heights are first underestimated (i.e., Case #9–11) and was then changed to overestimation with 100% of the EXH selected for training. The increased 25% of the EXH are attributed to this transition although the corresponding incremental number is only 124. Therefore, we can draw a conclusion that the non-extreme simulation is more sensitive to the EXH training dataset in the present research. Based on the statistical distribution of tree canopy height, we hypothesize that the impact of extreme point sampling on non-extreme simulation is closely associated

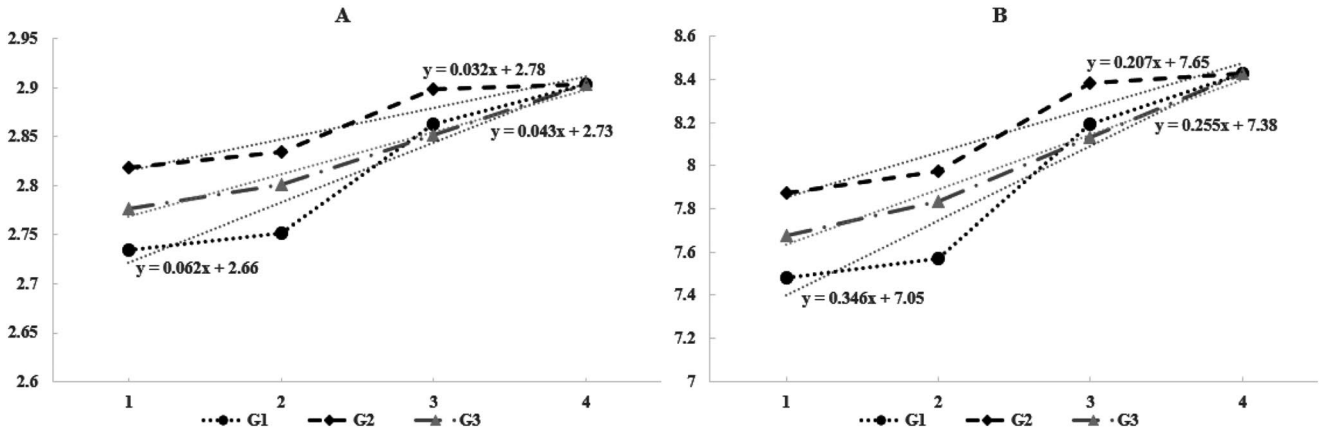

Fig. 3 Comparisons of two statistical indicators (A: RMSE, B: VE) between G1 (growing % in same pace for EXH and EXL), G2 (with a fixed 100% for EXH) and G3 (with a fixed 100% for EXL)

with the overall distribution of non-extreme heights. In this study, a majority of the non-extreme are prone to be relatively low, leading to the fact that more selected EXH training sample may cause a significant disturbance even though the number is critically less than the EXL at the same percentage level. This hypothesis can be verified through Fig. 4 which shows the relationship between non-extreme true canopy height and its residual for each configuration set. From A to J, the non-extreme true canopies are found to be biased to the low heights.

Figure 4 further reveals a coherently robust predictive ability of every configuration while some outstanding residuals make their performances various. A dispersed distribution of these outstanding residuals (>|5| m) helps to quantitatively define the impacts of extreme sampling by combining the comprehensive statistical assessment. The integrated analysis implies that the strongest predictive ability of the model for the non-extreme simulation is taken place at Case #1, with the least proportion (25%) of EXH and EXL selected for training. This evidence further verifies that less amount of extreme training sample has a smaller effect on the non-extreme simulation. However, the underlying mechanism of this influence may vary depending on the different statistical index. The variation trend of Bias statistic is distinct from others. In future research, particularly with biased distribution of target value, the selection of extreme training data should be determined by percentage rather than amount due to the substantially different sensitivity of model to each division. Although less extreme training data may lead to an inaccurate simulation of extreme division, the observation of this study apparently provides a contrary implication. Therefore, a trade-off consideration is desirable when choosing the best configuration of training dataset between the extreme and non-extreme.

Conclusion

This follow-up research provides a critical insight into the impact of extreme point sampling on the non-extreme simulation, particularly for a machine learning process. Two major research questions have been addressed. First, we confirmed that the selection of extreme training sample does affect the non-extreme simulation. In addition to this, we also further found that the non-extreme simulation is more sensitive to the EXH training data in this case study, leading to the fact that the EXH and EXL divisions contribute unequally to the process. Moreover, we also infer that this phenomenon probably depends on the biased distribution of non-extreme division. Therefore, an investigation of the 'noise' embedded in a set of simulation data also seems to be necessary at a preliminary stage before simulation. Second, numbers of extreme sampling data and statistical analysis (i.e., RMSE and VE) demonstrate that the change rates of model performance are unstable for different proportions of EXH and EXL. Therefore, the distinct sensitivities finally call upon an attention that extreme training sample should be adjusted on a basis of ratio or percentage rather than their amounts when applying stratified sampling. This research provides a critical insight into the extreme point sampling for a machine learning process. Although this study fills in the gaps stated above, there are certain limitations existing at the stages of data preparation and method development. First, the EXH and EXL divisions are predefined based on the quantile statistics, which could be involved with subjective consciousness. A more objective definition of "extreme" is expected. Second, the detailed information of variation of statistical value is still missing. The interval of percentage (25%) could be narrowed down to depict a clearer image of statistical change curve in future analysis. Lastly, although we believe this study area can represent those places where canopy heights are randomly distributed, we still expect that

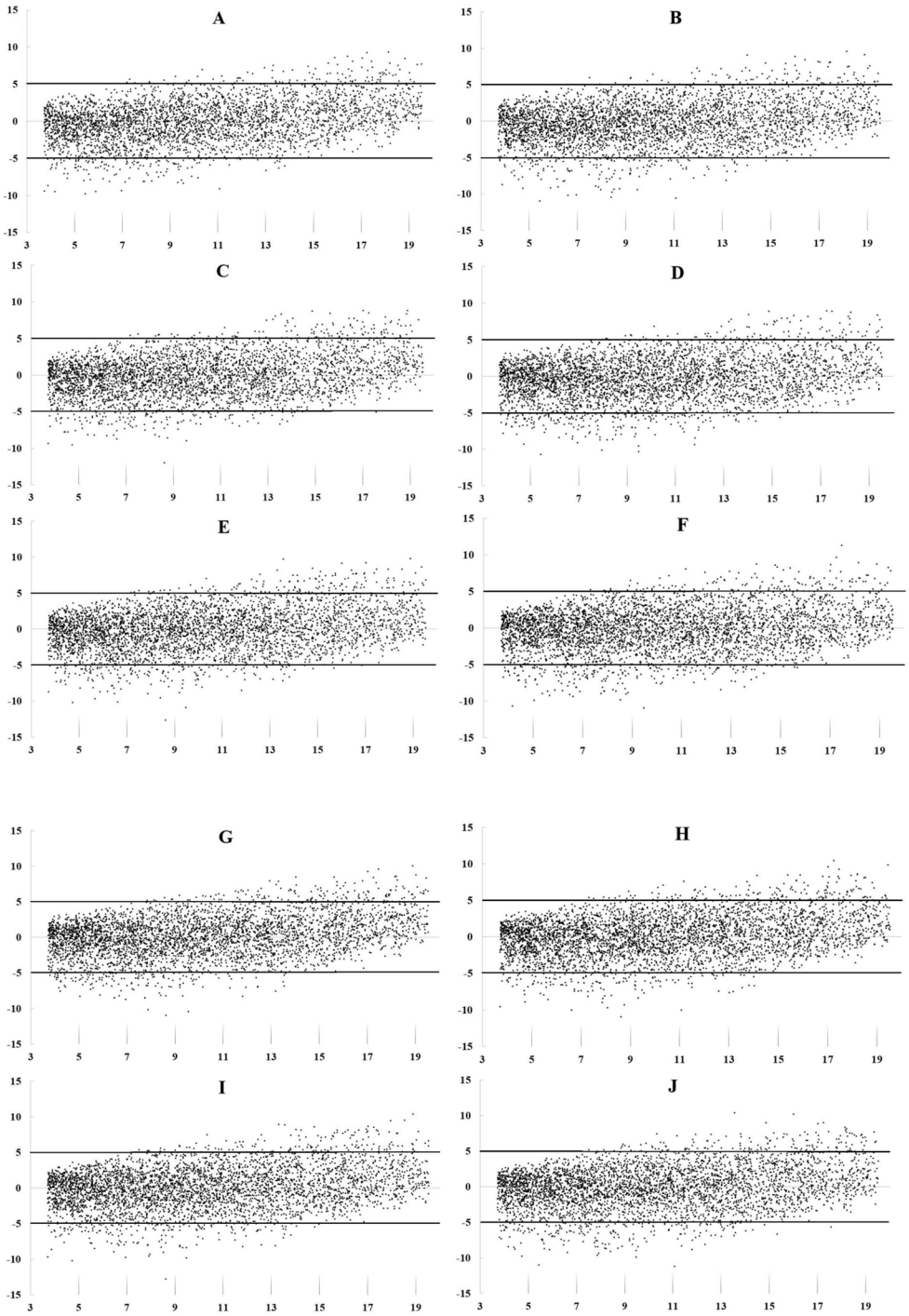

Fig. 4 Relationship between non-extreme true canopy height and its residual for each configuration set. (X-axis: true canopy height in meters, y-axis: residual in meters; A: 25% of EXL and 25% of EXH (hereafter, "25L-25H"), B: 25 L-100 H, C:50 L-50 H, D: 50 L-100 H, E: 75 L-75 H, F: 75 L-100 H, G: 100 L-25 H, H: 100 L-50 H, I: 100 L-75 H, J: 100 L-100 H.)

further research should cover more areas with various types of canopy heights.

Author contributions Hui Wang: Methodology, Data processing, Visualization, Writing-review & editing. Meixu Chen: Writing-review & editing. Zhe Wang: Data processing, Writing-review & editing. Li Huang: Writing-review & editing. Christopher C. Caudill: Supervision, Writing-review & editing. Shijin Qu: Methodology, Visualization, Writing-review & editing. Xiang Que: Writing-review & editing.

Funding This publication was made possible by the NSF Idaho EPSCoR Program and by the National Science Foundation under award number OIA1757324. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NSF. The authors also acknowledge the financial support from the National Natural Science Foundation of China (42202333).

Data availability Data and materials will be made available on request.

Declarations

Competing interests The authors declare no competing interests.

References

Belgiu M, Drăguţ L (2016) Random forest in remote sensing: a review of applications and future directions. ISPRS J Photogrammetry Remote Sens 114:24–31

Berry BJL, Marble DF (1968) Spatial analysis: a reader in statistical geography. Prentice-Hall

Boukerche A, Zheng L, Alfandi O (2020) Outlier detection: methods, models, and classification. ACM Comput Surv (CSUR) 53(3):1–37

Brunsdon C, Fotheringham S, Charlton M (1998) Geographically weighted regression. J Royal Stat Society: Ser D 47(3):431–443

Brus D, De Gruijter J (1997) Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). Geoderma 80(1–2):1–44

Byrd RH, Chin GM, Nocedal J, Wu Y (2012) Sample size selection in optimization methods for machine learning. Math Program 134(1):127–155

Dixon WJ (1950) Analysis of extreme values. Ann Math Stat 21(4):488–506

Dumelle M, Higham M, Ver Hoef JM, Olsen AR, Madsen L (2022) A comparison of design-based and model-based approaches for finite population spatial sampling and inference. Methods Ecol Evol 13(9):2018–2029

Dunn R, Harrison A (1993) Two-dimensional systematic sampling of land use. J Royal Stat Society: Ser C 42(4):585–601

Flood N, Danaher T, Gill T, Gillingham S (2013) An operational scheme for deriving standardised surface reflectance from Landsat TM/ETM + and SPOT HRG imagery for Eastern Australia. Remote Sens 5(1):83–109

Georganos S, Grippa T, Niang Gadiaga A, Linard C, Lennert M, Vanhuysse S, Kalogirou S (2021) Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. Geocarto Int 36(2):121–136

Gómez Puente SM, Van Eijck M, Jochems W (2013) A sampled literature review of design-based learning approaches: a search for key characteristics. Int J Technol Des Educ 23:717–732

Gregoire TG (1998) Design-based and model-based inference in survey sampling: appreciating the difference. Can J for Res 28(10):1429–1447

James R, Knaub J (1999) Model-based sampling, inference and imputation

Kalogirou S, Georganos S (2018) Spatial Machine Learning (Version 0.1.3) [Package]

Kwak SK, Kim JH (2017) Statistical data preparation: management of missing values and outliers. Korean J Anesthesiology 70(4):407

Masud MM, Gao J, Khan L, Han J, Thuraisingham B (2008) *A practical approach to classify evolving data streams: Training with limited amount of labeled data* Paper presented at the 2008 Eighth IEEE International Conference on Data Mining

Millard K, Richardson M (2015) On the importance of training data sample selection in random forest image classification: a case study in peatland ecosystem mapping. Remote Sens 7(7):8489–8515

Myneni R, Maggion S, Iaquinta J, Privette J, Gobron N, Pinty B, Williams D (1995a) Optical remote sensing of vegetation: modeling, caveats, and algorithms. Remote Sens Environ 51(1):169–188

Myneni RB, Hall FG, Sellers PJ, Marshak AL (1995b) The interpretation of spectral vegetation indexes. IEEE Trans Geoscience Remote Sens 33(2):481–486

Sayed A, Ibrahim A (2018) Recent developments in systematic sampling: a review. J Stat Theory Pract 12(2):290–310

Uçar MK, Nour M, Sindi H, Polat K (2020) The effect of training and testing process on machine learning in biomedical datasets. *Mathematical Problems in Engineering, 2020*

Vabalas A, Gowen E, Poliakoff E, Casson AJ (2019) Machine learning algorithm validation with a limited sample size. PLoS ONE 14(11):e0224365

Wadoux AM-C, Minasny B, McBratney AB (2020) Machine learning for digital soil mapping: applications, challenges and suggested solutions. Earth Sci Rev 210:103359

Wang J, Wise S, Haining R (1997) An integrated regionalization of earthquake, flood, and drought hazards in China. Trans GIS 2(1):25–44

Wang J, Haining R, Cao Z (2010) Sample surveying to estimate the mean of a heterogeneous surface: reducing the error variance through zoning. Int J Geogr Inf Sci 24(4):523–543

Wang J-F, Stein A, Gao B-B, Ge Y (2012) A review of spatial sampling. Spat Stat 2:1–14

Wang H, Seaborn T, Wang Z, Caudill CC, Link TE (2021) Modeling tree canopy height using machine learning over mixed vegetation landscapes. Int J Appl Earth Observation Geoinf 101:102353

Zafari A, Zurita-Milla R, Izquierdo-Verdiguier E (2019) Evaluating the performance of a random forest kernel for land cover classification. Remote Sens 11(5):575

Zeng Y, Hao D, Badgley G, Damm A, Rascher U, Ryu Y, Qiu H (2021) Estimating near-infrared reflectance of vegetation from hyperspectral data. Remote Sens Environ 267:112723

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.