

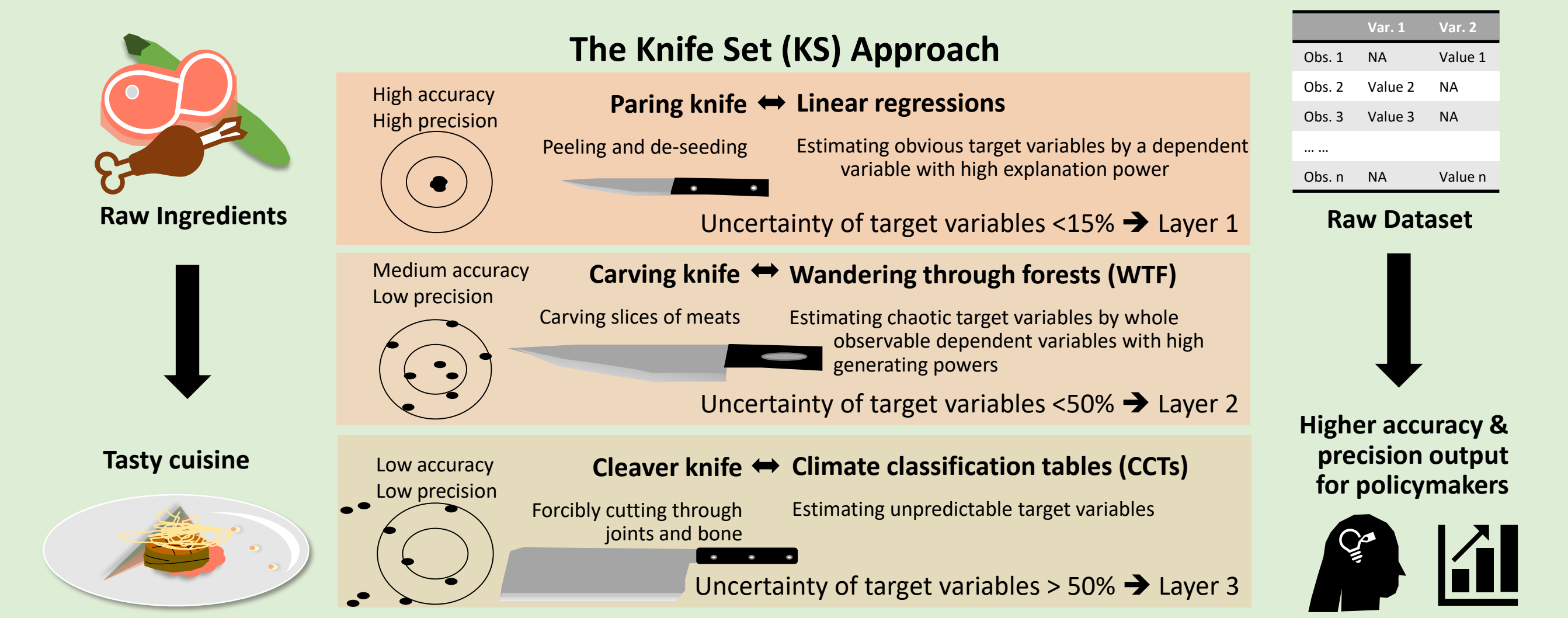
6 ^a Graduate School of Agriculture, Kyoto University, Kyoto, 606-8502, Japan

9 ^c Department of Forestry, College of Agriculture and Nature Resource, National Chung
10 Hsing University, Taichung, 402-204, Taiwan

^e Department of Computer Science & Information Engineering, National Taiwan University, Taipei, 106-216, Taiwan

17 ^g Faculty of Agriculture and Marine Sciences, Kochi University, Kochi, 783-8502,
18 Japan

* Corresponding author (email: gn03138868@gmail.com; Address: Kitashirakawa
Oiwake-cho, Sakyo-ku, Kyoto 606-8502, Japan; Tel: +81-75-753-6096; ORCID:
0000-0003-2719-7459)



- The KS approach provides an algorithmic framework that is not prone to overfitting.
- The study compiled the climatic classification table with CSSs for *Phyllostachys edulis* forests.
- Quantifying unobservable variables is a linchpin to chaotic estimation in the future.
- KS approach can be used for estimating unknown CSSs by datasets with missing data based on its core of CCTs and Gibbs sampling.

Highlights

1. KS approach provides an algorithmic framework that is not prone to overfitting.
2. KS approach has been applied in an extremely divergent case of *P. edulis* forests.
3. KS can be used for estimating unknown CSSs by datasets with missing data.
4. The study compiled the climatic classification table with CSSs for *P. edulis* forests.
5. Quantifying unobservable variables is a key for chaotic estimation in the future.

Abstract

Forest carbon stocks and sinks (CSSs) have been widely estimated using climate classification tables and linear regression (LR) models with common independent variables (IVs) such as the average diameter at breast height (DBH) of stems and root shoot ratio. However, this approach is relatively ineffective when the explanatory power of IVs is lower than that of unobservable variables. Various environmental and anthropogenic factors affect target variables that cause the correlation between them to be chaotic. Here, we designed a knife set (KS) approach combining LR models and the wandering through random forests (WTF) algorithm and applied it in a specific case of *Phyllostachys edulis* (Carrière) J. Houz. (*P. edulis*) forests, which have an irregular relationship between their belowground carbon (BGC) stocks and average DBH. We then validated the KS approach performed by cluster computing to estimate the aboveground carbon (AGC) and BGC stocks and the total net primary production (TNPP). The estimated CSSs were compared to the benchmark of the methodology that applied Tier 1 in the Intergovernmental Panel on Climate Change (IPCC) Guidelines for National Greenhouse Gas Inventories via 10-fold cross validation, and the KS approach significantly increased precision and accuracy of estimations. Our approach provides general insights to accurately estimate forest CSSs relying on evidence-based field data, even if some target variables are divergent in specific forest types. We also pointed out the reason why current fancy models containing machine learning (ML) or deep learning algorithms are not effective in predicting the target variables of certain chaotic systems is perhaps that the total explanatory power of observable variables is less than that of the total unobservable variables. Quantifying unobservable variables into observable variables is a linchpin of future works related to chaotic system estimation.

Keywords. Aboveground net primary production; belowground net primary production; Chaotic estimation; Hybrid machine learning; *Phyllostachys edulis*; Wandering through random forests

1. Introduction

Forest carbon stocks and sinks (CSSs) are usually estimated based on linear regression (LR) models. LR models work well for aboveground CSSs but are relatively ineffective for belowground CSSs. This is mainly because the relationships between the target variable (TV) and independent variables (IVs) are commonly chaotic or divergent and appear everywhere in natural ecosystems (Toker et al., 2020). There is large variability in belowground CSSs estimates among sites and high estimate uncertainties, which can be attributed to environmental and anthropogenic factors such as temperature, precipitation, soil chemical properties, biotic communities, nutrient availability, management intensity, and ecosystem productivity (Poeplau et al., 2011). Thus, if there are many unobservable IVs, it will increase the proportion of the variance in the target variable when using LR-based models to estimate belowground CSSs in forests (Thompson et al., 2006; Zhao et al., 2014).

An irregular belowground forest type mentioned above, *P. edulis* (Synonyms: *P. pubescens*, *P. heterocycla*) is a giant bamboo species, has a high growth rate, and tends to aggressively invade neighbouring vegetation in temperate zones; these invasions have been frequently observed in East Asian countries such as Japan and China (Okutomi et al., 1996; Wang et al., 2013; Lin et al., 2017; Kobayashi et al., 2018) due to *P. edulis* has rhizome systems, and can transport carbohydrates (Zhai et al., 2021) and nutrients (Shi et al., 2021) to others. This may be also one of the main reasons that belowground biomass, aboveground net primary production (ANPP), and belowground net primary production (BNPP) are divergent in *P. edulis* forests. Since *P. edulis* grows

1 53 rapidly, some studies have highlighted its high potential as an efficient carbon stock
2
3 54 (Wang et al., 2009; Yen and Lee, 2011; Nath et al., 2015; Chen et al., 2016) and carbon
4
5 55 sink (Lin et al., 2017). Efficient CSSs are vital to terrestrial ecosystems and should be
6
7 56 included in debates in the agriculture, forestry and other land use (AFOLU) sector of
8
9
10 57 the national inventory report (NIR), which is submitted to the United Nations
11
12 58 Framework Convention on Climate Change (UNFCCC) by all Annex I countries.
13
14 59 However, bamboos have not been sufficiently considered in NIRs (Lobovikov et al.,
15
16 60 2012; Yuen et al., 2017) because they are botanically classified as grasses and not trees
17
18 61 (Buckingham et al., 2011). Furthermore, the carbon sink of bamboos in the
19
20 62 Intergovernmental Panel on Climate Change (IPCC) Guidelines for the NIR has been
21
22 63 considered only for Africa and not for Asia and Central America (Katumbi et al., 2017;
23
24 64 Dong et al., 2019). Providing accurate estimates of CSSs is one of the current priorities
25
26
27
28
29 65 in climate policy.

30
31 66 For estimating the BGC stocks from the aboveground biomass and carbon
32
33 67 storage, mainstream methods are based on biomass expansion factors (BEFs). Such
34
35 68 BEFs have been widely recommended for estimating forest biomass and carbon storage
36
37 69 at a regional or national level (Brown and Lugo, 1984; Brown, 2002; Aalde et al., 2006;
38
39 70 Guo et al., 2010; Zhang et al., 2012; Dong et al., 2019). However, the above methods
40
41 71 are based on linear statistical models. The methods may be suited for estimating
42
43 72 aboveground carbon (AGC) stocks and ANPP but not for estimating BGC and BNPP
44
45 73 in specific cases of belowground structures (Kobayashi et al., 2022). The linear
46
47 74 statistical model does have limitations. The main reason is that the key factors affecting
48
49 75 the growth of forests are different in different regions, and the value range of the factors
50
51 76 is difficult to determine. That's why these methods often perform imprecisely because
52
53 77 target variables and IVs are distributed in divergent structures, as implied by the
54
55 78 existence of unobservable variables.
56
57
58
59
60
61
62
63
64
65

Hybrid algorithms and machine learning (ML) models are potential methods for dealing with divergent and non-monotonic correlations in many research fields (Geman and Geman, 1984; Stallkamp et al., 2012; Silver et al., 2016; Silver et al., 2017; Lee et al., 2018). The rapid development of ML in recent years has enabled the building of models and algorithms for reliably estimating the carbon accumulation potential (Cook-Patton et al., 2020) and soil carbon (SC) storage despite a divergence between target variables and IVs (Bou Kheir et al., 2010; Were et al., 2015; Wang et al., 2018). For instance, Wang et al. (2018) used a genetic algorithm and stepwise regression to select features as IVs and then used a boosted regression tree and random forest (RF) to predict soil organic carbon (SOC). The results of the study indicated that RF outperformed boosted regression tree in the prediction of SOC in rangelands and that feature selection increased model efficiency and accuracy. Furthermore, Were et al. (2015) compared the performance of support vector regression (SVR), artificial neural networks (ANNs), and RF for predicting the SOC across an Afromontane landscape. Even though ML models have been applied in predicting SOC, ML has rarely been combined with simple LR-based models to assess the CSSs in forests. The reason we combined LR and ML-based models is that ecological processes and relating values, including the masses of CSSs, are fundamentally controlled by universal physiological phenomena, physical structures and chemical processes; therefore, simple formulae may be better suited for the fundamental estimation models than the simple application of standard ML structures (which do not involve logic of physiological and physical backgrounds, may result in overfitting for unknown [extrapolation] situations).

As *P. edulis* forests have IVs with regular and monotonous correlations in the aboveground part but irregularly correlated IVs in the belowground part, they are suitable for designing a model that combines LR and ML to estimate target variables such as CSSs. Simple LR models based on correlations between morphological

characters and biomass, which are usually described as allometric models (Martin et al., 2013; Youkhana et al., 2017), had better performance than ML models when estimating highly correlated variables such as AGC and BA (Wang et al., 2019), of which the latter is defined as the total cross-sectional area of all stems or culms in a stand measured at breast height per unit of land area. In contrast, the above example of BGC reveals that most variables in the forest ecosystem may have a high level of complexity and that many unobservable variables could not be observed. Compared with highly correlated variable pairs, there are usually many factors affecting the target variable in the terrestrial carbon cycle when the target variable and the IV are relatively divergent.

To design a well-considered algorithm without overfitting, we observed the various steps of the carbon cycle and previous studies related to *P. edulis* forests. Then, we notice that collecting all variables in one single study is often impossible. Because studies commonly have focused on different purposes such as the estimation of aboveground biomass and the impacts on different management treatments (Wang et al., 2010; Chen et al., 2016), carbon dynamics and productivity (Isagi et al., 1997; Lin et al., 2017), and productivity and the rate of carbon dioxide absorption by applying fertilisers (Song et al., 2020), which causes the raw data sheet to have many missing values always. Having a model that accepts any form of missing data input is very friendly to policymakers, and its usability is also greatly improved. In this study, to overcome the above difficulties, we aimed to design a knife set (KS) approach that combined LR models and wandering through random forests (WTF) algorithm which was similarly used in previous study (Wang et al., 2019) and a generalised Multivariate Imputation by Chained Equations (MICE) method (van Buuren & Groothuis-Oudshoorn, 2011; Shah et al., 2014) that allows the use of nonlinear models by conducting the random forests (Breiman, 2001) and Gibbs sampling (Geman and Geman, 1984), and CCTs for estimating the AGC, BGC, ANPP, and BNPP of *P. edulis*

forests in East Asia.

2. Materials and Methods

2.1. Data collection from study sites and published research articles

2.1.1. Biomass investigation

Field investigations were conducted in *P. edulis* forests at the Forestry and Forest Products Research Institute (FFPRI), Kansai Research Centre (34°56'28.3"N 135°46'20.5"E), ~~Kyoto Prefecture, Japan,~~ and in the Huisun Forest Area, Experimental Forest Management Office, National Chung Hsing University (24°04'28.0"N 121°00'41.0"E), ~~Nantou County, Taiwan,~~ between 2008 and 2019 (Fig. S1). The biomass investigation contained two major parts, aboveground and belowground, of which the former included culms, branches, and leaves and the latter included stumps, rhizomes, and roots. More information on these sites is shown in Dataset S1.

2.1.2. Analysing carbon stocks in *P. edulis* forests

Samples for carbon storage collected in Taiwan via the above steps were milled into a powder using a grinder and sieved through a standard sieve of mesh size 0.149 mm (mesh No. 100). The obtained powders were analysed for their percent carbon content (PCCs) (the same as carbon fractions), which was the ratio of the carbon weight to dry weight in the plant bodies (Wang et al., 2009; Yen et al., 2010; Yen and Lee, 2011) and in the leaves, branches, culms, stumps, rhizomes, roots, and soil carbons (SCs) with an elemental analyser (CHNOS Elemental Analyser, Vario EL, Germany); we used between 3 and 9 duplicates per sample. The carbon storage in each component of *P. edulis* forests was calculated by multiplying the biomass by the corresponding PCCs shown in Dataset S1. Furthermore, the PCCs in Japan were used in previously

published data around Kyoto, Japan (Fukushima et al., 2015).

158

2.1.3. Analyses of soil nutrient contents and soil particle size

In total, 18 samples per bamboo forest study site were further analysed for carbon and nutrient contents at Mountain Ali and Huisun Experimental Forest Station in Taiwan. Nine soil core (100 ml) samples were collected for nutrient contents from 0-10 cm and 10-30 cm separately in the same *P. edulis* forests. At the same time, 3 cores were collected to determine the soil bulk density and calculate the total amount of soil nutrients at each study site. Soil pH (H₂O) (soil to water ratio, 1:1) and pH (KCl) (soil to KCl ratio, 1:2.5) were determined by a glass electrode (WTW/Germany PH537). The particle size analysis of the soil was measured using a hydrometer method. After the samples were naturally air-dried, roots and soil were separated and then sieved through a 2-mm sieve to analyse the C and nutrient contents (N, P, K, Ca, Mg, P). Total N was determined with the same elemental analyser (CHNOS Elemental Analyser, Vario EL, Germany) mentioned above. After the ammonium acetate method was used to measure cation exchange capacity, inductively coupled plasma optical emission spectrometry (inductively coupled plasma emission spectrometry, ICP-AES, Leeman Labs. the USA) was used to determine the contents of other effective nutrients (Ca, K, Mg, P) (Moore and Chapman, 1986).

2.1.4. Data collection: literature surveys

Fifty-five peer-reviewed journal articles on CSSs and many IVs were collected in *P. edulis* ecosystems from East Asia. Together with our field survey data described above, 83 observations from 361 plot data from 47 study sites were collected in a wide geographic area that covered the whole area of naturalised *P. edulis* forests (Dataset S1). In the database, there were 85 variables, including 6 management factors described by

dummy variables, 2 temporal factors, 36 environmental factors, and 43 growth traits overall described in Table 1 and defined in Table S2.

2.2. Constructing an algorithm for estimating AGC, BGC, ANPP, and BNPP

The KS flowchart shown in Fig. 1 illustrates the procedure used in this study for ensemble learning, the estimation of DVs and the details of practical guidelines in the R programming language via cluster or cloud computing approaches described in the Method S1. All the statistical analysis and the hybrid algorithm were performed using R-4.0.0 for Windows (64 bit) and the codes are provided in Method S1. We input the dataset of non-target variables including management, environmental factors, and growth traits as horizontal rows and observations as vertical columns in the matrix. The dataset was separated into a training dataset (90%) and validation dataset (10%) for k-fold cross-validation ($k=10$). The detail for determining the value of k is described in Method S1.

2.2.1. Linear Regression (LR)-Layer 1

The priority of the logic of statistical analysis cannot be higher than that of ecological physiology in LR models. Thus, we mined the variables that were highly correlated with each other, as supported by pieces of evidence of ecological physiology. Then, variables were partially inputted via several LR models based on the above evidence. The major reason for designing LR models to estimate some parts of missing values that formed the output of the layer 1 as the input of the layer 2 was because different studies have different purposes when applied to *P. edulis* forests and all the variables in each study site are not fully recorded in this data matrix.

To increase the stability of the simple algorithm performed by the data matrix collected

via our investigation and peer-reviewed journal articles, we used the Pearson correlation method (Pearson, 1901) to analyse the correlation between DVs and IVs (Table S3), constructed LR models for the target variables such as AGC, BGC, and ANPP, and then predicted them via LR models. Output dataset included target variables.

2.2.2. Wandering through the random forest (WTF) sub-algorithm - Layer 2

The predicted values came from layer 1 if the Pearson correlation coefficient $R^2 > 0.85$. And re-inputted these predicted values to the input of layer 2. A sub-algorithm of the KS, RF ensemble ML (Breiman, 2001) with GS (Geman and Geman, 1984) [also named wandering through the random forests sub-algorithm (WTF sub-algorithm) because it was like that the random forests algorithm wandered through the high-dimensional space] was used for 200 duplicates performed by cluster computing consisting of two i5-9400F CPU @ 2.90 GHz 2.90 GHz, one i3-8100 CPU @ 3.60 GHz 3.60 GHz, and four i9-9900K CPU @ 3.60 GHz 3.60 GHz. [It is also possible to operate cloud computing via Amazon Web Services (AWS) if the operator or policymaker wants to compute faster in this methodology.] The pseudocodes of the WTF sub-algorithm were described in Table S1. Following this, we found the most stable combination of duplicates, maximum iteration times of GS, and CART numbers of all matrices by minimising the quadratic loss function [similar to the lowest mean square error (MSE)] compared to the mean of matrices. Equation (1) illustrates the quadratic loss function of the mean of matrices:

$$\lambda(M) = \frac{1}{n} \sum_{i=1}^n (Pi - Mi)^2 \quad (1)$$

where λ is the quadratic loss function of training the WTF sub-algorithm; Pi is the predicted output value; Mi is the average value in each cell of the formed replicate datasets during iterations; and n is the number of observations.

234 The input validation dataset removed target variables into the trained ensemble
 235 matrix. The WTF sub-algorithm was run using the most stable combination to predict
 236 the target variables via cluster computing again, and all duplicate matrices (an N-
 237 dimensional array) were averaged as one dataset by stacking. Finally, the dataset that
 238 included the estimated values of target variables was output. The output of layer 2 was
 239 the input of layer 3.

240

241 2.2.3. Climate classification tables (CCTs)– Layer 3

242 To estimate target variables that could not be output from the LR models and
 243 the WTF sub-algorithm, we made CCTs for the ratio of BGC to AGC (Table S4), Tier
 244 1-estimated values of target variables (Table S5) via Dataset S1 followed the definition
 245 from the Köppen-Trewatha climate groups and IPCC’s method (Aalde et al., 2006) for
 246 *P. edulis* forests. Then, the estimated values in the rest of the target variables were input
 247 as the output of Layer

248

249 2.3. Statistical analysis & cross-validation of the knife set (KS) approach

250 The benchmark of the IPCC guideline and the hybrid algorithm that combined
 251 LR and RF with GS were compared by MAPE, RMSE, and MAE. If the target variables’
 252 MAPE, RSME, and MAE predicted by the LR models were higher than those predicted
 253 by the hybrid algorithm, we used the LR models’ predicted values of the target variables;
 254 otherwise, we used the hybrid algorithm’s predicted values of the target variables as the
 255 final output. Equations (2) to (5) illustrate the MAPE, RMSE, MAE, and ME,
 256 respectively:

$$257 \quad \text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{(Pi - Oi)}{Oi} \right| \quad (2)$$

258 where Pi and Oi are the predicted output and observed dataset values, respectively, and

n is the number of observations. MAPE is used to compare the percentage of error of the algorithm from each single estimated value that is not affected by the unit.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Pi - Oi)^2} \quad (3)$$

where Pi , Oi , n are the same as above. RMSE is used to compare the intuitive error between the algorithm's mean estimated values and the actual observations (that is, including the concept of the unit), which indicates that a lower RMSE signals a higher accuracy and precision of the algorithm and fewer outlier values.

$$MAE = \frac{1}{n} \sum_{i=1}^n |Pi - Oi| \quad (4)$$

where Pi , Oi , n are the same as above. MAE transfers the negative errors to positive and is used to intuitively compare the error between the actual value and the estimation from the algorithm (that is, the concept of a unit is included).

$$ME = \frac{1}{n} \sum_{i=1}^n (Pi - Oi) \quad (5)$$

where Pi , Oi , n are the same as above. ME is used to compare the error between the actual value and the values estimated by the algorithm. (that is, the concept of the unit is included, and the positive error and the negative error offset each other, which means that if the ME is low, the accuracy is high even though the precision may be low.) Through the above equations and parameters, we revealed the performance of the benchmark and the KS approach via 10-fold cross-validation. Further details of the KS algorithm are revealed in the Method S1.

3. Results

3.1. Correlations of target variables in carbon stocks with independent variables

3.1.1. AGC with BA

The noteworthy results showed that there was a strong significant correlation between AGC and BA ($R^2=0.878$, p -value <0.001) (Fig. 3a; Table S3) in *P. edulis* forests. AGC was significantly related to the mean diameter at breast height (DBH) or culm density even though their R^2 values were low (0.036 and 0.322, respectively) (Fig. 3b, 3c).

3.1.2. BGC with other IVs

As our speculation, BGC has no significant correlation with BA ($R^2 = 0.055$), culm density ($R^2 = 0.043$), and mean DBH ($R^2 = 0.0038$), whose distribution was divergent and irregular (Table S3; Fig. 4). BGC cannot be determined with the method of layer 1 because it has no suitable IVs in LR models (Table S3).

3.2. Correlations of target variables in carbon sinks with independent variables

3.2.1. ANPP with heterotrophic respiration (HR), total nitrogen (N) in the soil, and litterfall

The correlation between ANPP and heterotrophic respiration (HR) was significant ($R^2=0.918$, p -value <0.01) (Fig. 3d). However, litterfall was only slightly significantly correlated with HR (p -value <0.05) (Fig. 3e) and was not significantly correlated with ANPP (p -value >0.05) (Fig. 3f). Furthermore, ANPP was significantly correlated with total N in the soil ($R^2=0.981$, p -value <0.05) (Fig. 3g; Table S3).

3.2.2. BNPP with leaf net production (LNP)

Correlations between BNPP and leaf net production (LNP) were strongly significant ($R^2=0.966$, p -value <0.001) (Fig. 3h). We believe that soil availability

cations and nutrients may be related to BNPP, however there is not much data published on soil availability cations and BNPP (Dataset S1). This cannot be analysed using Pearson's correlation.

3.3. Performance of the knife set (KS) approach

The performance of this KS approach is better than that of the IPCC's Tier 1 method, which employs the default factors provided by the IPCC and has been widely used for estimating broadleaf forest target variables, as a benchmark certified by k-fold cross-validation (k=10). The results of k-fold cross-validation also indicate that if the target and IVs have a significant correlation, the performance of estimating the target variable via LR models is excellent (Table S6). Regardless of the mean absolute error (MAE), root-mean-square error (RMSE), mean error (ME) or mean absolute percentage error (MAPE), the KS is generally lower than the benchmark (IPCC Tier 1) in estimations of 10-fold cross validation (Table S6). Further, the precision and the accuracy of the KS are higher than that of the benchmark in estimations of 10-fold cross validation (Fig. 5 & 6), and are similar with holdout testing (Fig. 7).

4. Discussion

4.1. The estimation of target variables and their correlations with other independent variables used in linear regression and the output from the layer 1

To establish the LR models of the layer 1 of the KS approach, we collected various variables from peer-reviewed articles (Dataset S1), initially determined the annual carbon dynamics for different ecological zones of *P. edulis* forests (Fig. 2), and analysed the correlation between target variables and other IVs (Table S3).

4.1.1. AGC and BA

AGC was positively and significantly correlated with BA and had an $R^2 > 0.85$ across all studies (Fig. 3a; Table S3). When IVs and DVs are highly correlated with small errors, it is usually appropriate to construct linear models with higher accuracy and precision that satisfy this hybrid algorithm to save computing power. For example, BA is suitable for estimating AGC stocks in *P. edulis* forests. Therefore, the AGC and BA were suitable for constructing a simple and highly accurate linear model, and the results were input into the data matrix as layer 1 of the hybrid algorithm. Shinohara et al. (2014) also found the similar results and concluded similar linear models to estimate AGC and BGC via BA in *P. edulis* forests across Japan. Although AGC was significantly related to the mean diameter at breast height (DBH) or culm density, their R^2 values were low (Fig. 3b, 3c) and obtained similar results in *P. edulis* forests based on other research data (Yuen et al., 2017); this approach was suitable for building linear or monotonic regression models to estimate target variables which is not divergent to IVs.

4.1.2. BGC and other IVs

When the target variable and the IVs are divergent and irregular, various factors affecting the target variable exist in the real terrestrial carbon cycle, such as the BGC stocks of *P. edulis* forests, which are affected by annual rainfall (p -value < 0.05), LNP (p -value < 0.05), and selective cutting and weeding (p -value < 0.05) (Table S3). Even though the DV and the IVs were significantly correlated in the above case, their R^2 values were usually less than 0.6, which implied that there were many more unobservable variables and that the fitting of the linear models made by each pair of variables was imprecise and not suitable for estimating the target variable, although these variables still had statistical significance as long as the number of samples was sufficient (Table S3).

359

360 4.1.3. ANPP, heterotrophic respiration (HR), total nitrogen (N) in the soil, and

361 litterfall

362 ANPP as a target variable was significantly correlated with HR and slightly
363 significantly correlated with litterfall. This result might imply that the HR of
364 microbiome communities in the soil decomposed organic matter supplied from litterfall
365 increased and that the nutrient resupply to *P. edulis* forests caused the ANPP to increase.
366 Another possibility was that total nitrogen (N) in the soil, which increased due to the
367 decomposition of litter or fertilisation, drove the activity of the microbiome further to
368 activate ammonification and nitrification in the N cycle and caused the HR to increase.
369 N addition significantly increased the wood biomass, carbon stocks, and ANPP, and the
370 decomposition of SOC was confirmed early by controlled field experiments (Song et
371 al., 2020).

372 However, because it is difficult to obtain data on both ANPP and total N in the
373 soil at the same time, only 3 data points were collected in *P. edulis* forests (Fig. 3g).
374 Only 2 peer-reviewed studies recorded the total N in soil and the HR at the same study
375 site (Xiao et al., 2009; Xiao et al., 2010). ANPP had a slight significant correlation with
376 the total N in the soil. The relationships among soil nutrients, HR and ANPP have the
377 potential for discovering insights, and further related studies might be obtained in the
378 future, even if they have almost reached the threshold of significant correlations.

379

380 4.1.4. BNPP and leaf net production (LNP)

381 Correlations between BNPP and LNP were strongly significant (Fig. 3h)
382 perhaps implied that as more leaves formed, more roots and rhizomes grew because
383 leaves of *P. edulis* produced sugars and supplied them to the developing roots and
384 rhizomes.

Although the BGC might need to be estimated using other reliable methods, the above discussion rationally proved that the AGC could be estimated by the BA, ANPP by HR, and BNPP could be estimated by LNP via LR models applied in the layer 1 of the KS approach. If there were target variables that were not estimated by the first layer of the KS approach, the remaining target variables were also estimated using the following layer 2 and layer 3 (Fig. 1).

4.2. Complexity and unobservable variables in the carbon cycle of forest ecosystems and a potential method to overcome overfitting, the WTF sub-algorithm

Reasons why it is difficult to find independent variables of BGC with high explanatory power (Fig. 4) is because that carbon cycling and growth characteristics in *P. edulis* forests have complicated interactions with each other. For example, the photosynthetic systems in the leaves and epidermis of culms are activated by sunlight to respond to light reactions, a series of electron transfer processes occur, and then ATP and NADP are produced. Meanwhile, the stomata on leaf surfaces absorb carbon dioxides, which supply mesophylls to synthesise carbohydrates in the air and on earth (Wang et al., 2020). These carbohydrates provide culms to form structural carbohydrates (SC) and accumulate dry matter as a part of biomass (Yen and Lee, 2010; Yuen et al., 2017) and supply the respiratory demands of leaves, branches, culms, stumps, rhizomes, roots, and bamboo shoots. In particular, unlike other normal temperate forests, the carbohydrates produced by mature *P. edulis* may be transferred to other parts of young shoots through the rhizome system, resulting in the development of expanded belowground parts and the production of new bamboo shoots (Song et al., 2016), further helping young individuals resist abiotic stresses (Wang et al., 2020). When *P. edulis* forests continue to grow, coarse woody debris (such as old bamboo) and

litterfall gradually return to the soil. The belowground roots, rhizomes and stumps will be decomposed by fungi and microbes, and then carbon will return to the soil and the atmosphere as it is released from organic matter. Each part of the above process may be interfered with by multiple environmental and anthropogenic factors.

Because *P. edulis* forests have a special belowground system that may transfer carbohydrates (Song et al., 2016; Wang et al., 2020), water (Mei et al., 2019), and nutrients (Shi et al., 2021) to young individuals or develop the belowground system itself, their rhizomes and roots are not as stable as those in normal temperate forests and are easily affected by various anthropogenic and environmental factors, which has resulted in difficulty in accurately predicting BGC and irregular root shoot ratios (Table S3).

Each value of a variable is the result of all the other subtle variables interacting with each other. As long as there are infinitely various IVs, it is possible to fully deduce the value that the real target variable should be given, and this value is close to the means of observations. However, some of the current ML models often use only limited IVs to estimate the target variable, which leads to the problem of overfitting (It means the model is accurate in estimating target variables in validation but causes considerable deviation when it is applied in practice). To construct an algorithm that considers all the known variables as much as possible without overfitting, we used the WTF sub-algorithm in the 2nd layer of the KS approach to estimate the values of target variables for which the verification error rate was below 50% (Fig. 1). If there were target variables that were not estimated by the first two layers of the KS approach, the remaining target variables were also estimated using the CCTs (Table S4, S5) as the 3rd layer—the final output as results (Fig. 1).

4.3. Comparing the performance of the knife set (KS) approach with IPCC's Tier

1 method

Our empirical results (Table S6; Fig. 5, 6, 7) prove that the KS approach can improve the accuracy and precision in estimations of the carbon stocks and sinks of *P. edulis* forest without over-fitting. The reason why the KS performance is better than the results of IPCC Tier 1 is that the KS approach provides different and reasonable sub-algorithms for the different characteristics of target variables. For instance, target variables will be estimated by LR, WTF, and CCTs in the KS approach if the uncertainties of target variables are below 15%, between 15-50%, and over 50%, respectively. Another reason is that IPCC's Tier 1 approach may be simplifying assumptions about some carbon pools and can only provide a general but relevantly low precise and accurate approach (Such as IPCC's Tier 1 does not include reference data for scattered bamboo forests but borrow values from broadleaf forests). From observing the KS approach, if there is no significant relationship or a low R^2 between the target and IVs, the WTF sub-algorithm approach will be indispensable for estimating target variables to compensate for the shortcomings of LR models (Fig. 1). Furthermore, to reduce the disadvantage of the data matrix with many missing values and increase the stability and accuracy of the KS approach whose DVs and IVs have divergent but highly accurate correlations, we tried to identify the variables that were highly correlated with each other, as supported by ecophysiological evidence, and used LR models to estimate some of the missing values to form an output of the layer 1 for use as the input of the layer 2 for the WTF sub-algorithm.

Designing the WTF sub-algorithm by combining RF (Breiman et al., 2001) and Gibbs sampling (GS) (Geman and Geman, 1984) is suitable for treating the problem of divergent correlation between the target and IVs. The main reason is that any design that changes the weight of the IV always results in an overfit when the effect of the unobservable variable of the target variable is greater than that of the IV. Unless the

number of variables exceeds the interpolation threshold, it is possible to use mainstream ML without weight changing designs to train algorithms with higher accuracy and precision (Belkin et al., 2019; Belkin et al., 2020). The carbon sinks, belowground stocks, and carbon accumulation potential of forests are in line with the abovementioned conditions. It seems that with limited IVs, the RF algorithm elegantly avoids overfitting and multicollinearity because its variables and observations are all randomly sampled and recombined with a certain number of new submatrices for CARTs and then an ensemble is created as a result (Breiman, 2001).

In practical applications, estimating the target variables that have high uncertainties, such as the carbon accumulation potential in forests, also confirms the abovementioned perspective: RF is superior to other types of ML algorithms. For example, RF without feature selection defeated various other combinations of ML algorithms, including multilayer perceptron (MLP) (Rosenblatt et al., 1958), support vector machine (SVM) (Cortes and Vapnik, 1995), a gradient boosting decision tree called XGBoost (Chen and Guestrin, 2016), and RF (Breiman, 2001) along with feature selection methods such as SVM feature selection, RF-based feature selection, principal component analysis (PCA), and no feature selection, for estimating the carbon accumulation potential of 20 forests (Cook-Patton et al., 2020). However, the abovementioned research used 100 repetitive integrations as the Monte Carlo method. We intuitively guess that there is a stable recursive method that allows the ensemble result of the RF to “wander” more than 100 times in a high-dimensional space in an orderly manner and obtain the stable combination that minimises the loss function. Therefore, we designed the WTF sub-algorithm by combining RF and GS and making 200 repetitive integrations (This approximates the Monte Carlo method over 20,000 times).

1 488 However, a possibility of the above method is that in the process of RF, as long
2
3 489 as the out-of-bag (OOB) error rate exceeds 50% (random permutation method), the RF
4
5 490 will be unable to provide output in one of the target variables. After all, the advanced
6
7 491 and complex models or algorithms do not necessarily produce reliable results in some
8
9 492 cases (Adams & Piovesan, 2002), especially in chaotic systems. Therefore, if the simple
10
11 493 LR models of the KS approach and the WTF sub-algorithms are successively failing,
12
13 494 the last resort, the CCTs of the 3rd layer at least will guarantee the accuracy of the results.
14
15 495 The 3rd layer is using the average of all observations to estimate the target variable is
16
17 496 the only way to let the output at least in high accuracy although the precision is
18
19 497 commonly low. The KS uses different sub-methods and sub-algorithms for different
20
21 498 target variables so that the estimation result of the target variable is possibly higher
22
23 499 accurate than other methods.
24
25
26
27
28

29 500

30 31 501 **5. Conclusion**

32
33
34 502 We designed the KS approach as a new method to estimate chaotic target
35
36 503 variables (such as forest carbon stocks and sinks), and the performance of the KS
37
38 504 approach was significantly better than that of the IPCC's guideline for applied broadleaf
39
40 505 forest parameters, especially in terms of estimating irregular target variables for
41
42 506 policymakers. The linchpin is that the KS approach combined simple LR models, the
43
44 507 WTF sub-algorithm, and the CCTs for estimating different target variables that have
45
46 508 different characteristics. Because the KS approach has high operational flexibility, it
47
48 509 can also be applied in other forest types if the database is reliable. Based on collecting
49
50 510 *P. edulis* forest data from peer-reviewed articles and our field investigation, which
51
52 511 analysed soil nutrients and the PCCs of AGB and BGB to account for AGC and BGC,
53
54 512 we also found that AGC stocks and BA, BNPP and LNP, ANPP and HR, and ANPP
55
56 513 and total N in the soil correlated with each other. Furthermore, the reason why current
57
58
59
60
61
62
63
64
65

fancy models containing machine learning or deep learning algorithms are not effective in predicting the target variables of certain chaotic systems is perhaps that the total explanatory power of observable variables is less than that of the total unobservable variables. Before the advancement of science and technology in human society has transformed many unobservable variables that were originally difficult to quantify into observable variables, estimating target variables with the higher total explanatory power of observable variables by using simple linear models are the most efficient and accurate way. For estimating target variables with the lower total explanatory power of observable variables, the more effective independent variables that can be collected, the better. Then, using or designing methods with the good generating ability to estimate the divergent target become accurate speculatively.

Acknowledgements

We thank the agencies of Forestry and Forest Products Research Institute (FFPRI), Kyoto University (KU), Taiwan Forestry Research Institute (TFRI), Taiwan Forestry Bureau (TFB), and National Chung Hsing University (NCHU) for providing study sites and experimental equipment. We are very grateful to professor A. Osawa & N. Osawa, associate professor N. Okada, assistant professor M. Ataka, and Ms. N. Shiozaki at KU for environmental support. We thank retired professor B. H. Sheu and Dr. S. M. Hsu at NCHU, retired senior research fellow Dr. T. H. Chen at TFRI, and Dr. C. S. Hsu at Taiwan Research Institute (TRI) for contributing initial ideas, and professor C. C. Chang & S. H. Hsu at National Taiwan University (NTU) and Academia Sinica (AS) for conceptualising climate policy assessment. Further, we also thank Mr. M. Ohnuki, Dr. H. Schäfer, Dr. M. Kamar, Dr. R. Nakahata, Dr. A. Niazai, Ms. Y. Kihara, Dr. T. Tanabe, Mr. A. Yabuki, Mr. R. Fujii, Mr. Aye Ko Ko, Ms. Y. Xia, and Mr. T. Ikeda at KU, assistant professor B. A. T. M. Zinnatul at Begum Rokeya University, assistant

professor M. Mukai at University of Yamanashi, and professor D. Epron at University of Lorraine & KU for field works, providing post-script ideas, and rolling the corrections of the KS approach. We further thank Mr. Z. Y. Li at TFRI and retired professor C. H. Yen at NCHU for performing the soil experiments, Dr. W. Nakai at DATUM STUDIO Co. Ltd. for constructing the cluster and suggesting the method of validation, deputy director-general M. F. Chou and climate mitigation & adaptation specialist C. Y. Chang at Council of Agriculture, Executive Yuan, Taiwan (R.O.C.) for providing feedback on the user response from policymakers.

References

- Adams, J. M., Piovesan, G., 2002. Uncertainties in the role of land vegetation in the carbon cycle. *Chemosphere* 49, 805-819. [https://doi.org/10.1016/S0045-6535\(02\)00382-X](https://doi.org/10.1016/S0045-6535(02)00382-X)
- Belkin, M., Hsu, D., Ma, S., Mandal, S., 2019. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *PNAS* 116, 15849–15854. <https://doi.org/10.1073/pnas.1903070116>.
- Belkin, M., Hsu, D., Ma, S., Mandal, S., 2020. Reply to Loog et al.: Looking beyond the peaking phenomenon. *PNAS* 117, 10627–10627. <https://doi.org/10.1073/pnas.2003206117>.
- Bou Kheir, R., Greve, M. H., Bøcher, P. K., Greve, M. B., Larsen, R., McCloy, K., 2010. Predictive mapping of soil organic carbon in wet cultivated lands using classification-tree based models: The case study of Denmark. *J. Environ. Manage.* 91, 1150–1160. <https://doi.org/10.1016/j.jenvman.2010.01.001>.
- Breiman, L., 2001. Random Forests. *Mach. Learn.* 45, 5–32. <https://doi.org/10.1023/A:1010933404324>.
- Brown, B., Lugo, A. E., 1984. Biomass of tropical forests: A new estimate based on

forest volumes. *Science* 223, 1290–1293.
<https://doi.org/10.1126/science.223.4642.1290>.

Brown, S., 2002. Measuring carbon in forests: current status and future challenges. *Environ. Pollut.* 116, 363–372. [https://doi.org/10.1016/S0269-7491\(01\)00212-3](https://doi.org/10.1016/S0269-7491(01)00212-3).

Buckingham, K., Jepson, P., Wu, L. R., Ramanuja Rao, I. V., Jiang, S., Liese, W., Lou, Y. P., Fu, M. Y., 2011. The potential of bamboo is constrained by outmoded policy frames. *Ambio*. 40, 544–548. <https://doi.org/10.1007/s13280-011-0138-4>.

van Buuren, S., Groothuis-Oudshoorn, C. G. M., 2011. mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* 45(3), 1–67.
<https://doi.org/10.18637/jss.v045.i03>.

Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, 2016).
<https://doi.org/10.1145/2939672.2939785>.

Chen, T. H., Wang, D. H., Wang, S., 2016. The trend of growth characteristics of Moso bamboo (*Phyllostachys pubescens*) forests under an unmanaged condition in central Taiwan. *Taiwan J. For. Sci.* 31, 75–87.
https://kmweb.coa.gov.tw/files/document/365351/3e1488ef4c51aa766c3e83890677cbd3_v2.pdf (accessed 10 October 2021)

Cook-Patton, S. C., Leavitt, S. M., Gibbs, D., Harris, N. L., Lister, K., Anderson-Teixeira, K. J., Briggs, R. D., Chazdon, R. L., Crowther, T. W., Ellis, P. W., Griscom, H. P., Herrmann, V., Holl, k. D., Houghton, R. A., Larrosa, C., Lomax, G., Lucas, R., Madsen, P., Malhi, Y., Paquette, A., Parker, J. D., Paul, K., Routh, D., Roxburgh, S., Saatchi, S., van den Hoogen, J., Walker, W. S., Wheeler, C. E., Wood, S. A., Xu, L., Griscom, B. W., 2020. Mapping carbon accumulation potential from global natural forest regrowth. *Nature* 585, 545–550.

<https://doi.org/10.1038/s41586-020-2686-x>.
 Cortes, C., Vapnik, V., 1995. Support-vector networks. *Mach. Learn.* 20, 273–297.
<https://doi.org/10.1007/BF00994018>.
 Fukushima, K., Usui, N., Ogawa, R., Tokuchi, N., 2015. Impacts of Moso bamboo (*Phyllostachys pubescens*) invasion on dry matter and carbon and nitrogen stocks in a broad-leaved secondary forest located in Kyoto, western Japan. *Plant Species Biol.* 30, 81–95. <https://doi.org/10.1111/1442-1984.12066>.
 Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE/ACM Trans. Audio, Speech, Language Process.* PAMI-6, 721–741. <https://doi.org/10.1109/TPAMI.1984.4767596>.
 Guo, Z. D., Fang, J. Y., Pan, Y. D., Birdsey, R., 2010. Inventory-based estimates of forest biomass carbon stocks in China: A comparison of three methods. *For. Ecol. Manage.* 259, 1225–1231. <https://doi.org/10.1016/j.foreco.2009.09.047>.
 Isagi, Y., Kawahara, T., Kamo, K., Ito, H., 1997. Net production and carbon cycling in a bamboo *Phyllostachys pubescens* stand. *Plant Ecol.* 130, 41–52. <https://doi.org/10.1023/A:1009711814070>.
 Intergovernmental Panel on Climate Change, 2006. In 2006 IPCC Guidelines for National Greenhouse Gas Inventories. Volume 4: Agriculture, Forestry and Other Land Use. Institute for Global Environmental Strategies, Hayama, Kanagawa, Japan. <https://www.ipcc-nggip.iges.or.jp/public/2006gl/vol4.html> (accessed 10 October 2021).
 Intergovernmental Panel on Climate Change, 2019. 2019 Refinement to the 2006 IPCC Guidelines for National Greenhouse Gas Inventories. Volume 4: Agriculture, Forestry and Other Land Use. Intergovernmental Panel on Climate Change, Geneva, Switzerland. <https://www.ipcc-nggip.iges.or.jp/public/2019rf/index.html> (accessed 10 October 2021).

- 618 Katumbi, N. M., Kinyanjui, M. J., Jm, K., Mware, M. J., 2017. Biomass energy
619 resource of the highland bamboo (*Yushania alpina*) and its potential for sustainable
620 exploitation in southern aberdares Forest. J. Sustain. Bioenergy Syst. 7, 85–97.
621 <https://doi.org/10.4236/jsbs.2017.73007>.
- 622 Kendall, M. G., 1938. A new measure of rank correlation. Biometrika 30(1-2), 81–93.
623 <https://doi.org/10.1093/biomet/30.1-2.81>.
- 624 Kobayashi, K., Kitayama, K., Onoda, Y., 2018. A simple method to estimate the rate
625 of the bamboo expansion based on one-time measurement of spatial distribution of
626 culms. Ecol. Res. 33, 1137–1143. <https://doi.org/10.1007/s11284-018-1626-9>.
- 627 Kobayashi, K., Ohashi, M., Fujihara, M., Kitayama, K., Onoda, Y., 2022. Rhizomes
628 play significant roles in biomass accumulation, production and carbon turnover in
629 a stand of the tall bamboo *Phyllostachys edulis*, J. For. Res. (accepted)
630 <https://doi.org/10.1080/13416979.2022.2090669>.
- 631 Lee, H. Y., Chung, P. H., Wu, Y. C., Lin, T. H., Wen, T. H., 2018. Interactive spoken
632 content retrieval by deep reinforcement learning. IEEE/ACM Trans. Audio,
633 Speech, Language Process 26, 2447–2459.
634 <https://doi.org/10.1109/TASLP.2018.2852739>.
- 635 Lin, M. Y., Hsieh, I. F., Lin, P. H., Laplace, S., Ohashi, M., Chen, T. H., Kume, T.,
636 2017. Moso bamboo (*Phyllostachys pubescens*) forests as a significant carbon
637 sink? A case study based on 4-year measurements in central Taiwan. Ecol. Res. 32,
638 845– 857. <https://doi.org/10.1007/s11284-017-1497-5>.
- 639 Lobovikov, M., Schoene, D., Lou, Y. P., 2012. Bamboo in climate change and rural
640 livelihoods. Mitig. Adapt. Strateg. Glob. Chang. 17, 261–276.
641 <https://doi.org/10.1007/s11027-011-9324-8>.
- 642 Martin, A. P., Palmer, W. M., Byrt, C. S., Furbank, R. T., Grof, C. P., 2013. A holistic
643 high-throughput screening framework for biofuel feedstock assessment that

644 characterises variations in soluble sugars and cell wall composition in *Sorghum*
 645 *bicolor*. *Biotechnol. Biofuels* 6, 186. <https://doi.org/10.1186/1754-6834-6-186>.
 646 Mei, T., Fang, D., Röhl, A., Hölscher, D., 2019. Bamboo water transport assessed with
 647 deuterium tracing. *Forests* 10, 623. <https://doi.org/10.3390/f10080623>.
 648 Moore, P. D., Chapman, S. B., 1986. *Methods in Plant Ecology*. Blackwell Scientific
 649 Publications, London, GB, pp. 215–337.
 650 Nath, A. J., Lal, R., Das, A. K., Managing woody bamboos for carbon farming and
 651 carbon trading. *Glob. Ecol. Conserv.* 3, 654–663.
 652 <https://doi.org/10.1016/j.gecco.2015.03.002>.
 653 Okutomi, K., Shinoda, S., Fukuda, H., 1996. Causal analysis of the invasion of broad-
 654 leaved forest by bamboo in Japan. *J. Veg. Sci.* 7(5), 723–728.
 655 <https://doi.org/10.2307/3236383>.
 656 Pearson, K., 1901. LIII. On lines and planes of closest fit to systems of points in
 657 space. *Philos Mag* 2, 559–572. <https://doi.org/10.1080/14786440109462720>.
 658 Poeplau, C., Don, A., Vesterdal, L., Leifeld, J., Wesemael, B. V., Schumacher, J.,
 659 Gensior, A., 2011. Temporal dynamics of soil organic carbon after land-use change
 660 in the temperate zone – carbon response functions as a model approach. *Glob.*
 661 *Change Biol.* 17, 2415–2427. <https://doi.org/10.1111/j.1365-2486.2011.02408.x>.
 662 Rosenblatt, F., 1958. The perceptron: A probabilistic model for information storage
 663 and organization in the brain. *Psychol. Rev.* 65, 386–408.
 664 <https://doi.org/10.1037/h0042519>.
 665 Royston, P., 1982. Algorithm AS 181: The W test for normality. *J. R. Stat. Soc. Series*
 666 *C* 31, 176–180. <https://doi.org/10.2307/2347986>.
 667 Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., Hemingway, H., 2014.
 668 Comparison of random forest and parametric imputation models for imputing
 669 missing data using MICE: A CALIBER Study. *Am. J. Epidemiol.* 179(6), 764–774.

<https://doi.org/10.1093/aje/kwt312>.
 Shi, J., Mao, S., Wang, L., Ye, X., Wu, J., Wang, G., Chen, F., Yang, Q., 2021. Clonal
 integration driven by source-sink relationships is constrained by rhizome
 branching architecture in a running bamboo species (*Phyllostachys glauca*): A ¹⁵N
 assessment in the field. For. Ecol. Manage. 481, 118754.
<https://doi.org/10.1016/j.foreco.2020.118754>.
 Shinohara, Y., Kume, T., Ichihashi, R., Komatsu, H., Otsuki, K., 2014. Moso-bamboo
 forests in Japan: What are the effects of their area expansion on ecosystem
 services? J. Japanese For. Soc. 96, 351–361. <https://doi.org/10.4005/jjfs.96.351>.
 Silver, D., Huang, A. J., Maddison, C. J., Guez, A., Sifre, L., van der Driessche, G.,
 Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S.,
 Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M.,
 Kavukcuoglu, K., Graepel, T., Hassabis, D., 2016. Mastering the game of Go with
 deep neural networks and tree search. Nature 529, 484–489.
<https://doi.org/10.1038/nature16961>.
 Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A. J., Guez, A.,
 Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y. T., Lillicrap, T., Hui, F., Sifre,
 L., van den Driessche, G., Graepel, T., Hassabis, D., 2017. Mastering the game of
 Go without human knowledge. Nature 550, 354–359.
<https://doi.org/10.1038/nature24270>.
 Song, X. Z., Peng, C. H., Ciais, P., Li, Q., Xiang, W., Xiao, W., Zhou, G. M., Deng,
 L., 2020. Nitrogen addition increased CO₂ uptake more than non-CO₂ greenhouse
 gases emissions in a Moso bamboo forest. Sci. Adv. 6, eaaw5790.
<https://doi.org/10.1126/sciadv.aaw5790>.
 Song, X. Z., Peng, C. H., Zhou, G. M., Gu, H. H., Quan, L., Zhang, C., 2016.
 Dynamic allocation and transfer of non-structural carbohydrates, a possible

696 mechanism for the explosive growth of Moso bamboo (*Phyllostachys heterocycla*).
 697 Sci. Rep. 6, 25908. <https://doi.org/10.1038/srep25908>.
 698 Spearman, C., 1904. The proof and measurement of association between two things.
 699 Am. J. Psychol. 15, 72–101. <https://doi.org/10.2307/1412159>.
 700 Stallkamp, J., Schlipsing, M., Salmen, J., Igel, C., 2012. Man vs. computer:
 701 Benchmarking machine learning algorithms for traffic sign recognition. Neural
 702 Netw. 32, 323–332. <https://doi.org/10.1016/j.neunet.2012.02.016>.
 703 Thompson, J. R., Foster, D. R., Scheller, R., Kittredge, D., 2011 The influence of land
 704 use and climate change on forest biomass and composition in Massachusetts, USA.
 705 Ecol. Appl. 21, 2425–2444. <https://doi.org/10.1890/10-2383.1>.
 706 Toker, D., Sommer, F. T., D’Esposito, M., 2020. A simple method for detecting chaos
 707 in nature. Commun. Biol. 3, 11. <https://doi.org/10.1038/s42003-019-0715-9>.
 708 Wang, B., Waters, C., Orgill, S., Cowie, A., Clark, A., Liu, D. L., Simpson, M.,
 709 McGowen, I., Sides, T., 2018. Estimating soil organic carbon stocks using different
 710 modelling techniques in the semi-arid rangelands of eastern Australia. Ecol. Indic.
 711 88, 425–438. <https://doi.org/10.1016/j.ecolind.2018.01.049>.
 712 Wang, B., Wei, W. J., Liu, C. J., You, W. Z., Niu, X., Man, R. Z., 2013. Biomass and
 713 carbon stock in Moso bamboo forests in subtropical China: characteristics and
 714 implications. J. Trop. For. Sci. 25, 137–148. <http://www.jstor.org/stable/43595383>.
 715 Wang, S., Chen, T. H., Chang, H. C., Chung, H. Y., Li, T. I., Liu, C. P., 2009. The
 716 structures, aboveground biomass carbon storage of *Phyllostachys pubescens* stands
 717 in Huisun Experimental Forest Station and Shi-Zhuo. Quarterly J. For. Res. 34,
 718 17– 26. <https://doi.org/10.29898/SHBQ.200912.0002>.
 719 Wang, S., Chen, T. H., Chen, S. Y., Chung, H. Y., Liu, E. U., Li, T. I., Liu, C. P., 2010.
 720 Estimating aboveground biomass and carbon sequestration of Moso bamboo
 721 grown under selection cutting after two years. Quarterly J. For. Res. 32, 35–44.

722 <https://doi.org/10.29898/SHBQ.201009.0003>.
 723 Wang, S., Chen, T. H., Liu, E. U., Liu, C. P., 2020. Accessing the nursing behaviour of
 724 Moso bamboo (*Phyllostachys edulis*) on carbohydrates dynamics and
 725 photosystems. Sci. Rep. 10, 1015. <https://doi.org/10.1038/s41598-020-57643-1>.
 726 Wang, S., Kobayashi, K., Li, T. I., Hsu, C. S., Chen, T. H., Liu, C. P., Dannoura, M.,
 727 2019. Assessment of forest carbon stock based on management and environmental
 728 factors using the Multivariate Imputation by Chained Equations (MICE) and
 729 Multiple Regression (MR) approaches – a case study on Moso bamboo
 730 (*Phyllostachys edulis*). The esj Kinki Region Conference.
 731 [https://www.researchgate.net/publication/365489594_Assessment_of_forest_carbon](https://www.researchgate.net/publication/365489594_Assessment_of_forest_carbon_stock_based_on_management_and_environmental_factors_using_the_Multivariate_Imputation_By_Chained_Equations_MICE_and_Multiple_Regression_MR_approaches_-_a_case_study_on_Moso)
 732 [n_stock_based_on_management_and_environmental_factors_using_the_Multivari](https://www.researchgate.net/publication/365489594_Assessment_of_forest_carbon_stock_based_on_management_and_environmental_factors_using_the_Multivariate_Imputation_By_Chained_Equations_MICE_and_Multiple_Regression_MR_approaches_-_a_case_study_on_Moso)
 733 [ate_Imputation_By_Chained_Equations_MICE_and_Multiple_Regression_MR_ap](https://www.researchgate.net/publication/365489594_Assessment_of_forest_carbon_stock_based_on_management_and_environmental_factors_using_the_Multivariate_Imputation_By_Chained_Equations_MICE_and_Multiple_Regression_MR_approaches_-_a_case_study_on_Moso)
 734 [proaches_-_a_case_study_on_Moso](https://www.researchgate.net/publication/365489594_Assessment_of_forest_carbon_stock_based_on_management_and_environmental_factors_using_the_Multivariate_Imputation_By_Chained_Equations_MICE_and_Multiple_Regression_MR_approaches_-_a_case_study_on_Moso).
 735 Were, K., Bui, D. T., Dick, Ø. B., Singh, B. R., 2015. A comparative assessment of
 736 support vector regression, artificial neural networks, and random forests for
 737 predicting and mapping soil organic carbon stocks across an Afrotropical
 738 landscape. Ecol. Indic. 52, 394–403. <https://doi.org/10.1016/j.ecolind.2014.12.028>.
 739 Xiao, F. M., Fan, S. H., Wang, S. L., Guan, F. Y., Yu, X. J., Shen, Z. Q., 2010.
 740 Estimation of the carbon balance in Moso bamboo and Chinese fir plantation
 741 ecosystem. Scientia Silvae Sinicae 46, 59–65. [https://doi.org/10.11707/j.1001-](https://doi.org/10.11707/j.1001-7488.20101108)
 742 [7488.20101108](https://doi.org/10.11707/j.1001-7488.20101108).
 743 Xiao, F. M., Fan, S. H., Wang, S. L., Xiong, C. Y., Shen, Z. Q., 2009. Soil carbon
 744 cycle of *Phyllostachys edulis* plantation in Huitong region, Hunan Province.
 745 Scientia Silvae Sinicae 45, 11–15. <https://doi.org/10.11707/j.1001-7488.20090603>.
 746 Yen, T. M., Ji, Y. J., Lee, J. S., 2010. Estimating biomass production and carbon
 747 storage for a fast-growing makino bamboo (*Phyllostachys makinoi*) plant based on

the diameter distribution model. *For. Ecol. Manage.* 260, 339–344.
<https://doi.org/10.1016/j.foreco.2010.04.021>.

Yen, T. M., Lee, J. S., 2011. Comparing aboveground carbon sequestration between Moso bamboo (*Phyllostachys heterocycla*) and China fir (*Cunninghamia lanceolata*) forests based on the allometric model. *For. Ecol. Manage.* 261, 995–1002. <https://doi.org/10.1016/j.foreco.2010.12.015>.

Youkhana, A. H., Ogoshi, R. M., Kiniry, J. R., Meki, M. N., Nakahata, M. H., Crow, S. E., 2017. Allometric models for predicting aboveground biomass and carbon stock of tropical perennial C₄ grasses in Hawaii. *Front. Plant Sci.* 8, 650. <https://doi.org/10.3389/fpls.2017.00650>.

Yuen, J. Q., Fung, T., Ziegler, A. D., 2017. Carbon stocks in bamboo ecosystems worldwide: Estimates and uncertainties. *For. Ecol. Manage.* 393, 113–138. <https://doi.org/10.1016/j.foreco.2017.01.017>.

Zhang, H., Guan, D. S., Song, M. W., 2012. Biomass and carbon storage of Eucalyptus and Acacia plantations in the Pearl River Delta, South China. *For. Ecol. Manage.* 277, 90–97. <https://doi.org/10.1016/j.foreco.2012.04.016>.

Zhao, J. L., Kang, F. F., Wang, L. X., Yu, X. W., Zhao, W. H., Song, X. S., Zhang, Y. Z., Chen, F., Sun, Y., He, T. F., Han, H. R., 2014. Patterns of biomass and carbon distribution across a chronosequence of Chinese pine (*Pinus tabulaeformis*) forests. *PLoS One* 9, e94966. <https://doi.org/10.1371/journal.pone.0094966>.

Funding: This work was supported by JSPS KAKENHI [grant numbers JP 15H04513, JP 19J11336, and JP 20J15519], and by the Taiwan Forestry Research Institute, Council of Agriculture, Executive Yuan, Taiwan (R.O.C.) [grant number 99AS-8.3.2-F-G1].

Video

1 774 The briefing video was revealed at <https://www.youtube.com/watch?v=l8SKe-K6ypY>

2
3 775

4
5 776 **Data and materials availability**

6
7
8 777 All data needed to evaluate the conclusions in the paper are present in the paper and/or

9
10 778 the Supplementary Materials. We revealed R codes of the WTF sub-algorithm for

11
12 779 estimating irregular target variables at https://github.com/gn03138868/WTF_sub-

13
14
15 780 [algorithm](https://github.com/gn03138868/WTF_sub-)

16
17 781

Abstract

Forest carbon stocks and sinks (CSSs) have been widely estimated using climate classification tables and linear regression (LR) models with common independent variables (IVs) such as the average diameter at breast height (DBH) of stems and root shoot ratio. However, this approach is relatively ineffective when the explanatory power of IVs is lower than that of unobservable variables. Various environmental and anthropogenic factors affect target variables that cause the correlation between them to be chaotic. Here, we designed a knife set (KS) approach combining LR models and the wandering through random forests (WTF) algorithm and applied it in a specific case of *Phyllostachys edulis* (Carrière) J. Houz. (*P. edulis*) forests, which have an irregular relationship between their belowground carbon (BGC) stocks and average DBH. We then validated the KS approach performed by cluster computing to estimate the aboveground carbon (AGC) and BGC stocks and the total net primary production (TNPP). The estimated CSSs were compared to the benchmark of the methodology that applied Tier 1 in the Intergovernmental Panel on Climate Change (IPCC) Guidelines for National Greenhouse Gas Inventories via 10-fold cross validation, and the KS approach significantly increased precision and accuracy of estimations. Our approach provides general insights to accurately estimate forest CSSs relying on evidence-based field data, even if some target variables are divergent in specific forest types. We also pointed out the reason why current fancy models containing machine learning (ML) or deep learning algorithms are not effective in predicting the target variables of certain chaotic systems is perhaps that the total explanatory power of observable variables is less than that of the total unobservable variables. Quantifying unobservable variables into observable variables is a linchpin of future works related to chaotic system estimation.

Keywords. Aboveground net primary production; belowground net primary production; Chaotic estimation; Hybrid machine learning; *Phyllostachys edulis*; Wandering through random forests

1. Introduction

Forest carbon stocks and sinks (CSSs) are usually estimated based on linear regression (LR) models. LR models work well for aboveground CSSs but are relatively ineffective for belowground CSSs. This is mainly because the relationships between the target variable (TV) and independent variables (IVs) are commonly chaotic or divergent and appear everywhere in natural ecosystems (Toker et al., 2020). There is large variability in belowground CSSs estimates among sites and high estimate uncertainties, which can be attributed to environmental and anthropogenic factors such as temperature, precipitation, soil chemical properties, biotic communities, nutrient availability, management intensity, and ecosystem productivity (Poeplau et al., 2011). Thus, if there are many unobservable IVs, it will increase the proportion of the variance in the target variable when using LR-based models to estimate belowground CSSs in forests (Thompson et al., 2006; Zhao et al., 2014).

An irregular belowground forest type mentioned above, *P. edulis* (Synonyms: *P. pubescens*, *P. heterocycla*) is a giant bamboo species, has a high growth rate, and tends to aggressively invade neighbouring vegetation in temperate zones; these invasions have been frequently observed in East Asian countries such as Japan and China (Okutomi et al., 1996; Wang et al., 2013; Lin et al., 2017; Kobayashi et al., 2018) due to *P. edulis* has rhizome systems, and can transport carbohydrates (Zhai et al., 2021) and nutrients (Shi et al., 2021) to others. This may be also one of the main reasons that belowground biomass, aboveground net primary production (ANPP), and belowground net primary production (BNPP) are divergent in *P. edulis* forests. Since *P. edulis* grows

1 53 rapidly, some studies have highlighted its high potential as an efficient carbon stock
2
3 54 (Wang et al., 2009; Yen and Lee, 2011; Nath et al., 2015; Chen et al., 2016) and carbon
4
5 55 sink (Lin et al., 2017). Efficient CSSs are vital to terrestrial ecosystems and should be
6
7 56 included in debates in the agriculture, forestry and other land use (AFOLU) sector of
8
9
10 57 the national inventory report (NIR), which is submitted to the United Nations
11
12 58 Framework Convention on Climate Change (UNFCCC) by all Annex I countries.
13
14 59 However, bamboos have not been sufficiently considered in NIRs (Lobovikov et al.,
15
16 60 2012; Yuen et al., 2017) because they are botanically classified as grasses and not trees
17
18 61 (Buckingham et al., 2011). Furthermore, the carbon sink of bamboos in the
19
20 62 Intergovernmental Panel on Climate Change (IPCC) Guidelines for the NIR has been
21
22 63 considered only for Africa and not for Asia and Central America (Katumbi et al., 2017;
23
24 64 Dong et al., 2019). Providing accurate estimates of CSSs is one of the current priorities
25
26
27
28
29 65 in climate policy.

30
31 66 For estimating the BGC stocks from the aboveground biomass and carbon
32
33 67 storage, mainstream methods are based on biomass expansion factors (BEFs). Such
34
35 68 BEFs have been widely recommended for estimating forest biomass and carbon storage
36
37 69 at a regional or national level (Brown and Lugo, 1984; Brown, 2002; Aalde et al., 2006;
38
39 70 Guo et al., 2010; Zhang et al., 2012; Dong et al., 2019). However, the above methods
40
41 71 are based on linear statistical models. The methods may be suited for estimating
42
43 72 aboveground carbon (AGC) stocks and ANPP but not for estimating BGC and BNPP
44
45 73 in specific cases of belowground structures (Kobayashi et al., 2022). The linear
46
47 74 statistical model does have limitations. The main reason is that the key factors affecting
48
49 75 the growth of forests are different in different regions, and the value range of the factors
50
51 76 is difficult to determine. That's why these methods often perform imprecisely because
52
53 77 target variables and IVs are distributed in divergent structures, as implied by the
54
55 78 existence of unobservable variables.
56
57
58
59
60
61
62
63
64
65

Hybrid algorithms and machine learning (ML) models are potential methods for dealing with divergent and non-monotonic correlations in many research fields (Geman and Geman, 1984; Stallkamp et al., 2012; Silver et al., 2016; Silver et al., 2017; Lee et al., 2018). The rapid development of ML in recent years has enabled the building of models and algorithms for reliably estimating the carbon accumulation potential (Cook-Patton et al., 2020) and soil carbon (SC) storage despite a divergence between target variables and IVs (Bou Kheir et al., 2010; Were et al., 2015; Wang et al., 2018). For instance, Wang et al. (2018) used a genetic algorithm and stepwise regression to select features as IVs and then used a boosted regression tree and random forest (RF) to predict soil organic carbon (SOC). The results of the study indicated that RF outperformed boosted regression tree in the prediction of SOC in rangelands and that feature selection increased model efficiency and accuracy. Furthermore, Were et al. (2015) compared the performance of support vector regression (SVR), artificial neural networks (ANNs), and RF for predicting the SOC across an Afromontane landscape. Even though ML models have been applied in predicting SOC, ML has rarely been combined with simple LR-based models to assess the CSSs in forests. The reason we combined LR and ML-based models is that ecological processes and relating values, including the masses of CSSs, are fundamentally controlled by universal physiological phenomena, physical structures and chemical processes; therefore, simple formulae may be better suited for the fundamental estimation models than the simple application of standard ML structures (which do not involve logic of physiological and physical backgrounds, may result in overfitting for unknown [extrapolation] situations).

As *P. edulis* forests have IVs with regular and monotonous correlations in the aboveground part but irregularly correlated IVs in the belowground part, they are suitable for designing a model that combines LR and ML to estimate target variables such as CSSs. Simple LR models based on correlations between morphological

characters and biomass, which are usually described as allometric models (Martin et al., 2013; Youkhana et al., 2017), had better performance than ML models when estimating highly correlated variables such as AGC and BA (Wang et al., 2019), of which the latter is defined as the total cross-sectional area of all stems or culms in a stand measured at breast height per unit of land area. In contrast, the above example of BGC reveals that most variables in the forest ecosystem may have a high level of complexity and that many unobservable variables could not be observed. Compared with highly correlated variable pairs, there are usually many factors affecting the target variable in the terrestrial carbon cycle when the target variable and the IV are relatively divergent.

To design a well-considered algorithm without overfitting, we observed the various steps of the carbon cycle and previous studies related to *P. edulis* forests. Then, we notice that collecting all variables in one single study is often impossible. Because studies commonly have focused on different purposes such as the estimation of aboveground biomass and the impacts on different management treatments (Wang et al., 2010; Chen et al., 2016), carbon dynamics and productivity (Isagi et al., 1997; Lin et al., 2017), and productivity and the rate of carbon dioxide absorption by applying fertilisers (Song et al., 2020), which causes the raw data sheet to have many missing values always. Having a model that accepts any form of missing data input is very friendly to policymakers, and its usability is also greatly improved. In this study, to overcome the above difficulties, we aimed to design a knife set (KS) approach that combined LR models and wandering through random forests (WTF) algorithm which was similarly used in previous study (Wang et al., 2019) and a generalised Multivariate Imputation by Chained Equations (MICE) method (van Buuren & Groothuis-Oudshoorn, 2011; Shah et al., 2014) that allows the use of nonlinear models by conducting the random forests (Breiman, 2001) and Gibbs sampling (Geman and Geman, 1984), and CCTs for estimating the AGC, BGC, ANPP, and BNPP of *P. edulis*

forests in East Asia.

2. Materials and Methods

2.1. Data collection from study sites and published research articles

2.1.1. Biomass investigation

Field investigations were conducted in *P. edulis* forests at the Forestry and Forest Products Research Institute (FFPRI), Kansai Research Centre (34°56'28.3"N 135°46'20.5"E), ~~Kyoto Prefecture, Japan,~~ and in the Huisun Forest Area, Experimental Forest Management Office, National Chung Hsing University (24°04'28.0"N 121°00'41.0"E), ~~Nantou County, Taiwan,~~ between 2008 and 2019 (Fig. S1). The biomass investigation contained two major parts, aboveground and belowground, of which the former included culms, branches, and leaves and the latter included stumps, rhizomes, and roots. More information on these sites is shown in Dataset S1.

2.1.2. Analysing carbon stocks in *P. edulis* forests

Samples for carbon storage collected in Taiwan via the above steps were milled into a powder using a grinder and sieved through a standard sieve of mesh size 0.149 mm (mesh No. 100). The obtained powders were analysed for their percent carbon content (PCCs) (the same as carbon fractions), which was the ratio of the carbon weight to dry weight in the plant bodies (Wang et al., 2009; Yen et al., 2010; Yen and Lee, 2011) and in the leaves, branches, culms, stumps, rhizomes, roots, and soil carbons (SCs) with an elemental analyser (CHNOS Elemental Analyser, Vario EL, Germany); we used between 3 and 9 duplicates per sample. The carbon storage in each component of *P. edulis* forests was calculated by multiplying the biomass by the corresponding PCCs shown in Dataset S1. Furthermore, the PCCs in Japan were used in previously

published data around Kyoto, Japan (Fukushima et al., 2015).

158

2.1.3. Analyses of soil nutrient contents and soil particle size

In total, 18 samples per bamboo forest study site were further analysed for carbon and nutrient contents at Mountain Ali and Huisun Experimental Forest Station in Taiwan. Nine soil core (100 ml) samples were collected for nutrient contents from 0-10 cm and 10-30 cm separately in the same *P. edulis* forests. At the same time, 3 cores were collected to determine the soil bulk density and calculate the total amount of soil nutrients at each study site. Soil pH (H₂O) (soil to water ratio, 1:1) and pH (KCl) (soil to KCl ratio, 1:2.5) were determined by a glass electrode (WTW/Germany PH537). The particle size analysis of the soil was measured using a hydrometer method. After the samples were naturally air-dried, roots and soil were separated and then sieved through a 2-mm sieve to analyse the C and nutrient contents (N, P, K, Ca, Mg, P). Total N was determined with the same elemental analyser (CHNOS Elemental Analyser, Vario EL, Germany) mentioned above. After the ammonium acetate method was used to measure cation exchange capacity, inductively coupled plasma optical emission spectrometry (inductively coupled plasma emission spectrometry, ICP-AES, Leeman Labs. the USA) was used to determine the contents of other effective nutrients (Ca, K, Mg, P) (Moore and Chapman, 1986).

2.1.4. Data collection: literature surveys

Fifty-five peer-reviewed journal articles on CSSs and many IVs were collected in *P. edulis* ecosystems from East Asia. Together with our field survey data described above, 83 observations from 361 plot data from 47 study sites were collected in a wide geographic area that covered the whole area of naturalised *P. edulis* forests (Dataset S1). In the database, there were 85 variables, including 6 management factors described by

dummy variables, 2 temporal factors, 36 environmental factors, and 43 growth traits overall described in Table 1 and defined in Table S2.

2.2. Constructing an algorithm for estimating AGC, BGC, ANPP, and BNPP

The KS flowchart shown in Fig. 1 illustrates the procedure used in this study for ensemble learning, the estimation of DVs and the details of practical guidelines in the R programming language via cluster or cloud computing approaches described in the Method S1. All the statistical analysis and the hybrid algorithm were performed using R-4.0.0 for Windows (64 bit) and the codes are provided in Method S1. We input the dataset of non-target variables including management, environmental factors, and growth traits as horizontal rows and observations as vertical columns in the matrix. The dataset was separated into a training dataset (90%) and validation dataset (10%) for k-fold cross-validation ($k=10$). The detail for determining the value of k is described in Method S1.

2.2.1. Linear Regression (LR)-Layer 1

The priority of the logic of statistical analysis cannot be higher than that of ecological physiology in LR models. Thus, we mined the variables that were highly correlated with each other, as supported by pieces of evidence of ecological physiology. Then, variables were partially inputted via several LR models based on the above evidence. The major reason for designing LR models to estimate some parts of missing values that formed the output of the layer 1 as the input of the layer 2 was because different studies have different purposes when applied to *P. edulis* forests and all the variables in each study site are not fully recorded in this data matrix.

To increase the stability of the simple algorithm performed by the data matrix collected

via our investigation and peer-reviewed journal articles, we used the Pearson correlation method (Pearson, 1901) to analyse the correlation between DVs and IVs (Table S3), constructed LR models for the target variables such as AGC, BGC, and ANPP, and then predicted them via LR models. Output dataset included target variables.

2.2.2. Wandering through the random forest (WTF) sub-algorithm - Layer 2

The predicted values came from layer 1 if the Pearson correlation coefficient $R^2 > 0.85$. And re-inputted these predicted values to the input of layer 2. A sub-algorithm of the KS, RF ensemble ML (Breiman, 2001) with GS (Geman and Geman, 1984) [also named wandering through the random forests sub-algorithm (WTF sub-algorithm) because it was like that the random forests algorithm wandered through the high-dimensional space] was used for 200 duplicates performed by cluster computing consisting of two i5-9400F CPU @ 2.90 GHz 2.90 GHz, one i3-8100 CPU @ 3.60 GHz 3.60 GHz, and four i9-9900K CPU @ 3.60 GHz 3.60 GHz. [It is also possible to operate cloud computing via Amazon Web Services (AWS) if the operator or policymaker wants to compute faster in this methodology.] The pseudocodes of the WTF sub-algorithm were described in Table S1. Following this, we found the most stable combination of duplicates, maximum iteration times of GS, and CART numbers of all matrices by minimising the quadratic loss function [similar to the lowest mean square error (MSE)] compared to the mean of matrices. Equation (1) illustrates the quadratic loss function of the mean of matrices:

$$\lambda(M) = \frac{1}{n} \sum_{i=1}^n (Pi - Mi)^2 \quad (1)$$

where λ is the quadratic loss function of training the WTF sub-algorithm; Pi is the predicted output value; Mi is the average value in each cell of the formed replicate datasets during iterations; and n is the number of observations.

234 The input validation dataset removed target variables into the trained ensemble
 235 matrix. The WTF sub-algorithm was run using the most stable combination to predict
 236 the target variables via cluster computing again, and all duplicate matrices (an N-
 237 dimensional array) were averaged as one dataset by stacking. Finally, the dataset that
 238 included the estimated values of target variables was output. The output of layer 2 was
 239 the input of layer 3.

240

241 2.2.3. Climate classification tables (CCTs)– Layer 3

242 To estimate target variables that could not be output from the LR models and
 243 the WTF sub-algorithm, we made CCTs for the ratio of BGC to AGC (Table S4), Tier
 244 1-estimated values of target variables (Table S5) via Dataset S1 followed the definition
 245 from the Köppen-Trewatha climate groups and IPCC’s method (Aalde et al., 2006) for
 246 *P. edulis* forests. Then, the estimated values in the rest of the target variables were input
 247 as the output of Layer

248

249 2.3. Statistical analysis & cross-validation of the knife set (KS) approach

250 The benchmark of the IPCC guideline and the hybrid algorithm that combined
 251 LR and RF with GS were compared by MAPE, RMSE, and MAE. If the target variables’
 252 MAPE, RSME, and MAE predicted by the LR models were higher than those predicted
 253 by the hybrid algorithm, we used the LR models’ predicted values of the target variables;
 254 otherwise, we used the hybrid algorithm’s predicted values of the target variables as the
 255 final output. Equations (2) to (5) illustrate the MAPE, RMSE, MAE, and ME,
 256 respectively:

$$257 \quad \text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{(Pi - Oi)}{Oi} \right| \quad (2)$$

258 where Pi and Oi are the predicted output and observed dataset values, respectively, and

n is the number of observations. MAPE is used to compare the percentage of error of the algorithm from each single estimated value that is not affected by the unit.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Pi - Oi)^2} \quad (3)$$

where Pi , Oi , n are the same as above. RMSE is used to compare the intuitive error between the algorithm's mean estimated values and the actual observations (that is, including the concept of the unit), which indicates that a lower RMSE signals a higher accuracy and precision of the algorithm and fewer outlier values.

$$MAE = \frac{1}{n} \sum_{i=1}^n |Pi - Oi| \quad (4)$$

where Pi , Oi , n are the same as above. MAE transfers the negative errors to positive and is used to intuitively compare the error between the actual value and the estimation from the algorithm (that is, the concept of a unit is included).

$$ME = \frac{1}{n} \sum_{i=1}^n (Pi - Oi) \quad (5)$$

where Pi , Oi , n are the same as above. ME is used to compare the error between the actual value and the values estimated by the algorithm. (that is, the concept of the unit is included, and the positive error and the negative error offset each other, which means that if the ME is low, the accuracy is high even though the precision may be low.) Through the above equations and parameters, we revealed the performance of the benchmark and the KS approach via 10-fold cross-validation. Further details of the KS algorithm are revealed in the Method S1.

3. Results

3.1. Correlations of target variables in carbon stocks with independent variables

3.1.1. AGC with BA

The noteworthy results showed that there was a strong significant correlation between AGC and BA ($R^2=0.878$, p -value <0.001) (Fig. 3a; Table S3) in *P. edulis* forests. AGC was significantly related to the mean diameter at breast height (DBH) or culm density even though their R^2 values were low (0.036 and 0.322, respectively) (Fig. 3b, 3c).

3.1.2. BGC with other IVs

As our speculation, BGC has no significant correlation with BA ($R^2 = 0.055$), culm density ($R^2 = 0.043$), and mean DBH ($R^2 = 0.0038$), whose distribution was divergent and irregular (Table S3; Fig. 4). BGC cannot be determined with the method of layer 1 because it has no suitable IVs in LR models (Table S3).

3.2. Correlations of target variables in carbon sinks with independent variables

3.2.1. ANPP with heterotrophic respiration (HR), total nitrogen (N) in the soil, and litterfall

The correlation between ANPP and heterotrophic respiration (HR) was significant ($R^2=0.918$, p -value <0.01) (Fig. 3d). However, litterfall was only slightly significantly correlated with HR (p -value <0.05) (Fig. 3e) and was not significantly correlated with ANPP (p -value >0.05) (Fig. 3f). Furthermore, ANPP was significantly correlated with total N in the soil ($R^2=0.981$, p -value <0.05) (Fig. 3g; Table S3).

3.2.2. BNPP with leaf net production (LNP)

Correlations between BNPP and leaf net production (LNP) were strongly significant ($R^2=0.966$, p -value <0.001) (Fig. 3h). We believe that soil availability

cations and nutrients may be related to BNPP, however there is not much data published on soil availability cations and BNPP (Dataset S1). This cannot be analysed using Pearson's correlation.

3.3. Performance of the knife set (KS) approach

The performance of this KS approach is better than that of the IPCC's Tier 1 method, which employs the default factors provided by the IPCC and has been widely used for estimating broadleaf forest target variables, as a benchmark certified by k-fold cross-validation (k=10). The results of k-fold cross-validation also indicate that if the target and IVs have a significant correlation, the performance of estimating the target variable via LR models is excellent (Table S6). Regardless of the mean absolute error (MAE), root-mean-square error (RMSE), mean error (ME) or mean absolute percentage error (MAPE), the KS is generally lower than the benchmark (IPCC Tier 1) in estimations of 10-fold cross validation (Table S6). Further, the precision and the accuracy of the KS are higher than that of the benchmark in estimations of 10-fold cross validation (Fig. 5 & 6), and are similar with holdout testing (Fig. 7).

4. Discussion

4.1. The estimation of target variables and their correlations with other independent variables used in linear regression and the output from the layer 1

To establish the LR models of the layer 1 of the KS approach, we collected various variables from peer-reviewed articles (Dataset S1), initially determined the annual carbon dynamics for different ecological zones of *P. edulis* forests (Fig. 2), and analysed the correlation between target variables and other IVs (Table S3).

4.1.1. AGC and BA

AGC was positively and significantly correlated with BA and had an $R^2 > 0.85$ across all studies (Fig. 3a; Table S3). When IVs and DVs are highly correlated with small errors, it is usually appropriate to construct linear models with higher accuracy and precision that satisfy this hybrid algorithm to save computing power. For example, BA is suitable for estimating AGC stocks in *P. edulis* forests. Therefore, the AGC and BA were suitable for constructing a simple and highly accurate linear model, and the results were input into the data matrix as layer 1 of the hybrid algorithm. Shinohara et al. (2014) also found the similar results and concluded similar linear models to estimate AGC and BGC via BA in *P. edulis* forests across Japan. Although AGC was significantly related to the mean diameter at breast height (DBH) or culm density, their R^2 values were low (Fig. 3b, 3c) and obtained similar results in *P. edulis* forests based on other research data (Yuen et al., 2017); this approach was suitable for building linear or monotonic regression models to estimate target variables which is not divergent to IVs.

4.1.2. BGC and other IVs

When the target variable and the IVs are divergent and irregular, various factors affecting the target variable exist in the real terrestrial carbon cycle, such as the BGC stocks of *P. edulis* forests, which are affected by annual rainfall (p -value < 0.05), LNP (p -value < 0.05), and selective cutting and weeding (p -value < 0.05) (Table S3). Even though the DV and the IVs were significantly correlated in the above case, their R^2 values were usually less than 0.6, which implied that there were many more unobservable variables and that the fitting of the linear models made by each pair of variables was imprecise and not suitable for estimating the target variable, although these variables still had statistical significance as long as the number of samples was sufficient (Table S3).

359

360 4.1.3. ANPP, heterotrophic respiration (HR), total nitrogen (N) in the soil, and

361 litterfall

362 ANPP as a target variable was significantly correlated with HR and slightly
363 significantly correlated with litterfall. This result might imply that the HR of
364 microbiome communities in the soil decomposed organic matter supplied from litterfall
365 increased and that the nutrient resupply to *P. edulis* forests caused the ANPP to increase.
366 Another possibility was that total nitrogen (N) in the soil, which increased due to the
367 decomposition of litter or fertilisation, drove the activity of the microbiome further to
368 activate ammonification and nitrification in the N cycle and caused the HR to increase.
369 N addition significantly increased the wood biomass, carbon stocks, and ANPP, and the
370 decomposition of SOC was confirmed early by controlled field experiments (Song et
371 al., 2020).

372 However, because it is difficult to obtain data on both ANPP and total N in the
373 soil at the same time, only 3 data points were collected in *P. edulis* forests (Fig. 3g).
374 Only 2 peer-reviewed studies recorded the total N in soil and the HR at the same study
375 site (Xiao et al., 2009; Xiao et al., 2010). ANPP had a slight significant correlation with
376 the total N in the soil. The relationships among soil nutrients, HR and ANPP have the
377 potential for discovering insights, and further related studies might be obtained in the
378 future, even if they have almost reached the threshold of significant correlations.

379

380 4.1.4. BNPP and leaf net production (LNP)

381 Correlations between BNPP and LNP were strongly significant (Fig. 3h)
382 perhaps implied that as more leaves formed, more roots and rhizomes grew because
383 leaves of *P. edulis* produced sugars and supplied them to the developing roots and
384 rhizomes.

Although the BGC might need to be estimated using other reliable methods, the above discussion rationally proved that the AGC could be estimated by the BA, ANPP by HR, and BNPP could be estimated by LNP via LR models applied in the layer 1 of the KS approach. If there were target variables that were not estimated by the first layer of the KS approach, the remaining target variables were also estimated using the following layer 2 and layer 3 (Fig. 1).

4.2. Complexity and unobservable variables in the carbon cycle of forest ecosystems and a potential method to overcome overfitting, the WTF sub-algorithm

Reasons why it is difficult to find independent variables of BGC with high explanatory power (Fig. 4) is because that carbon cycling and growth characteristics in *P. edulis* forests have complicated interactions with each other. For example, the photosynthetic systems in the leaves and epidermis of culms are activated by sunlight to respond to light reactions, a series of electron transfer processes occur, and then ATP and NADP are produced. Meanwhile, the stomata on leaf surfaces absorb carbon dioxides, which supply mesophylls to synthesise carbohydrates in the air and on earth (Wang et al., 2020). These carbohydrates provide culms to form structural carbohydrates (SC) and accumulate dry matter as a part of biomass (Yen and Lee, 2010; Yuen et al., 2017) and supply the respiratory demands of leaves, branches, culms, stumps, rhizomes, roots, and bamboo shoots. In particular, unlike other normal temperate forests, the carbohydrates produced by mature *P. edulis* may be transferred to other parts of young shoots through the rhizome system, resulting in the development of expanded belowground parts and the production of new bamboo shoots (Song et al., 2016), further helping young individuals resist abiotic stresses (Wang et al., 2020). When *P. edulis* forests continue to grow, coarse woody debris (such as old bamboo) and

litterfall gradually return to the soil. The belowground roots, rhizomes and stumps will be decomposed by fungi and microbes, and then carbon will return to the soil and the atmosphere as it is released from organic matter. Each part of the above process may be interfered with by multiple environmental and anthropogenic factors.

Because *P. edulis* forests have a special belowground system that may transfer carbohydrates (Song et al., 2016; Wang et al., 2020), water (Mei et al., 2019), and nutrients (Shi et al., 2021) to young individuals or develop the belowground system itself, their rhizomes and roots are not as stable as those in normal temperate forests and are easily affected by various anthropogenic and environmental factors, which has resulted in difficulty in accurately predicting BGC and irregular root shoot ratios (Table S3).

Each value of a variable is the result of all the other subtle variables interacting with each other. As long as there are infinitely various IVs, it is possible to fully deduce the value that the real target variable should be given, and this value is close to the means of observations. However, some of the current ML models often use only limited IVs to estimate the target variable, which leads to the problem of overfitting (It means the model is accurate in estimating target variables in validation but causes considerable deviation when it is applied in practice). To construct an algorithm that considers all the known variables as much as possible without overfitting, we used the WTF sub-algorithm in the 2nd layer of the KS approach to estimate the values of target variables for which the verification error rate was below 50% (Fig. 1). If there were target variables that were not estimated by the first two layers of the KS approach, the remaining target variables were also estimated using the CCTs (Table S4, S5) as the 3rd layer—the final output as results (Fig. 1).

4.3. Comparing the performance of the knife set (KS) approach with IPCC's Tier

1 method

Our empirical results (Table S6; Fig. 5, 6, 7) prove that the KS approach can improve the accuracy and precision in estimations of the carbon stocks and sinks of *P. edulis* forest without over-fitting. The reason why the KS performance is better than the results of IPCC Tier 1 is that the KS approach provides different and reasonable sub-algorithms for the different characteristics of target variables. For instance, target variables will be estimated by LR, WTF, and CCTs in the KS approach if the uncertainties of target variables are below 15%, between 15-50%, and over 50%, respectively. Another reason is that IPCC's Tier 1 approach may be simplifying assumptions about some carbon pools and can only provide a general but relevantly low precise and accurate approach (Such as IPCC's Tier 1 does not include reference data for scattered bamboo forests but borrow values from broadleaf forests). From observing the KS approach, if there is no significant relationship or a low R^2 between the target and IVs, the WTF sub-algorithm approach will be indispensable for estimating target variables to compensate for the shortcomings of LR models (Fig. 1). Furthermore, to reduce the disadvantage of the data matrix with many missing values and increase the stability and accuracy of the KS approach whose DVs and IVs have divergent but highly accurate correlations, we tried to identify the variables that were highly correlated with each other, as supported by ecophysiological evidence, and used LR models to estimate some of the missing values to form an output of the layer 1 for use as the input of the layer 2 for the WTF sub-algorithm.

Designing the WTF sub-algorithm by combining RF (Breiman et al., 2001) and Gibbs sampling (GS) (Geman and Geman, 1984) is suitable for treating the problem of divergent correlation between the target and IVs. The main reason is that any design that changes the weight of the IV always results in an overfit when the effect of the unobservable variable of the target variable is greater than that of the IV. Unless the

number of variables exceeds the interpolation threshold, it is possible to use mainstream ML without weight changing designs to train algorithms with higher accuracy and precision (Belkin et al., 2019; Belkin et al., 2020). The carbon sinks, belowground stocks, and carbon accumulation potential of forests are in line with the abovementioned conditions. It seems that with limited IVs, the RF algorithm elegantly avoids overfitting and multicollinearity because its variables and observations are all randomly sampled and recombined with a certain number of new submatrices for CARTs and then an ensemble is created as a result (Breiman, 2001).

In practical applications, estimating the target variables that have high uncertainties, such as the carbon accumulation potential in forests, also confirms the abovementioned perspective: RF is superior to other types of ML algorithms. For example, RF without feature selection defeated various other combinations of ML algorithms, including multilayer perceptron (MLP) (Rosenblatt et al., 1958), support vector machine (SVM) (Cortes and Vapnik, 1995), a gradient boosting decision tree called XGBoost (Chen and Guestrin, 2016), and RF (Breiman, 2001) along with feature selection methods such as SVM feature selection, RF-based feature selection, principal component analysis (PCA), and no feature selection, for estimating the carbon accumulation potential of 20 forests (Cook-Patton et al., 2020). However, the abovementioned research used 100 repetitive integrations as the Monte Carlo method. We intuitively guess that there is a stable recursive method that allows the ensemble result of the RF to “wander” more than 100 times in a high-dimensional space in an orderly manner and obtain the stable combination that minimises the loss function. Therefore, we designed the WTF sub-algorithm by combining RF and GS and making 200 repetitive integrations (This approximates the Monte Carlo method over 20,000 times).

1 488 However, a possibility of the above method is that in the process of RF, as long
2
3 489 as the out-of-bag (OOB) error rate exceeds 50% (random permutation method), the RF
4
5 490 will be unable to provide output in one of the target variables. After all, the advanced
6
7 491 and complex models or algorithms do not necessarily produce reliable results in some
8
9 492 cases (Adams & Piovesan, 2002), especially in chaotic systems. Therefore, if the simple
10
11 493 LR models of the KS approach and the WTF sub-algorithms are successively failing,
12
13 494 the last resort, the CCTs of the 3rd layer at least will guarantee the accuracy of the results.
14
15 495 The 3rd layer is using the average of all observations to estimate the target variable is
16
17 496 the only way to let the output at least in high accuracy although the precision is
18
19 497 commonly low. The KS uses different sub-methods and sub-algorithms for different
20
21 498 target variables so that the estimation result of the target variable is possibly higher
22
23 499 accurate than other methods.
24
25
26
27
28

29 500

30 31 501 **5. Conclusion** 32

33
34 502 We designed the KS approach as a new method to estimate chaotic target
35
36 503 variables (such as forest carbon stocks and sinks), and the performance of the KS
37
38 504 approach was significantly better than that of the IPCC's guideline for applied broadleaf
39
40 505 forest parameters, especially in terms of estimating irregular target variables for
41
42 506 policymakers. The linchpin is that the KS approach combined simple LR models, the
43
44 507 WTF sub-algorithm, and the CCTs for estimating different target variables that have
45
46 508 different characteristics. Because the KS approach has high operational flexibility, it
47
48 509 can also be applied in other forest types if the database is reliable. Based on collecting
49
50 510 *P. edulis* forest data from peer-reviewed articles and our field investigation, which
51
52 511 analysed soil nutrients and the PCCs of AGB and BGB to account for AGC and BGC,
53
54 512 we also found that AGC stocks and BA, BNPP and LNP, ANPP and HR, and ANPP
55
56 513 and total N in the soil correlated with each other. Furthermore, the reason why current
57
58
59
60
61
62
63
64
65

fancy models containing machine learning or deep learning algorithms are not effective in predicting the target variables of certain chaotic systems is perhaps that the total explanatory power of observable variables is less than that of the total unobservable variables. Before the advancement of science and technology in human society has transformed many unobservable variables that were originally difficult to quantify into observable variables, estimating target variables with the higher total explanatory power of observable variables by using simple linear models are the most efficient and accurate way. For estimating target variables with the lower total explanatory power of observable variables, the more effective independent variables that can be collected, the better. Then, using or designing methods with the good generating ability to estimate the divergent target become accurate speculatively.

Acknowledgements

We thank the agencies of Forestry and Forest Products Research Institute (FFPRI), Kyoto University (KU), Taiwan Forestry Research Institute (TFRI), Taiwan Forestry Bureau (TFB), and National Chung Hsing University (NCHU) for providing study sites and experimental equipment. We are very grateful to professor A. Osawa & N. Osawa, associate professor N. Okada, assistant professor M. Ataka, and Ms. N. Shiozaki at KU for environmental support. We thank retired professor B. H. Sheu and Dr. S. M. Hsu at NCHU, retired senior research fellow Dr. T. H. Chen at TFRI, and Dr. C. S. Hsu at Taiwan Research Institute (TRI) for contributing initial ideas, and professor C. C. Chang & S. H. Hsu at National Taiwan University (NTU) and Academia Sinica (AS) for conceptualising climate policy assessment. Further, we also thank Mr. M. Ohnuki, Dr. H. Schäfer, Dr. M. Kamar, Dr. R. Nakahata, Dr. A. Niazai, Ms. Y. Kihara, Dr. T. Tanabe, Mr. A. Yabuki, Mr. R. Fujii, Mr. Aye Ko Ko, Ms. Y. Xia, and Mr. T. Ikeda at KU, assistant professor B. A. T. M. Zinnatul at Begum Rokeya University, assistant

professor M. Mukai at University of Yamanashi, and professor D. Epron at University of Lorraine & KU for field works, providing post-script ideas, and rolling the corrections of the KS approach. We further thank Mr. Z. Y. Li at TFRI and retired professor C. H. Yen at NCHU for performing the soil experiments, Dr. W. Nakai at DATUM STUDIO Co. Ltd. for constructing the cluster and suggesting the method of validation, deputy director-general M. F. Chou and climate mitigation & adaptation specialist C. Y. Chang at Council of Agriculture, Executive Yuan, Taiwan (R.O.C.) for providing feedback on the user response from policymakers.

References

- Adams, J. M., Piovesan, G., 2002. Uncertainties in the role of land vegetation in the carbon cycle. *Chemosphere* 49, 805-819. [https://doi.org/10.1016/S0045-6535\(02\)00382-X](https://doi.org/10.1016/S0045-6535(02)00382-X)
- Belkin, M., Hsu, D., Ma, S., Mandal, S., 2019. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *PNAS* 116, 15849–15854. <https://doi.org/10.1073/pnas.1903070116>.
- Belkin, M., Hsu, D., Ma, S., Mandal, S., 2020. Reply to Loog et al.: Looking beyond the peaking phenomenon. *PNAS* 117, 10627–10627. <https://doi.org/10.1073/pnas.2003206117>.
- Bou Kheir, R., Greve, M. H., Bøcher, P. K., Greve, M. B., Larsen, R., McCloy, K., 2010. Predictive mapping of soil organic carbon in wet cultivated lands using classification-tree based models: The case study of Denmark. *J. Environ. Manage.* 91, 1150–1160. <https://doi.org/10.1016/j.jenvman.2010.01.001>.
- Breiman, L., 2001. Random Forests. *Mach. Learn.* 45, 5–32. <https://doi.org/10.1023/A:1010933404324>.
- Brown, B., Lugo, A. E., 1984. Biomass of tropical forests: A new estimate based on

566 forest volumes. *Science* 223, 1290–1293.
 567 <https://doi.org/10.1126/science.223.4642.1290>.
 568 Brown, S., 2002. Measuring carbon in forests: current status and future challenges.
 569 *Environ. Pollut.* 116, 363–372. [https://doi.org/10.1016/S0269-7491\(01\)00212-3](https://doi.org/10.1016/S0269-7491(01)00212-3).
 570 Buckingham, K., Jepson, P., Wu, L. R., Ramanuja Rao, I. V., Jiang, S., Liese, W., Lou,
 571 Y. P., Fu, M. Y., 2011. The potential of bamboo is constrained by outmoded policy
 572 frames. *Ambio*. 40, 544–548. <https://doi.org/10.1007/s13280-011-0138-4>.
 573 van Buuren, S., Groothuis-Oudshoorn, C. G. M., 2011. mice: Multivariate Imputation
 574 by Chained Equations in R. *J. Stat. Softw.* 45(3), 1–67.
 575 <https://doi.org/10.18637/jss.v045.i03>.
 576 Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. In *Proc. 22nd*
 577 *ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* 785–794
 578 (Association for Computing Machinery, 2016).
 579 <https://doi.org/10.1145/2939672.2939785>.
 580 Chen, T. H., Wang, D. H., Wang, S., 2016. The trend of growth characteristics of
 581 Moso bamboo (*Phyllostachys pubescens*) forests under an unmanaged condition in
 582 central Taiwan. *Taiwan J. For. Sci.* 31, 75–87.
 583 [https://kmweb.coa.gov.tw/files/document/365351/3e1488ef4c51aa766c3e8389067](https://kmweb.coa.gov.tw/files/document/365351/3e1488ef4c51aa766c3e83890677cbd3_v2.pdf)
 584 [7cbd3_v2.pdf](https://kmweb.coa.gov.tw/files/document/365351/3e1488ef4c51aa766c3e83890677cbd3_v2.pdf) (accessed 10 October 2021)
 585 Cook-Patton, S. C., Leavitt, S. M., Gibbs, D., Harris, N. L., Lister, K., Anderson-
 586 Teixeira, K. J., Briggs, R. D., Chazdon, R. L., Crowther, T. W., Ellis, P. W.,
 587 Griscom, H. P., Herrmann, V., Holl, k. D., Houghton, R. A., Larrosa, C., Lomax,
 588 G., Lucas, R., Madsen, P., Malhi, Y., Paquette, A., Parker, J. D., Paul, K., Routh,
 589 D., Roxburgh, S., Saatchi, S., van den Hoogen, J., Walker, W. S., Wheeler, C. E.,
 590 Wood, S. A., Xu, L., Griscom, B. W., 2020. Mapping carbon accumulation
 591 potential from global natural forest regrowth. *Nature* 585, 545–550.

<https://doi.org/10.1038/s41586-020-2686-x>.
 Cortes, C., Vapnik, V., 1995. Support-vector networks. *Mach. Learn.* 20, 273–297.
<https://doi.org/10.1007/BF00994018>.
 Fukushima, K., Usui, N., Ogawa, R., Tokuchi, N., 2015. Impacts of Moso bamboo (*Phyllostachys pubescens*) invasion on dry matter and carbon and nitrogen stocks in a broad-leaved secondary forest located in Kyoto, western Japan. *Plant Species Biol.* 30, 81–95. <https://doi.org/10.1111/1442-1984.12066>.
 Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE/ACM Trans. Audio, Speech, Language Process.* PAMI-6, 721–741. <https://doi.org/10.1109/TPAMI.1984.4767596>.
 Guo, Z. D., Fang, J. Y., Pan, Y. D., Birdsey, R., 2010. Inventory-based estimates of forest biomass carbon stocks in China: A comparison of three methods. *For. Ecol. Manage.* 259, 1225–1231. <https://doi.org/10.1016/j.foreco.2009.09.047>.
 Isagi, Y., Kawahara, T., Kamo, K., Ito, H., 1997. Net production and carbon cycling in a bamboo *Phyllostachys pubescens* stand. *Plant Ecol.* 130, 41–52. <https://doi.org/10.1023/A:1009711814070>.
 Intergovernmental Panel on Climate Change, 2006. In 2006 IPCC Guidelines for National Greenhouse Gas Inventories. Volume 4: Agriculture, Forestry and Other Land Use. Institute for Global Environmental Strategies, Hayama, Kanagawa, Japan. <https://www.ipcc-nggip.iges.or.jp/public/2006gl/vol4.html> (accessed 10 October 2021).
 Intergovernmental Panel on Climate Change, 2019. 2019 Refinement to the 2006 IPCC Guidelines for National Greenhouse Gas Inventories. Volume 4: Agriculture, Forestry and Other Land Use. Intergovernmental Panel on Climate Change, Geneva, Switzerland. <https://www.ipcc-nggip.iges.or.jp/public/2019rf/index.html> (accessed 10 October 2021).

- 618 Katumbi, N. M., Kinyanjui, M. J., Jm, K., Mware, M. J., 2017. Biomass energy
619 resource of the highland bamboo (*Yushania alpina*) and its potential for sustainable
620 exploitation in southern aberdares Forest. J. Sustain. Bioenergy Syst. 7, 85–97.
621 <https://doi.org/10.4236/jsbs.2017.73007>.
- 622 Kendall, M. G., 1938. A new measure of rank correlation. Biometrika 30(1-2), 81–93.
623 <https://doi.org/10.1093/biomet/30.1-2.81>.
- 624 Kobayashi, K., Kitayama, K., Onoda, Y., 2018. A simple method to estimate the rate
625 of the bamboo expansion based on one-time measurement of spatial distribution of
626 culms. Ecol. Res. 33, 1137–1143. <https://doi.org/10.1007/s11284-018-1626-9>.
- 627 Kobayashi, K., Ohashi, M., Fujihara, M., Kitayama, K., Onoda, Y., 2022. Rhizomes
628 play significant roles in biomass accumulation, production and carbon turnover in
629 a stand of the tall bamboo *Phyllostachys edulis*, J. For. Res. (accepted)
630 <https://doi.org/10.1080/13416979.2022.2090669>.
- 631 Lee, H. Y., Chung, P. H., Wu, Y. C., Lin, T. H., Wen, T. H., 2018. Interactive spoken
632 content retrieval by deep reinforcement learning. IEEE/ACM Trans. Audio,
633 Speech, Language Process 26, 2447–2459.
634 <https://doi.org/10.1109/TASLP.2018.2852739>.
- 635 Lin, M. Y., Hsieh, I. F., Lin, P. H., Laplace, S., Ohashi, M., Chen, T. H., Kume, T.,
636 2017. Moso bamboo (*Phyllostachys pubescens*) forests as a significant carbon
637 sink? A case study based on 4-year measurements in central Taiwan. Ecol. Res. 32,
638 845– 857. <https://doi.org/10.1007/s11284-017-1497-5>.
- 639 Lobovikov, M., Schoene, D., Lou, Y. P., 2012. Bamboo in climate change and rural
640 livelihoods. Mitig. Adapt. Strateg. Glob. Chang. 17, 261–276.
641 <https://doi.org/10.1007/s11027-011-9324-8>.
- 642 Martin, A. P., Palmer, W. M., Byrt, C. S., Furbank, R. T., Grof, C. P., 2013. A holistic
643 high-throughput screening framework for biofuel feedstock assessment that

644 characterises variations in soluble sugars and cell wall composition in *Sorghum*
 645 *bicolor*. *Biotechnol. Biofuels* 6, 186. <https://doi.org/10.1186/1754-6834-6-186>.
 646 Mei, T., Fang, D., Röhl, A., Hölscher, D., 2019. Bamboo water transport assessed with
 647 deuterium tracing. *Forests* 10, 623. <https://doi.org/10.3390/f10080623>.
 648 Moore, P. D., Chapman, S. B., 1986. *Methods in Plant Ecology*. Blackwell Scientific
 649 Publications, London, GB, pp. 215–337.
 650 Nath, A. J., Lal, R., Das, A. K., Managing woody bamboos for carbon farming and
 651 carbon trading. *Glob. Ecol. Conserv.* 3, 654–663.
 652 <https://doi.org/10.1016/j.gecco.2015.03.002>.
 653 Okutomi, K., Shinoda, S., Fukuda, H., 1996. Causal analysis of the invasion of broad-
 654 leaved forest by bamboo in Japan. *J. Veg. Sci.* 7(5), 723–728.
 655 <https://doi.org/10.2307/3236383>.
 656 Pearson, K., 1901. LIII. On lines and planes of closest fit to systems of points in
 657 space. *Philos Mag* 2, 559–572. <https://doi.org/10.1080/14786440109462720>.
 658 Poeplau, C., Don, A., Vesterdal, L., Leifeld, J., Wesemael, B. V., Schumacher, J.,
 659 Gensior, A., 2011. Temporal dynamics of soil organic carbon after land-use change
 660 in the temperate zone – carbon response functions as a model approach. *Glob.*
 661 *Change Biol.* 17, 2415–2427. <https://doi.org/10.1111/j.1365-2486.2011.02408.x>.
 662 Rosenblatt, F., 1958. The perceptron: A probabilistic model for information storage
 663 and organization in the brain. *Psychol. Rev.* 65, 386–408.
 664 <https://doi.org/10.1037/h0042519>.
 665 Royston, P., 1982. Algorithm AS 181: The W test for normality. *J. R. Stat. Soc. Series*
 666 *C* 31, 176–180. <https://doi.org/10.2307/2347986>.
 667 Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., Hemingway, H., 2014.
 668 Comparison of random forest and parametric imputation models for imputing
 669 missing data using MICE: A CALIBER Study. *Am. J. Epidemiol.* 179(6), 764–774.

<https://doi.org/10.1093/aje/kwt312>.
 Shi, J., Mao, S., Wang, L., Ye, X., Wu, J., Wang, G., Chen, F., Yang, Q., 2021. Clonal
 integration driven by source-sink relationships is constrained by rhizome
 branching architecture in a running bamboo species (*Phyllostachys glauca*): A ¹⁵N
 assessment in the field. For. Ecol. Manage. 481, 118754.
<https://doi.org/10.1016/j.foreco.2020.118754>.
 Shinohara, Y., Kume, T., Ichihashi, R., Komatsu, H., Otsuki, K., 2014. Moso-bamboo
 forests in Japan: What are the effects of their area expansion on ecosystem
 services? J. Japanese For. Soc. 96, 351–361. <https://doi.org/10.4005/jjfs.96.351>.
 Silver, D., Huang, A. J., Maddison, C. J., Guez, A., Sifre, L., van der Driessche, G.,
 Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S.,
 Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M.,
 Kavukcuoglu, K., Graepel, T., Hassabis, D., 2016. Mastering the game of Go with
 deep neural networks and tree search. Nature 529, 484–489.
<https://doi.org/10.1038/nature16961>.
 Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A. J., Guez, A.,
 Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y. T., Lillicrap, T., Hui, F., Sifre,
 L., van den Driessche, G., Graepel, T., Hassabis, D., 2017. Mastering the game of
 Go without human knowledge. Nature 550, 354–359.
<https://doi.org/10.1038/nature24270>.
 Song, X. Z., Peng, C. H., Ciais, P., Li, Q., Xiang, W., Xiao, W., Zhou, G. M., Deng,
 L., 2020. Nitrogen addition increased CO₂ uptake more than non-CO₂ greenhouse
 gases emissions in a Moso bamboo forest. Sci. Adv. 6, eaaw5790.
<https://doi.org/10.1126/sciadv.aaw5790>.
 Song, X. Z., Peng, C. H., Zhou, G. M., Gu, H. H., Quan, L., Zhang, C., 2016.
 Dynamic allocation and transfer of non-structural carbohydrates, a possible

696 mechanism for the explosive growth of Moso bamboo (*Phyllostachys heterocycla*).
697 Sci. Rep. 6, 25908. <https://doi.org/10.1038/srep25908>.
698 Spearman, C., 1904. The proof and measurement of association between two things.
699 Am. J. Psychol. 15, 72–101. <https://doi.org/10.2307/1412159>.
700 Stallkamp, J., Schlipsing, M., Salmen, J., Igel, C., 2012. Man vs. computer:
701 Benchmarking machine learning algorithms for traffic sign recognition. Neural
702 Netw. 32, 323–332. <https://doi.org/10.1016/j.neunet.2012.02.016>.
703 Thompson, J. R., Foster, D. R., Scheller, R., Kittredge, D., 2011 The influence of land
704 use and climate change on forest biomass and composition in Massachusetts, USA.
705 Ecol. Appl. 21, 2425–2444. <https://doi.org/10.1890/10-2383.1>.
706 Toker, D., Sommer, F. T., D’Esposito, M., 2020. A simple method for detecting chaos
707 in nature. Commun. Biol. 3, 11. <https://doi.org/10.1038/s42003-019-0715-9>.
708 Wang, B., Waters, C., Orgill, S., Cowie, A., Clark, A., Liu, D. L., Simpson, M.,
709 McGowen, I., Sides, T., 2018. Estimating soil organic carbon stocks using different
710 modelling techniques in the semi-arid rangelands of eastern Australia. Ecol. Indic.
711 88, 425–438. <https://doi.org/10.1016/j.ecolind.2018.01.049>.
712 Wang, B., Wei, W. J., Liu, C. J., You, W. Z., Niu, X., Man, R. Z., 2013. Biomass and
713 carbon stock in Moso bamboo forests in subtropical China: characteristics and
714 implications. J. Trop. For. Sci. 25, 137–148. <http://www.jstor.org/stable/43595383>.
715 Wang, S., Chen, T. H., Chang, H. C., Chung, H. Y., Li, T. I., Liu, C. P., 2009. The
716 structures, aboveground biomass carbon storage of *Phyllostachys pubescens* stands
717 in Huisun Experimental Forest Station and Shi-Zhuo. Quarterly J. For. Res. 34,
718 17– 26. <https://doi.org/10.29898/SHBQ.200912.0002>.
719 Wang, S., Chen, T. H., Chen, S. Y., Chung, H. Y., Liu, E. U., Li, T. I., Liu, C. P., 2010.
720 Estimating aboveground biomass and carbon sequestration of Moso bamboo
721 grown under selection cutting after two years. Quarterly J. For. Res. 32, 35–44.

- 722 <https://doi.org/10.29898/SHBQ.201009.0003>.
- 723 Wang, S., Chen, T. H., Liu, E. U., Liu, C. P., 2020. Accessing the nursing behaviour of
- 724 Moso bamboo (*Phyllostachys edulis*) on carbohydrates dynamics and
- 725 photosystems. Sci. Rep. 10, 1015. <https://doi.org/10.1038/s41598-020-57643-1>.
- 726 Wang, S., Kobayashi, K., Li, T. I., Hsu, C. S., Chen, T. H., Liu, C. P., Dannoura, M.,
- 727 2019. Assessment of forest carbon stock based on management and environmental
- 728 factors using the Multivariate Imputation by Chained Equations (MICE) and
- 729 Multiple Regression (MR) approaches – a case study on Moso bamboo
- 730 (*Phyllostachys edulis*). The esj Kinki Region Conference.
- 731 [https://www.researchgate.net/publication/365489594_Assessment_of_forest_carbon](https://www.researchgate.net/publication/365489594_Assessment_of_forest_carbon_stock_based_on_management_and_environmental_factors_using_the_Multivariate_Imputation_By_Chained_Equations_MICE_and_Multiple_Regression_MR_approaches_-_a_case_study_on_Moso)
- 732 [n_stock_based_on_management_and_environmental_factors_using_the_Multivari](https://www.researchgate.net/publication/365489594_Assessment_of_forest_carbon_stock_based_on_management_and_environmental_factors_using_the_Multivariate_Imputation_By_Chained_Equations_MICE_and_Multiple_Regression_MR_approaches_-_a_case_study_on_Moso)
- 733 [ate_Imputation_By_Chained_Equations_MICE_and_Multiple_Regression_MR_ap](https://www.researchgate.net/publication/365489594_Assessment_of_forest_carbon_stock_based_on_management_and_environmental_factors_using_the_Multivariate_Imputation_By_Chained_Equations_MICE_and_Multiple_Regression_MR_approaches_-_a_case_study_on_Moso)
- 734 [proaches_-_a_case_study_on_Moso](https://www.researchgate.net/publication/365489594_Assessment_of_forest_carbon_stock_based_on_management_and_environmental_factors_using_the_Multivariate_Imputation_By_Chained_Equations_MICE_and_Multiple_Regression_MR_approaches_-_a_case_study_on_Moso).
- 735 Were, K., Bui, D. T., Dick, Ø. B., Singh, B. R., 2015. A comparative assessment of
- 736 support vector regression, artificial neural networks, and random forests for
- 737 predicting and mapping soil organic carbon stocks across an Afrotropical
- 738 landscape. Ecol. Indic. 52, 394–403. <https://doi.org/10.1016/j.ecolind.2014.12.028>.
- 739 Xiao, F. M., Fan, S. H., Wang, S. L., Guan, F. Y., Yu, X. J., Shen, Z. Q., 2010.
- 740 Estimation of the carbon balance in Moso bamboo and Chinese fir plantation
- 741 ecosystem. Scientia Silvae Sinicae 46, 59–65. [https://doi.org/10.11707/j.1001-](https://doi.org/10.11707/j.1001-7488.20101108)
- 742 [7488.20101108](https://doi.org/10.11707/j.1001-7488.20101108).
- 743 Xiao, F. M., Fan, S. H., Wang, S. L., Xiong, C. Y., Shen, Z. Q., 2009. Soil carbon
- 744 cycle of *Phyllostachys edulis* plantation in Huitong region, Hunan Province.
- 745 Scientia Silvae Sinicae 45, 11–15. <https://doi.org/10.11707/j.1001-7488.20090603>.
- 746 Yen, T. M., Ji, Y. J., Lee, J. S., 2010. Estimating biomass production and carbon
- 747 storage for a fast-growing makino bamboo (*Phyllostachys makinoi*) plant based on

the diameter distribution model. *For. Ecol. Manage.* 260, 339–344.
<https://doi.org/10.1016/j.foreco.2010.04.021>.

Yen, T. M., Lee, J. S., 2011. Comparing aboveground carbon sequestration between Moso bamboo (*Phyllostachys heterocycla*) and China fir (*Cunninghamia lanceolata*) forests based on the allometric model. *For. Ecol. Manage.* 261, 995–1002. <https://doi.org/10.1016/j.foreco.2010.12.015>.

Youkhana, A. H., Ogoshi, R. M., Kiniry, J. R., Meki, M. N., Nakahata, M. H., Crow, S. E., 2017. Allometric models for predicting aboveground biomass and carbon stock of tropical perennial C₄ grasses in Hawaii. *Front. Plant Sci.* 8, 650. <https://doi.org/10.3389/fpls.2017.00650>.

Yuen, J. Q., Fung, T., Ziegler, A. D., 2017. Carbon stocks in bamboo ecosystems worldwide: Estimates and uncertainties. *For. Ecol. Manage.* 393, 113–138. <https://doi.org/10.1016/j.foreco.2017.01.017>.

Zhang, H., Guan, D. S., Song, M. W., 2012. Biomass and carbon storage of Eucalyptus and Acacia plantations in the Pearl River Delta, South China. *For. Ecol. Manage.* 277, 90–97. <https://doi.org/10.1016/j.foreco.2012.04.016>.

Zhao, J. L., Kang, F. F., Wang, L. X., Yu, X. W., Zhao, W. H., Song, X. S., Zhang, Y. Z., Chen, F., Sun, Y., He, T. F., Han, H. R., 2014. Patterns of biomass and carbon distribution across a chronosequence of Chinese pine (*Pinus tabulaeformis*) forests. *PLoS One* 9, e94966. <https://doi.org/10.1371/journal.pone.0094966>.

Funding: This work was supported by JSPS KAKENHI [grant numbers JP 15H04513, JP 19J11336, and JP 20J15519], and by the Taiwan Forestry Research Institute, Council of Agriculture, Executive Yuan, Taiwan (R.O.C.) [grant number 99AS-8.3.2-F-G1].

Video

1 774 The briefing video was revealed at <https://www.youtube.com/watch?v=l8SKe-K6ypY>

2
3 775

4
5 776 **Data and materials availability**

6
7
8 777 All data needed to evaluate the conclusions in the paper are present in the paper and/or

9
10 778 the Supplementary Materials. We revealed R codes of the WTF sub-algorithm for

11
12 779 estimating irregular target variables at https://github.com/gn03138868/WTF_sub-

13
14
15 780 [algorithm](https://github.com/gn03138868/WTF_sub-)

16
17 781

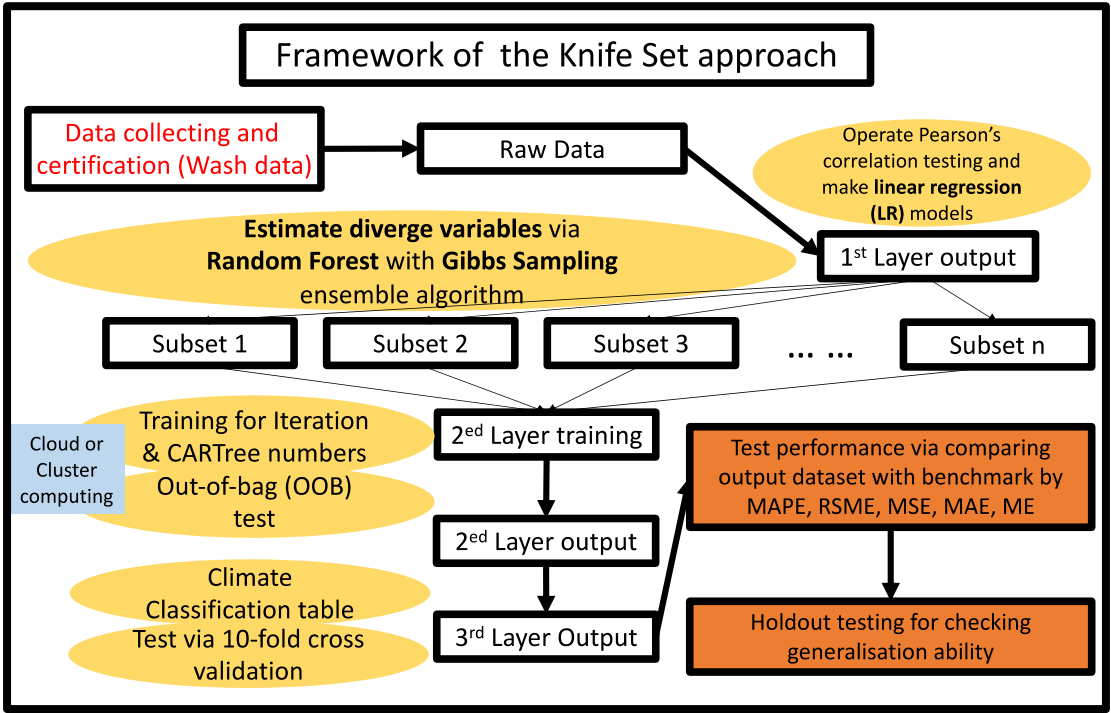


Fig. 1. The framework of knife set (KS) approach for estimating CSSs in forests.

The MAPE means mean absolute percentage error. The RMSE means root mean square error. The MSE means mean square error. The MAE means mean absolute error. The ME means mean error. The procedure is detailly described in Method S1. Pseudocode is described in Table S1.

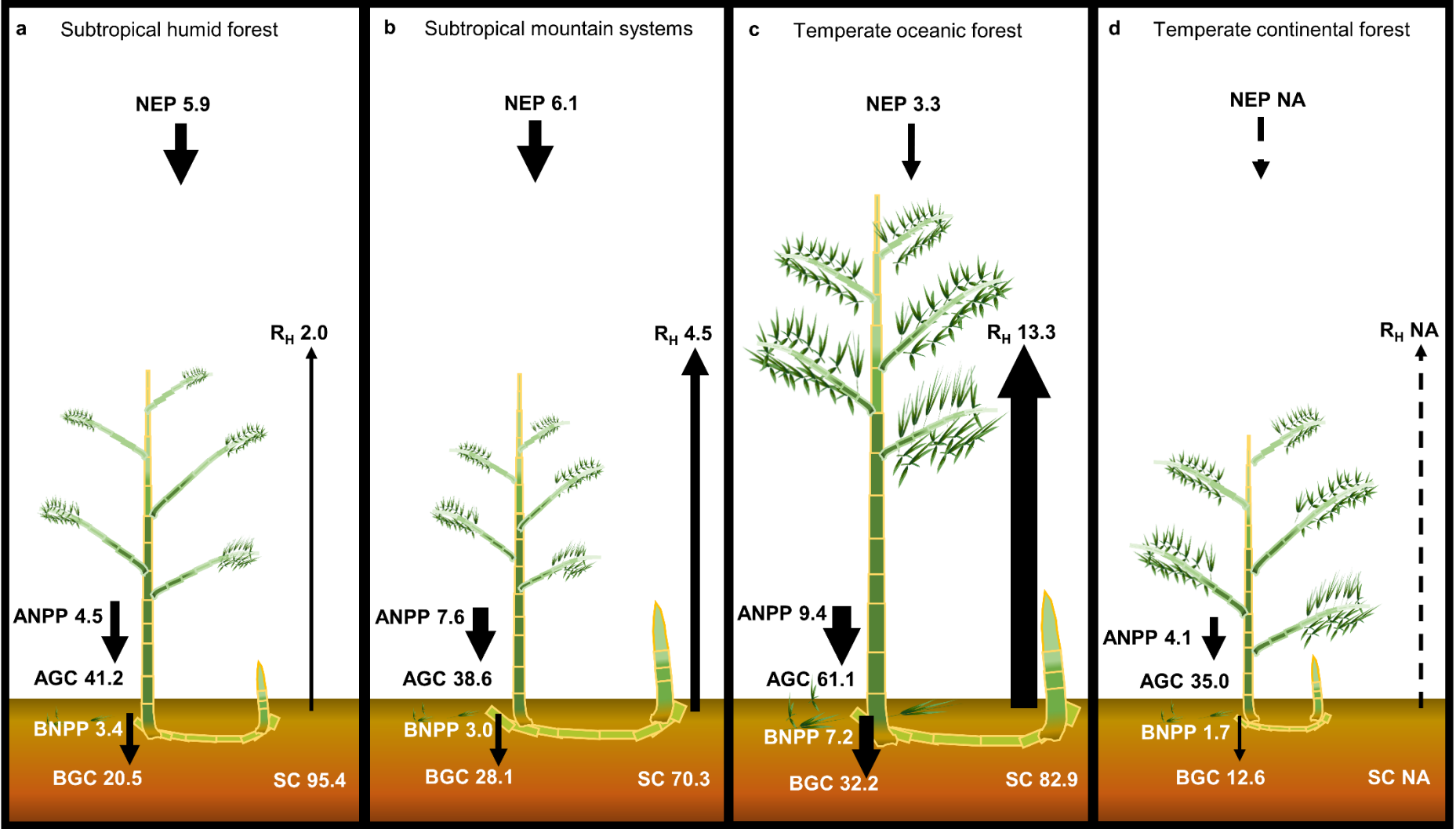


Fig. 2. Annual carbon dynamic for each ecological zone of *P. edulis* forests. All units are Mg C ha^{-1} for AGC, BGC, and SC storages, and are $\text{Mg C ha}^{-1} \text{ year}^{-1}$ for ANPP, BNPP, heterotopic respiration (R_H), and net ecosystem productivity (NEP). (a) Subtropical humid forest. (b) Subtropical mountain systems. (c) Temperate oceanic forest. (d) Temperate continental forest. The figure area of each organ is adjusted and scaled according to the real biomass ratio of each climatic zone, and the width of the arrows is adjusted in proportion to the actual carbon sinks and sources. Values are obtained from Dataset S1.

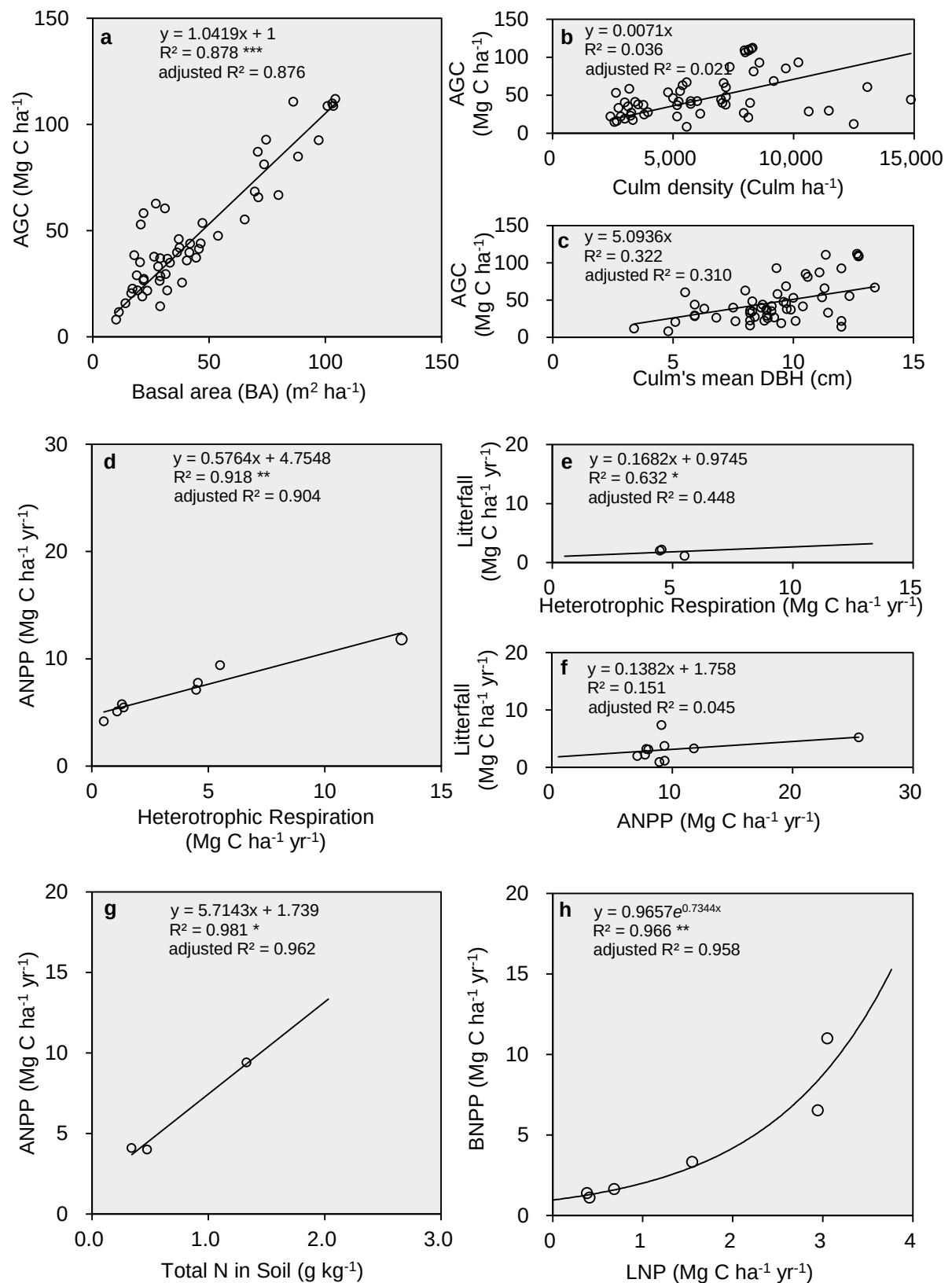


Fig. 3. Significant correlation between target and IVs used in layer 1. (a) Correlation between AGC storage and BA. (b) Correlation between AGC and culm density. (c) Correlation between AGC and culms' mean DBH. (d) Correlation between ANPP and heterotrophic respiration. (e) Correlation between litterfall and heterotrophic respiration. (f) Correlation between litterfall and ANPP. (g) Correlation between total nitrogen in soil and ANPP. (h) Correlation between leaf net production (LNP) and BNPP. Label “*” means $p < 0.05$, “**” means $p < 0.01$, and “***” means $p < 0.001$. All observations are obtained from Dataset S1.

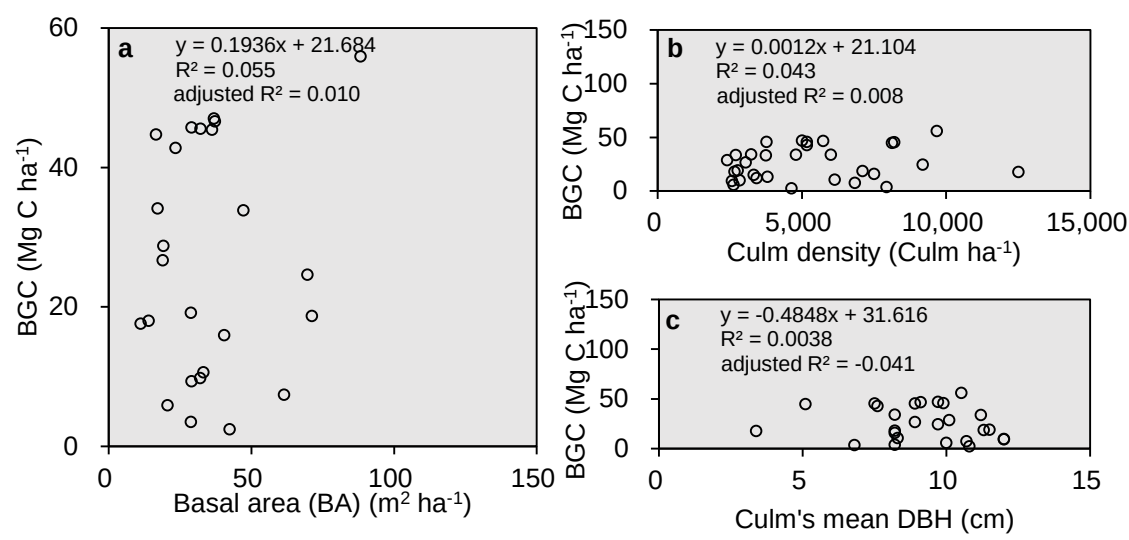


Fig. 4. Correlation between BGC stocks and other IVs. (a) Correlation between BGC and BA. (b) Correlation between BGC and culm density. (c) Correlation between BGC and culms' mean DBH. Open circles represent observations obtained from Dataset S1.

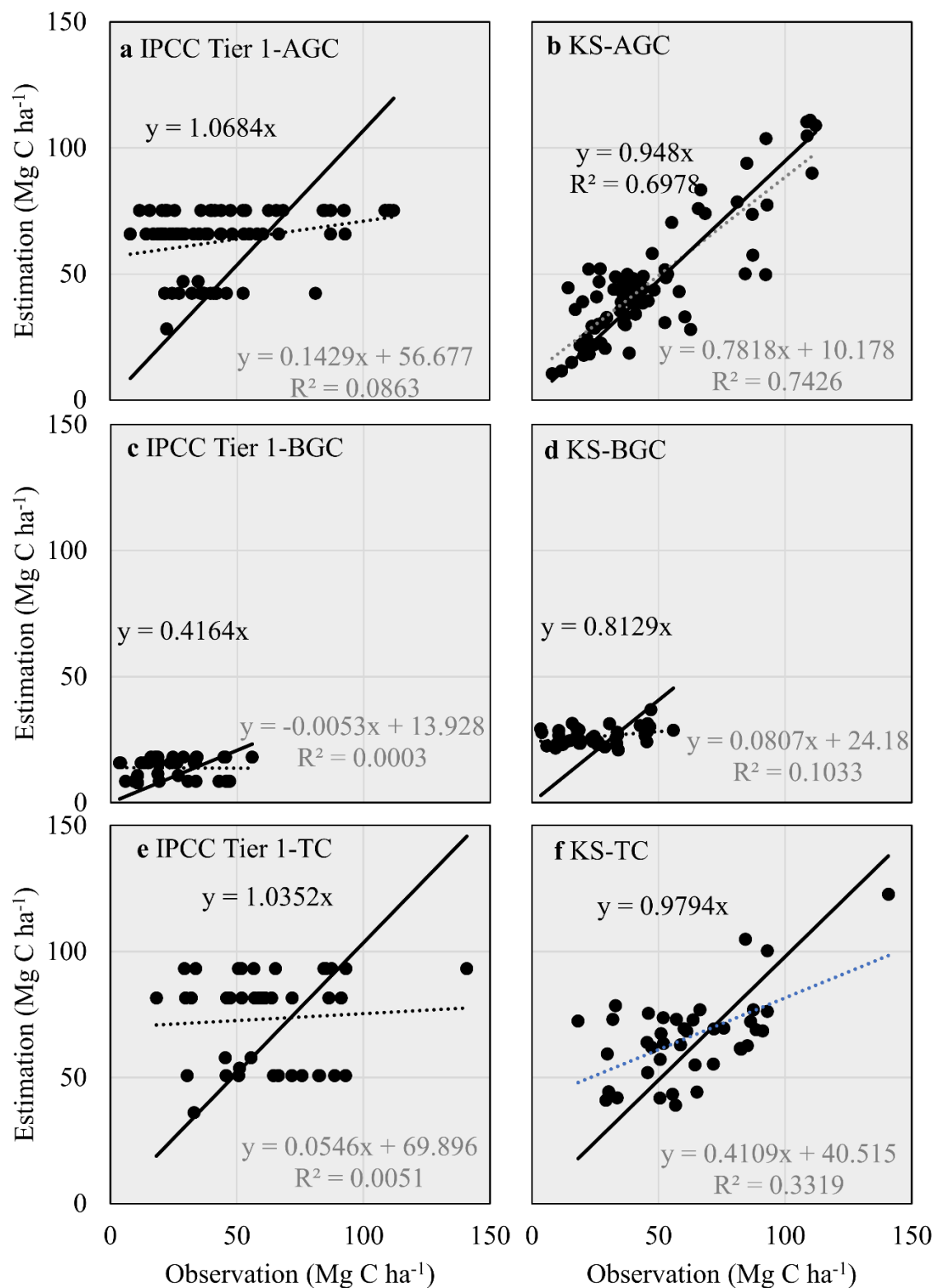


Fig. 5. Correlation between observation and estimation in forest carbon stocks of *P. edulis*. (a) Correlation between observation and estimation of AGC (n=77) via IPCC's Tier 1 method. (b) Correlation between observation and estimation of AGC via KS approach. (c) Correlation between observation and estimation of BGC (n=39) via IPCC's Tier 1 method. (d) Correlation between observation and estimation of BGC via KS approach. (e) Correlation between observation and estimation of total carbon (TC) (n=40) via IPCC's Tier 1 method. (f) Correlation between observation and estimation of TC via KS approach. The slope in black font represents the estimated accuracy (intercept is zero). The R^2 value in grey and dashed line font represent the estimated precision.

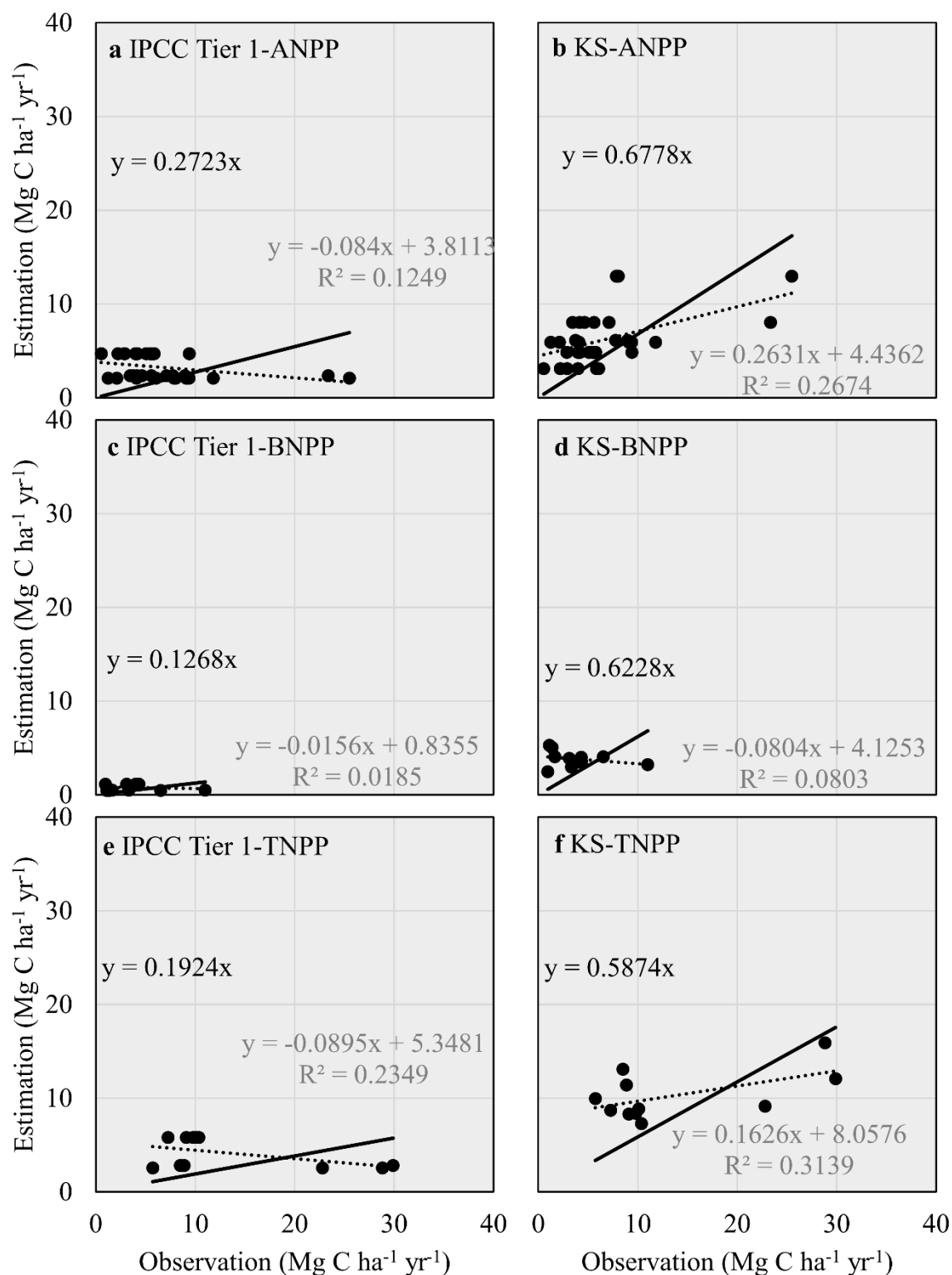


Fig. 6. Correlation between observation and estimation in forest carbon sinks of *P. edulis*. (a) Correlation between observation and estimation of ANPP (n=33) via IPCC's Tier 1 method. (b) Correlation between observation and estimation of ANPP via KS approach. (c) Correlation between observation and estimation of BNPP (n=11) via IPCC's Tier 1 method. (d) Correlation between observation and estimation of BNPP via KS approach. (e) Correlation between observation and estimation of TNPP (n=11) via IPCC's Tier 1 method. (f) Correlation between observation and estimation of TNPP via KS approach. The slope in black font represents the estimated accuracy (intercept is zero). The R^2 value in grey and dashed line font represent the estimated precision.

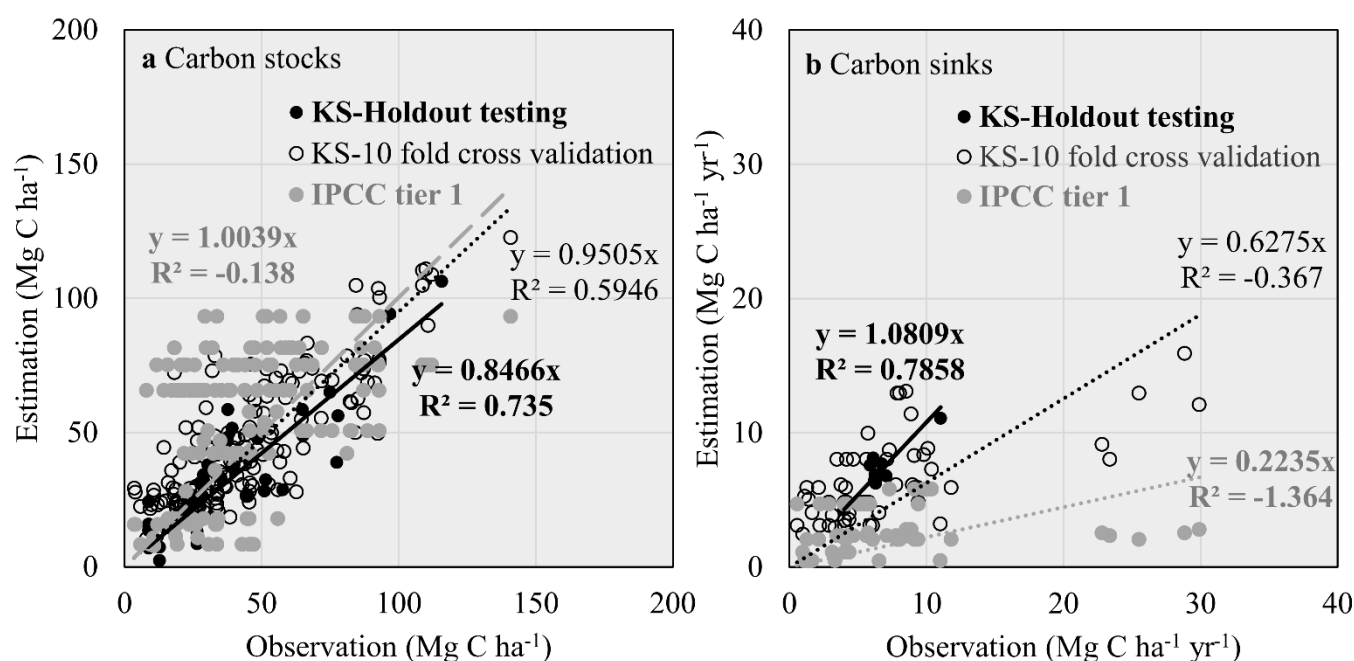


Fig. 7. Correlation between observation and estimation in forest CSSs of *P. edulis* in holdout testing, k-fold validation, and benchmark (IPCC tier 1 method). (a) Correlation between observation and estimation of carbon stocks in holdout testing (n=38), 10-fold validation (n=156), and benchmark (IPCC tier 1 method) (n=156). (b) Correlation between observation and estimation of carbon stocks in holdout testing (n=9), 10-fold validation (n=55), and benchmark (IPCC tier 1 method) (n=55). The slope represents the estimated accuracy (intercept is zero). The R^2 value represents the estimated precision. The bold black font represents the holdout testing of the KS approach. The black font represents the 10-fold cross validation of the KS approach. The grey font represents the benchmark (IPCC tier 1 method).

Table 1. The category of factors used in knife set approach in this study

category of factors	Variables
Time (2 factors)	beginning year, finished year.
Management (6 factors)	managed, fertilized, weeded & selective cut, shoot dug, clear cutting, and mixed with other forests.
Environment (36 factors)	Mean annual temperature, warmth index, annual rainfall, snow, elevation, relative humidity, sunshine duration, wind speed, water content (soil), pH (soil-KCl), pH (soil-H ₂ O), total N (soil), total P (soil), total K (soil), available P ₂ O ₅ (soil), available SiO ₂ (soil), cation exchange capacity, CEC (soil), K ⁺ (soil), Ca ²⁺ (soil), Mg ²⁺ (soil), Sand (soil), Silt (soil), Clay (soil), N (litter), Ca (litter), K (litter), Mg (litter), P (litter), Si (storage in aboveground biomass), Si (storage in belowground biomass), Si (storage in soil), Si (primary sink in plant annually), Si (net sink in plant annually), Si (return to soil), relative luminosity.
Growth trait (43 factors)	culm density, mean DBH, mean height, BA, LAI (Fisheye lens), LAI (leaf area scanner), leaves C, branches C, culms C, fine roots C, coarse root C, rhizomes C, stump C, soil C (0-10 cm), soil C (10-30 cm), foliage, branches, culms, AGC, root_shoot ratio, roots, rhizomes, stumps, BGC, TC, litter layer, SC, Undergrowth, TEC, LNP, BNP, CNP, litterfall, ANPP, RoNP, RhNP, SNP, BNPP, TNPP, soil respiration (SR), heterotrophic respiration (HR), and net ecosystem production (NEP).

Supplementary materials

Estimating divergent forest carbon stocks and sinks via a knife set approach

Shitephen Wang^{1,*}, Keito Kobayashi^{1,2}, Satoru Takanashi², Chiung-Pin Liu³,

Dian-Rong Li^{4,†}, San-Wan Chen^{5,†}, Yu-Ting Cheng⁶,

Kai Moriguchi⁷, Masako Dannoura¹

¹Graduate School of Agriculture, Kyoto University, Kyoto, 606-8502, Japan

²Kansai Research Centre, Forestry and Forest Products Research Institute, Kyoto, 612- 0855, Japan

³Department of Forestry, College of Agriculture and Nature Resource, National Chung Hsing University,
Taichung, 402-204, Taiwan

⁴Department of Electrical Engineering, National Taiwan Normal University, Taipei, 106-308, Taiwan

⁵Department of Computer Science & Information Engineering, National Taiwan University, Taipei, 106-216,
Taiwan

⁶Greater New York City Area, Médecins Sans Frontières (MSF), New York, 10006, the USA

⁷Faculty of Agriculture and Marine Sciences, Kochi University, Kochi, 783-8502, Japan

*Email: gn03138868@gmail.com

†These authors contributed equally to this work

These supplementary materials are consists of the methods, figures, and tables in the main text.

Contents of supplementary materials

Dataset S1. Raw data matrix of target and IVs collected from <i>P. edulis</i>.....	s3
Method S1. The practical guideline of the KS approach in R programming language via cluster or cloud computing.....	s3
Fig. S1. Location of the study sites in Huisun Experimental Forest Station and in Forestry and Forest Products Research Institute (FFPRI), Kansai Research Centre.	s16
Fig. S2. Frameworks of raw data for further analysis of the KS approach.	s17
Fig. S3. Correlation between dig rate and mean absolute different ratio of each dug value (%) in the original datasheet estimated via random forest with Gibbs sampling.....	s17
Fig. S4. LR models for estimating AGC and BGC stocks.	s17
Fig. S5. Relationships between $\lambda(M)$ and the different combinations of duplications, max iteration, and nTrees numbers in one of k folds.....	s18
Fig. S6. Meaning of indices for the performance of algorithms.	s18
Fig. S7. Flowchart of statistical process of different methods.	s19
Fig. S8. Frameworks of holdout dataset for testing the generalisation ability of the KS approach.....	s20
Table S1. Pseudocode of the WTF sub-algorithm in the KS approach.	s21
Table S2. Common terms and definitions of variables associated with photosynthesis, respiration, and carbon cycles in this these studied <i>P. edulis</i> forests.....	s22
Table S3. Results of Pearson's correlation test between dependent and IVs.....	s26
Table S4. Ratio of below-ground biomass to above-ground biomass (R) in <i>P. edulis</i> forests.....	s28
Table S5. Tier 1 estimated biomass values of <i>P. edulis</i> referred to the method from table 4.7-4.10 (values are approximate; use only for Tier 1) in 2006 IPCC Guidelines for National Greenhouse Gas Inventories.	s29
Table S6. Performance of k-fold cross-validation between benchmark (IPCC Tier 1) and KS approach.	s30

Dataset S1. Raw data matrix of target and IVs collected from *P. edulis*.

(<https://figshare.com/s/88ae9f2f8f968b706c2b>)

Method S1. The practical guideline of the KS approach in R programming language via cluster or cloud computing.

1. Introduction

Target variables in the forest and natural environment are easily diverging and irregular. Furthermore, many studies have their own research aims, so the existing studies in the world often have different IVs and the data sheets have characteristics that contain quite a lot of missing values. For conquering the problems mentioned above, we design a KS approach that contains multitaskable functions and can be operated easily. The flowchart of the KS is described in Fig. 1 of the main text of “Estimating divergent forest carbon stocks and sinks via a knife set approach”. Below we elaborate a very simple and low-cost method to practice the process of the algorithm that combines simple LR, ML (Gibbs sampling combined random forests), and CCTs to ensure that policymakers of any age in any country can fully understand and operate this algorithm by themselves easily.

2. Pre-treatment and Environment setting

2.1. Raw data collection

Raw data was collected from google scholar, Nature Publishing Group, Science (American Association for the Advancement of Science, AAAS), ScienceDirect, and so on for English peer-reviewed journal articles. However, not all critical and reliable data are published in English. Therefore, according to different species that need to be estimated, look for peer-reviewed articles in different languages and countries, and it is better to have local native speakers in the research team to correctly understand and collect data. For instance, *P. edulis* (Moso bamboo) forests are mainly distributed in China, Japan, South Korea and Taiwan, so our research team has native speakers of simplified Chinese, Japanese, Korean, and traditional Chinese. Further, we collected local peer-reviewed research articles from CNKI (<https://h5.cnki.net/#/home>), J-stage (<https://www.jstage.jst.go.jp/browse/-char/ja>), KISS (<http://kiss.kstudy.com/search/sch-result.asp>), and airtiti Library (<https://www.airitilibrary.com/Search/alJnlbrowse>).

After collecting peer-reviewed articles, carefully read articles, input data in an excel file, save it as .csv for further analysis, and double-check for prevent mistyping. The raw data used in this study is as Dataset S1, and the framework of template shown as Fig. S2. At meanwhile of reading journal researches and making raw data sheet, record or define all variables in a table. (For instance, in the case of this research, common terms of variables defined as Table S2.)

2.2. Pre-treatment of k -fold cross validation

k -fold cross validation is a common method used to evaluate ML models on a limited raw datasheet. The method is to randomly split the data into k sets, and then regard one set as validation dataset, and the rest of $k-1$ sets as training dataset, and repeat this process until each set is regarded as validation dataset once. And the prediction results are compared with the original raw data for performance comparison. Therefore, before the whole process starts, the value of k must be determined even though there is no hard rule for this value.

Further, in previous studies, k was usually pre-set to 3 or 10 (divided into 3 or 10 validation dataset).

Under normal circumstances, studies rarely test the stability of a data matrix. However, the simple test of the stability of the data matrix is carried out to determine the value of k because there are only 81 observations in this study (81 observations for training and k -fold cross validation, and 24 observations for holdout testing.).

2.2.1. Stability test in practice

For determining the k , we designed a dig testing which is randomly digging original data sheet in different percentage (such as 2%, 5%, 10%, 20%, and so on), then calculate the mean absolute different ratio of each dug value (MADREDV) as Eq (S1), and draw the correlation between dig rate and mean absolute different ratio of each dug value (MADREDV) (%).

$$\text{MADREDV} = \frac{1}{n} \sum_{i=1}^n \left| \frac{dnrvPi - Oi}{Oi} \right| \quad \text{Eq(S1)}$$

where $dnrvPi$ is the dug and re-estimated value; Oi is the original dug values collected from peer-reviewed articles; and n is the number of dug observations.

The results of stability test in this study show that if the dig rate over 10%, the MDREDV will sharply increase over 50% (Fig. S3). Hence, we decided $k=10$ in this study and let the MDREDV will below 50%. (Of course, it is possible directly skip the above test, let $k=10$ or 3, depending on the number of observations. If observation size is large, $k=3$ will be better because of saving computing power.)

2.2.2. Separate data sheet into k -folds

Through the above test, it is possible to decide a suitable k (For instance, in this study, we decided $k=10$.) which means the data sheet has to translate qualitative data into dummy variables and operate out of bags for validation (It is a kind of permutation analysis method). Then randomly re-arrange orders of observations in raw datasheet, and evenly divided observations of it into 10 .csv files (10% observations per file) as validation dataset. Grab the rest of each validation dataset as training dataset (90% observations per file) prepared for further analysis.

2.3. Environment setting

Because most of the computer operating systems used are Micro Soft (MS) Windows, the environment settings in this process will all take MS Windows as an example and operate via R statistics and R studio.

2.3.1. Install R in cluster computing (Windows)

Download R from <https://cran.r-project.org/mirrors.html> and R studio from <https://rstudio.com/products/rstudio/download/>, then install them in every single computer in the cluster used to run R codes and operate the following algorithm.

2.3.2. Install R in cloud computing (Amazon Web Services, AWS) (optional)

If research fund is enough, it is also possible using cloud computing to speed up all calculations. Below we introduce how to operate Amazon Web Services (AWS) for cloud computing.

- 2.3.2.1. Visit AWS web site <https://aws.amazon.com/> , click Amazon EC2, apply a new account, and then log in.
- 2.3.2.2. Press “Launch instance” (Then you will see whole of the instance types)
- 2.3.2.3. Go to “Community AMIs” (AMIs means Amazon Machine Images).
- 2.3.2.4. Press “Ubuntu” and search “rstudio”
- 2.3.2.5. Choose “RStudio-1.3.1073_R-4.0.2_CUDA-10.1_cuDNN-7.6.5_ubuntu-18.04-LTS-64bit - ami-00b6908a3a1faec83 “
- 2.3.2.6. Choose an instance type you want. [If you need very strong computing power and have enough funding, it is also possible to choose 128 vCPUs and 3904 (GiB) Memory.]. Then, press “Next: Configure Instance Details”.
- 2.3.2.7. Check the instance details, then press “Next: Add storage”.
- 2.3.2.8. Check the detail of storage. (If your data matrix is really large, it is better to choose suitable storage size. e.g. 100 GiB), press “Next: Add Tags”, and then press “Next: Configure Security Group”.
- 2.3.2.9. Choose SSH for Type. Then, press “Add Rule”, choose “Custom TCP” for Type, set 80 for Port Range. Further, press “Add Rule”, choose “Custom TCP” for Type, set 8787 for Port Range. (Make sure the source of 80 and 8787 both are “0.0.0.0/0, ::/0”). In this page’s end, press “Review and Launch”.
- 2.3.2.10. After check the details of instance, press “Launch”. Then, the user interface will ask you to choose an existing key pair or create a new key pair. (The step is quite important because the key can use to connect via SSH). If it is the first time to launch instance, create a new key pair. Then, press “Launch Instance”.
- 2.3.2.11. Wait for 2 to 8 minutes until the Instance State turn yellow to green light.
- 2.3.2.12. Copy the Public DNS (IPv4) and paste to URL bar, then press enter for entering the user interface of your new launched instance.
- 2.3.2.13. Input Username “rstudio” and Password which is preset your instance ID provided by AWS (such as this kind of format i-09d7c7c7e455d5612).
- 2.3.2.14. After enter the R studio user interface, the following steps are the same as usual computers for operating R. (Due to the user interface of AWS is frequently changing, the latest process will

continually update on GitHub from time to time if the first author is still alive, at https://github.com/gn03138868/WTF_sub-algorithm)

2.4. Input data

After collecting raw data, saving file as .csv, and building environment for running codes, open R studio software and run the code as below:

```
inputrawdata.R10.k.fold.X.training
<- read.csv("D:/Folder/rawdata k-fold X training.csv",
# Define the pathway of your input .csv files #
          header=T, row.names = 1,
# It means the first row is title #
          sep=",")
# Separate each value via comma #
          head(inputrawdata.R10.k.fold.X.training)
# Show the input raw data #
          summary(inputrawdata.R10.k.fold.X.training)
# Basic information of input raw data #
          str(inputrawdata.R10.k.fold.X.training)
# Show the structure of input raw data #
```

The name defined after data entry does not have a mandatory form. It is possible named logically and easy to recognise as well. Furthermore, sometimes, different countries' pathway in R perhaps has different styles. If you are not sure about it, please google it and refer other pathways' styles.

2.5. Install packages in R

Install packages in R before running the KS.

```
install.packages("mice")
require(mice)
# Borrow the framework from MICE for our wandering through #
# the random forests (WTF) sub-algorithm #
```

At the beginning, we wanted to write the entire WTF architecture by ourselves, but found that there would be thousands or even tens of thousands lines of codes, which was very inconvenient for policymakers to understand and apply. Fortunately, in the R community, there are packages that have been released and similar frameworks can be applied for our research purpose. **MICE** is an excellent one of them. [The R package mice imputes incomplete multivariate data by chained equations constructed by van Buuren and Groothuis-Oudshoorn (2011). The original purpose in designing MICE was to use chained equations to impute missing values in the data sheet. Following this, Shah et al. (2014) used the framework of MICE and combined random forests to impute missing values. In our purpose is to estimate the unknow and chaotic target variables related to forest CSSs.]

```
install.packages("Hmisc")
require(Hmisc)
# Package for Pearson's correlation test #

install.packages("corrplot")
require(corrplot)
# Package for correlation matrix #
```

If the total number of variables is below 50, it will be more convenient to use `corrplot`. If not, `Hmisc` must to be installed. We will elaborate more details about Pearson's correlation in section 3.

```
install.packages("ggplot2")
require(ggplot2)
install.packages("GGally")
require(GGally)
install.packages("ggfortify")
library(ggfortify)
# Install packages for visualisation #
```

It is possible to draw different kinds of figures after installing above packages in R.

2.6. Data observation

After finish input data and install packages, observe the data for checking missing values.

```
md.pattern(inputrawdata.R10.k.fold.X.training)
# Observe missing values #
```

3. Feature selection for LR models – 1st layer

3.1. Single features selecting (correlation tests)

We wanted to use stepwise for feature selection for making multilinear regression model initially. However, our data has a special character - a lot of missing values exist in data sheet, which means not every research investigate usual features in their own research. And after several times modification and rolling correction improved by conferences and seminars, we decided to construct simple LR models for estimating target variables which could be predicted by IVs of the 1st layer of the KS and must be in high accuracy and precision and to leave the irregular and divergent target variables to the next layer – the WTF sub-algorithm for estimating the target variables contained a lot of unobservable variables. Hence, we use Pearson (Pearson, 1901), Spearman (Spearman, 1906), and Kendall (Kendall, 1938) correlation methods to select features of each IV. The R codes describe as below:

```
cor.test(~Variable A + Variable B,
Data's name,
method = "spearman", "pearson" or "kendall")

# Correlation tests #
```

where the **Variable A** is target variable, the **Variable B** is IV, **Data's name** is the name of data you input in R, **method = "_____"** is for determining the method who want to use for correlation test.

If the Pearson correlation coefficient $r > 0.85$ and $p < 0.01$, construct LR models for the target (dependent) variables. If not, transfer the estimation work to the next layer of the KS.

3.2. Constructing LR models

The IVs with high precision and accuracy for target variables can be screened out after analysing all the correlation between target and IVs. Then construct a simple LR model for each target variable as the Eq (S2) and (S3) below.

$$Y = aZ + b \quad \text{Eq(S2)}$$

$$Y = c \times e^{aZ} \quad \text{Eq(S3)}$$

where Y is a target variable, Z is an IV, a is a coefficient of the IV, b is constant, e is Euler's number, and c is the coefficient of e .

The above equations can be easily made in Microsoft Excel even if in R as the same, which software policymakers use depending on which softwares they are familiar with.

```
LRmodelsName = lm(Variavle B~ Variable A,
data = Data's name,
na.rm = TRUE)

# Construct simple LR model as Eq(S2) #
```

where the **RLmodelsName** is the Regression models you named, **lm(Variavle B~ Variable A)** is a function code in R statistics software.

```
LRmodelsName = lm(Variavle B~ exp(Variable A),
data = Data's name,
na.rm = TRUE)

# Construct simple (exponential) LR model as Eq(S3) #
```

where the **exp(Variable A)** is an exponential function code in R statistics software.

```
summary(LRmodelsName)

# Statistics summary #

# If the  $R^2 < 0.85$  or  $> -0.85$ , it is good for constructing LR models. #
```

Select IVs which can be used to estimate target variable in high accuracy and precision via above simple correlation tests.

3.3. Diagnosis of LR models

After constructing leaner regression models, it is better to do some diagnosis than never.

Before operating the diagnosis, draw the figure of distribution first.

```
qplot(x = Variable B,  
      y = Variable A,  
      data = Data's name, na.rm=TRUE)  
# Observing distributions of target and IVs. #
```

where the **Variable A** is target variable, the **Variable B** is IV, **Data's name** is the name of data you input in R, **na.rm=TRUE** is for negating missing values.

Theoretically, after the regression model is fitted, its residual value must conform to the three assumptions of normality, independence, and homogeneity of variance, so the next three tests are used. Check whether these three conditions are satisfied or not.

```
shapiro.test(LRmodelsName$residual)  
# Normality test #
```

where the **LRmodelsName** is the name of LR models constructed above.

If the p -value is high (perhaps over 0.05 depended on the confidence level), it cannot reject the null hypothesis which means the assumption of normal distribution of residual value is reasonable. If the p -value is low, the normal distribution of the residual value is unreasonable.

```
durbinWatsonTest(LRmodelsName)  
# Independence test #
```

If the p -value is high (perhaps over 0.05 depended on the confidence level), it cannot reject the null hypothesis which means the independence of the residual value is reasonable. If the p -value is low, the independence of the residual value is unreasonable.

```
ncvTest(LRmodelsName)  
# Homogeneity of variance test #
```

If the p -value is high (perhaps over 0.05 depended on the confidence level), it cannot reject the null hypothesis which means the homogeneity of variance of the residual value is reasonable. If the p -value is low, the homogeneity of variance of the residual value is unreasonable.

Through the above R codes and processes, simple LR models can be prepared for estimating target variables in the 1st layer of the KS approach if above conditions are satisfied. Input estimated values, which obtained from those simple LR models, into the washed data matrix as an input matrix for the 2^{ed} layer's training.

4. WTF sub-algorithm – the 2^{ed} layer

4.1. Background of the WTF sub-algorithm

A case of simple LR models has good performance in estimating AGC stocks but awful in belowground (Fig. S4). The major reason is that target variables are always affected by a lot of environmental and anthropogenic factors (Table S2 in main text) caused the correlation between the dependent and IVs are easily irregular even divergent and chaotic.

For conquering the problem, we used a ML method called the WTF sub-algorithm, in the 2^{ed} layer to solve the problems that the 1st layer cannot handle. The WTF sub-algorithm is composed of random forests and a Markov chain Monte Carlo (MCMC) type algorithm - Gibbs sampling. The pseudocode of the WTF sub-algorithm describes in Table S1.

Through the WTF sub-algorithm, target variables with complex IVs can be estimated higher accurately and precision than simple LR models. Below we describe the process of implementing R statistical operations.

4.2. The 2^{ed} layer in practice

For finding the lowest loss function of combination in duplication, CARtrees, and iteration numbers, we input training dataset for running R codes via cluster computing or cloud computing. If we write the framework of the WTF algorithm completely by ourselves, it will cost quite a lot of time and need to debug. Fortunately, “mice” package provides the corresponding Gibbs sampling architecture, and can be bridged to the “randomForest” package. So here borrowed and combined these two packages as the basis for running the WTF sub-algorithms. We appreciate to these research teams, who wrote these two packages, and give them respect. Now it is possible run this WTF sub-algorithm with just a few lines of codes.

4.2.1. After making some simple LR models, estimating the NA values, and inputting in original datasheet as a new one for the 2^{ed} layer, load the new datasheet into Rstudio.

```
inputrawdata.R10.k.fold.X.training
<- read.csv("D:/Folder/rawdata k-fold X training for L2.csv",
# Define the pathway of your input .csv files #
header=T, row.names = 1, # It means the first row is title #
sep=",") # Separate each value via comma #
head(inputrawdata.R10.k.fold.X.training) # Show the input raw data #
summary(inputrawdata.R10.k.fold.X.training)
# Basic information of input raw data #
str(inputrawdata.R10.k.fold.X.training)
# Show the structure of input raw data #
```

4.2.2. Then, start to training out the minimised loss function combination of duplication, CARtrees, and iteration numbers. The implemented R code is as follows.

```
## m=X; maxit=Y; trees=Z ##
## It is better to try different combinations of X, Y, and Z based on different data ## characteristics
##
mice.data.R10.k.fold.k.training.1.mXmYtZ
  <- mice(inputrawdata.R10.10.fold.2.training.1, m = X,
    # generate 200 data sheets #
      # It is possible to generate more than 100 sheets# # if you like #
      maxit = Y,          # max iteration times#
      method = "rf",      # CART means Classification And Regression Tree #
      # rf means random forest #
      # (It is also possible choose other methods such as LRs #
      # logistic regression, cart, random forest, bootstrap if you'd already
    # installed related packages) #
      ntree=Z,
    # ntree means CARtrees number in random forest #
      seed = 188)
  # set.seed()Let the sample survey data be the same every time #
summary(mice.data.R10.k.fold.k.training.1.m200m150t100)
save.image("Z:/... ../result_file_name.RData")
# This is for defending the pathway of output results #
## Performed by i5-9400F CPU @ 2.90GHz 2.90 GHz ##
## From 14Aug2020 13:53 to 16Aug2020 23:53 ETA 32hr06min ##
## These two lines are optional, but if there is a record, ##
## it is possible to record the calculation cost of the running algorithm. ##
```

4.2.3. After running the above codes, call results out for calculating.

```
miceX.R10.k.fold.k.training.1.mXmYtZ <-
  complete(mice.data.R10.k.fold.2.training.1.m200m10t10, X)
# first data sheet called as mice1#
# repeat this code to call all duplicates X = 1, 2, 3, ..., X #
  head(miceX.R10.k.fold.k.training.1.mXmYtZ)

  miceX.R10.k.fold.k.training.1.mXmYtZ
  [is.na(miceX.R10.k.fold.k.training.1.mXmYtZ)]<-0
# replace NA to 0 for further calculations #
```

4.2.4. After call out results, calculate MSE of different X.

```
mice.mean1_10.R10.k.fold.k.training.1.mXmYtZ<-  
(mice1.R10.k.fold.k.training.1.mXmYtZ  
+mice2.R10.k.fold.k.training.1.mXmYtZ  
+mice3.R10.k.fold.k.training.1.mXmYtZ  
+ ...  
+miceX.R10.k.fold.k.training.1.mXmYtZ  
)/X  
# Take the means of sum of the matrix, #  
# which is the largest number of duplication X, #  
# as the mean of maximum duplicates dataset values (Mi). #  
head(mice.mean1_10.R10.k.fold.k.training.1.mXmYtZ)
```

4.2.5. After calculating the MSE of different combination of X, Y, and Z, then draw the relationships between $\lambda(M)$ and the different combination of them (Fig. S5).

4.2.6. Choose the combination with the lowest $\lambda(M)$ for further testing set after all combination of $\lambda(M)$ have converged. For instance, in the case of Fig. S5, the combination with the lowest $\lambda(M)$ is X = 150, Y=150, and Z=100.

After above process, we get a trained combination for one of k folds in training set.

5. k -fold cross validation¹

5.1. Combine original training and validation dataset as one .csv file after get the best combination of duplications, iterations, and numbers of CARTrees.

5.2. Then, input variables via simple LR models made from step 3.2. [In this step, it is possible to use Excel, MATLAB, or R which the operator prefers to use.]

5.3. Run the same process from step 4.2.1. to 4.2.4. and get the mean of output matrix from the 2^{ed} layer.

6. Climate classification table – the 3rd layer

If the mean of output matrix has missing values inside, which means the output target variable have not pass the OOB (out-of-bag) test mentioned in Table S1, need to follow IPCC Guidelines for National Greenhouse Gas Inventories and make Table S4 and S5 in main text prepared from Dataset S1. Then, input the values into the output of the 2^{ed} layer as the output of the 3rd layer which is also the final output of the KS. (At least, using the average of the global data can ensure that the accuracy of the estimated results is high.)

7. Final output compared with benchmark

¹ Do not to grab "testing data set" into the WTF sub-algorithm (the 2^{er} layer) for training or to find parameters.

7.1. Obtain the benchmark

Use the IPCC Guidelines for National Greenhouse Gas Inventories and re-estimate the target variables to replace the original data matrix as the benchmark data matrix.

7.2. Performance calculating

Calculate the mean absolute percentage error (MAPE), root mean square error (RMSE), mean absolute error (MAE), mean error (ME), which mentioned in main text of Equation (2) to (4), of the benchmark and the estimated KS output provided by k -fold cross validation. Of course, it will also get each observation's absolute percentage error (APE), absolute error (AE), and error. For further statistical process. (Meaning of indices for the performance of algorithms describe in Fig. S6.)

7.3. Statistical process of different algorithms

After obtain the benchmark and performance calculating, save the results data as a .csv file. And choose one of the most familiar statistic software of yourself and input the data into the software for further process. The flowchart of statistics process of comparing the performance between different algorithms as shown in Fig. S7. In this study, we use R statistic software to operate it.

7.3.1. Shapiro-Wilk normality test

This step is for check that the group is normal distribution or not (Royston, 1982). If the p -value > 0.05 , the data is normal distribution; if the p -value < 0.05 , the data is not normal distribution.

7.3.2. Homogeneity of variances' test in multiple groups

It is better to check the homogeneity of variances before recognising those each group is significant different or not via Bartlett's or Levene's test. If the result of homogeneity of variances' test is homogeneity, use one-way ANOVA; if not, use Welch's Heteroscedastic F Test.

7.3.2.1. Bartlett's test

Bartlett's test is suit for normal distributions groups. If the p -value > 0.05 , variances are homogeneity; if the p -value < 0.05 , variances are heterogeneity.

7.3.2.2. Levene's test

Levene's test is suitable for other types' distribution. If the p -value > 0.05 , variances are homogeneity; if the p -value < 0.05 , variances are heterogeneity.

7.3.3. one-way ANOVA or Welch's Heteroscedastic F Test

If the p -value < 0.05 , difference is statistically significant; if not, difference is statistically insignificant ($\alpha = 0.05$).

7.3.4. Post-hoc test

If variances in one-way ANOVA are significant different, just operate HSD, LSD (for same samples' number) or Scheffe test (for different samples' number). Or if the variables are heterogeneity and the results of Welch's Heteroscedastic F Test is significant different, the post-hoc test is better to use Games-Howell method.

8. Holdout testing (optional)

Holdout testing is used to test whether the model has overfitting (That is, it performs well when training the model, but it performs poorly in empirical applications.). Holdout testing for validating the performance is very important if we want to observe the generalisation ability of ML models (such as the convolutional neural network [CNN]). The network of CNN becomes a very flexible function mapper by adding more neurones in the hidden layer. This raises the risk of overfitting. However, it is elegance to prevent this kind of overfitting by out-of-bag (OBB) test in the step of random forests (Breiman, 2001) of the WTF sub-algorithm of the KS approach. Although some studies suggest that holdout testing is unnecessary to use when the number of observations is not enough, we still strongly recommend that if the number of samples is sufficient, it is nice to operate it. After all, each dataset has different characteristics.

The testing is good to use if operators have a large dataset. The dataset will split 20% to 30% of observations for holdout testing generally and isolated from the training and validation dataset. (Some studies will call training dataset, validating dataset, and testing dataset the same as training data set, validation dataset and holdout dataset in this study. It is because we want to prevent misunderstanding in the study in which the training dataset splits into a training set and a validating set in OOB test in the WTF sub-algorithm based on the Gibbs sampling and random forests.)

8.1. Attach observations of holdout dataset to the training dataset.

8.2. Remove nearly all target variables and IVs, and retain those traits which policymakers have.

8.3. Operate the KS approach.

8.4. Get the output of target variables.

8.5. Calculate the MAPE, RMSE, MAE, and ME as the same as step 7.2.

8.6. If the performance of holdout testing is better than k -fold cross validation or there is no significant difference, the model is not overfitting. Congratulations!

9. Estimating target variables by the KS approach

9.1. Attach observations of unknow places which policymakers perhaps just have limited traits of them. (Input lower than $1/k$ % observations when operate the KS approach once by once.)

9.2. Operate the KS approach.

9.3. Get the output of target variables.

10. Final Visualisation

After running all the processes above, it is possible to visualise the result in table or box plot depended on the preference of the operator. In this study, we reveal the result in Table S6. The latest version of the process will

continually be updated on GitHub (https://github.com/gn03138868/WTF_sub-algorithm). If there are some bugs in the code script, please report to the first author or corresponding author. We sincerely thank you.

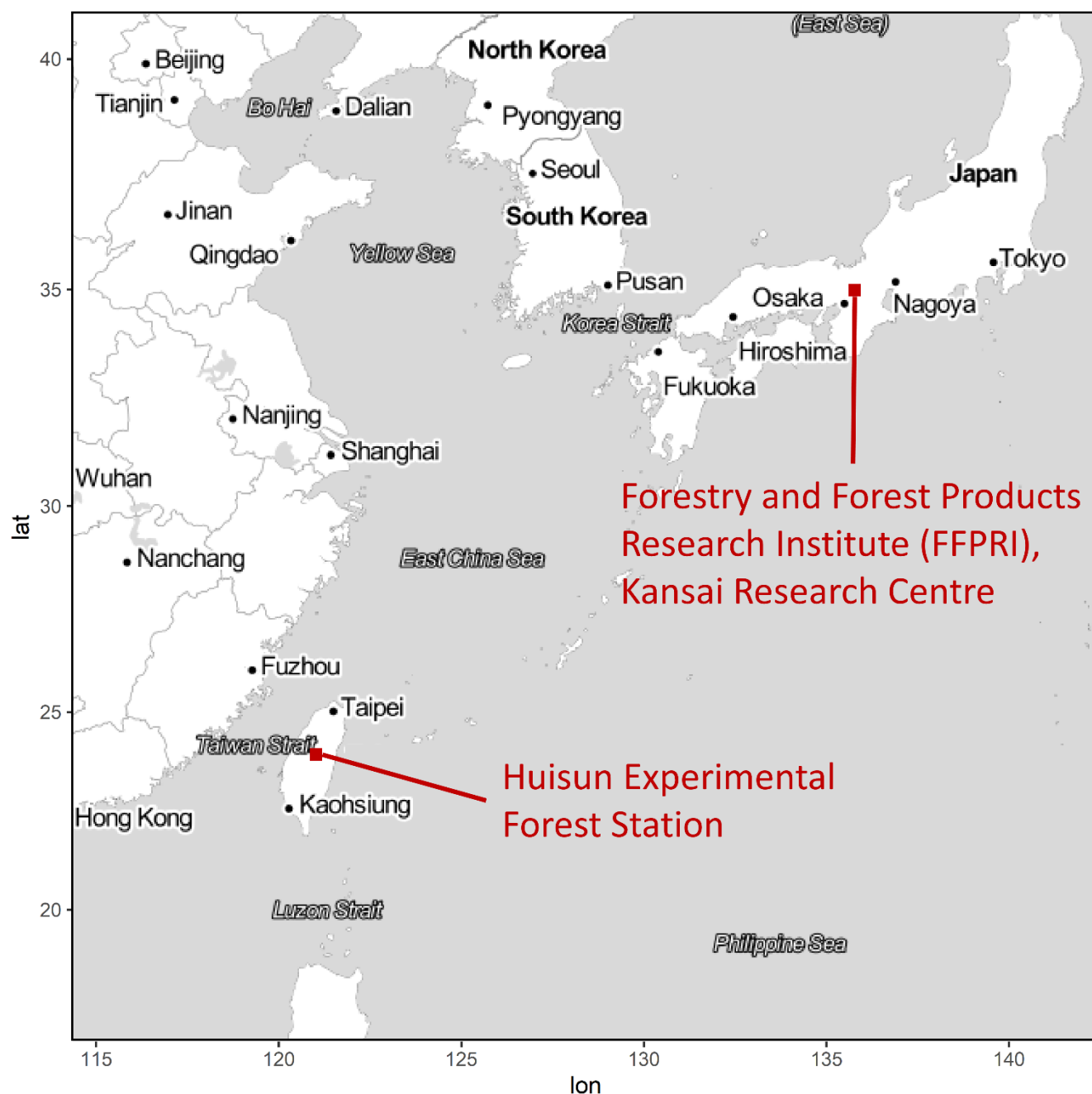


Fig. S1. Location of the study sites in Huisun Experimental Forest Station and in Forestry and Forest Products Research Institute (FFPRI), Kansai Research Centre.

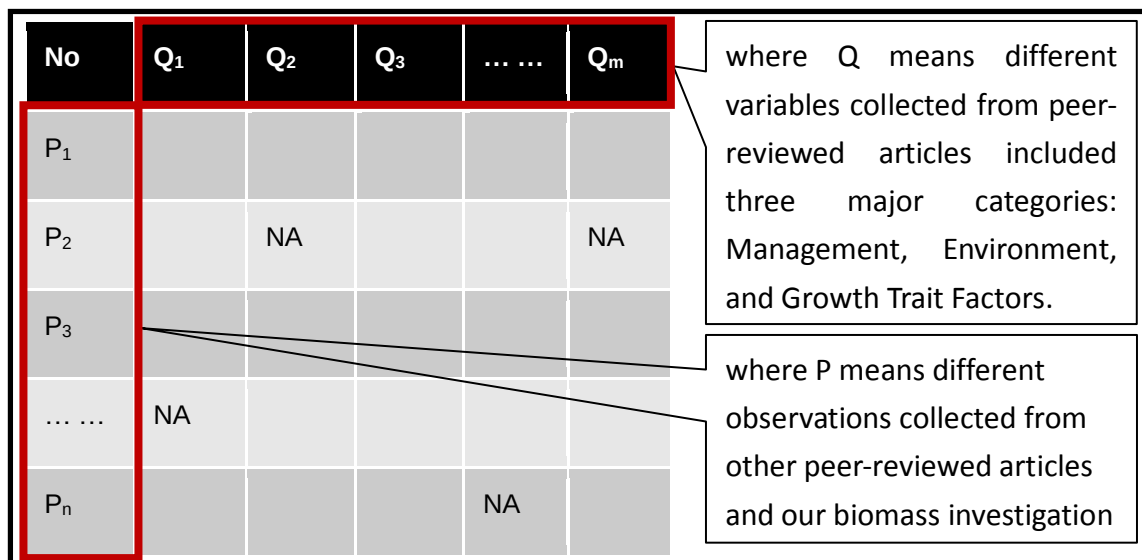


Fig. S2. Frameworks of raw data for further analysis of the KS approach.

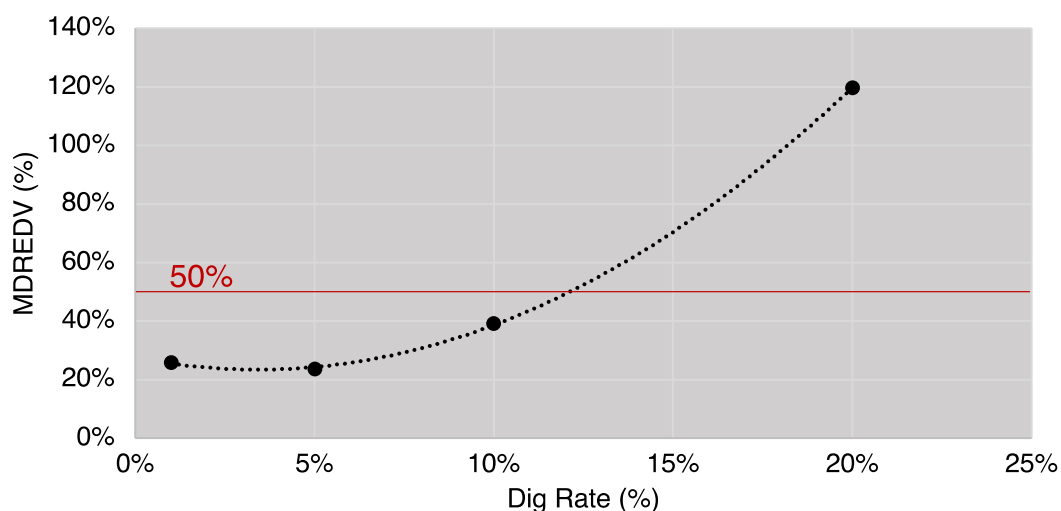


Fig. S3. Correlation between dig rate and mean absolute different ratio of each dug value (%) in the original datasheet estimated via random forest with Gibbs sampling. Gibbs sampling is a Markov chain Monte Carlo (MCMC) algorithm for obtaining a sequence of observations which are approximated from a multivariate probability distribution when direct sampling is difficult.

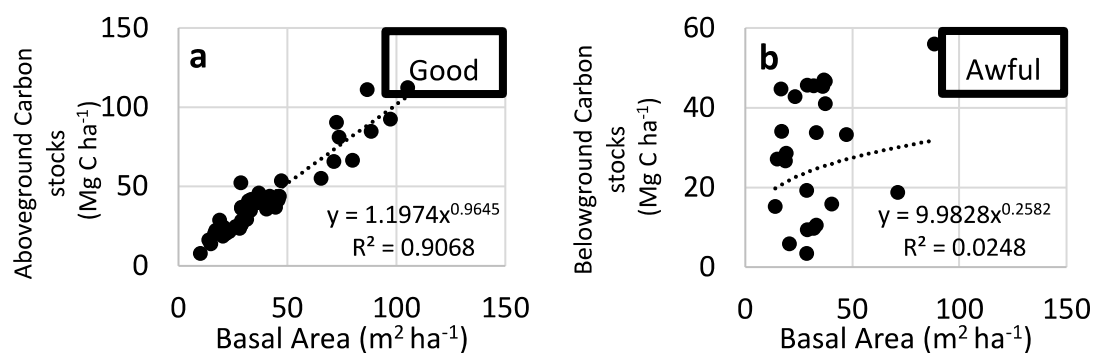


Fig. S4. LR models for estimating AGC and BGC stocks. (a) A simple LR model for estimating AGC with the BA as an IV. (b) A LR model for estimating BGC with the BA as an IV.

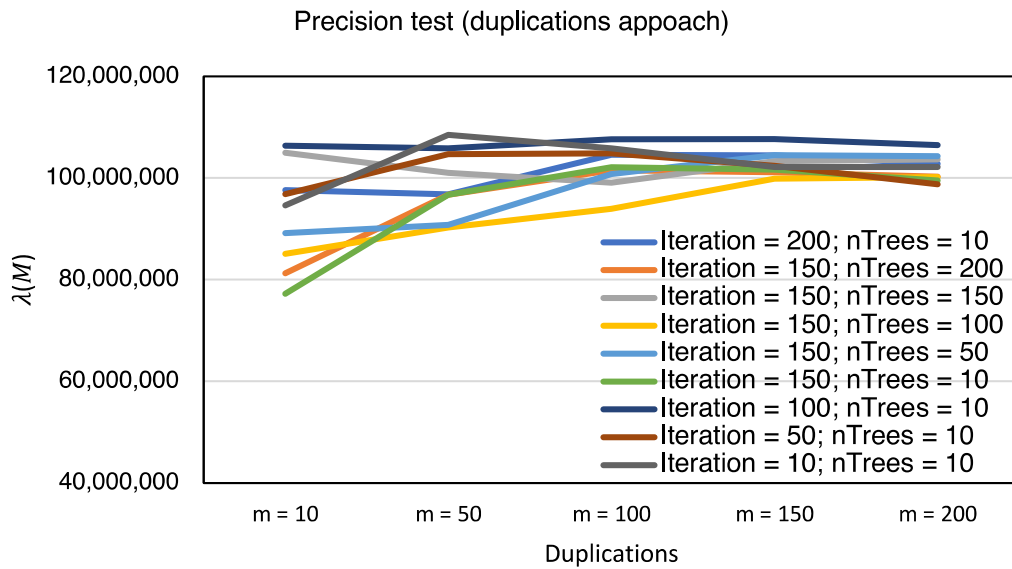


Fig. S5. Relationships between $\lambda(M)$ and the different combinations of duplications, max iteration, and nTrees numbers in one of k folds.

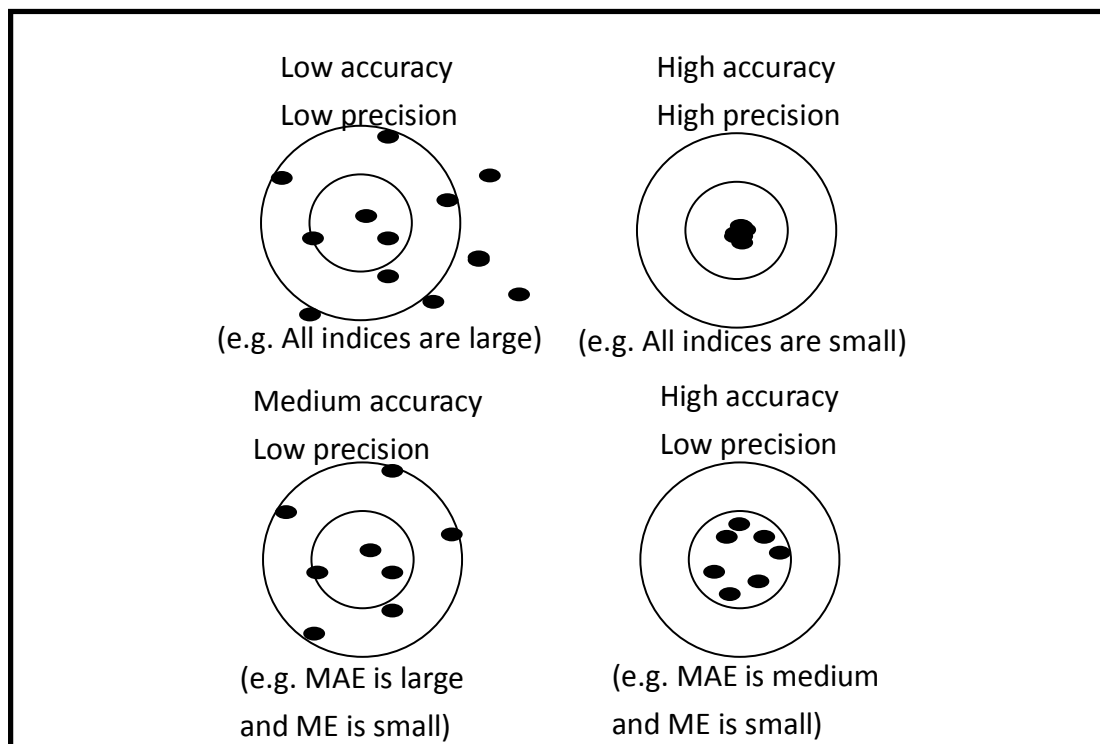


Fig. S6. Meaning of indices for the performance of algorithms.

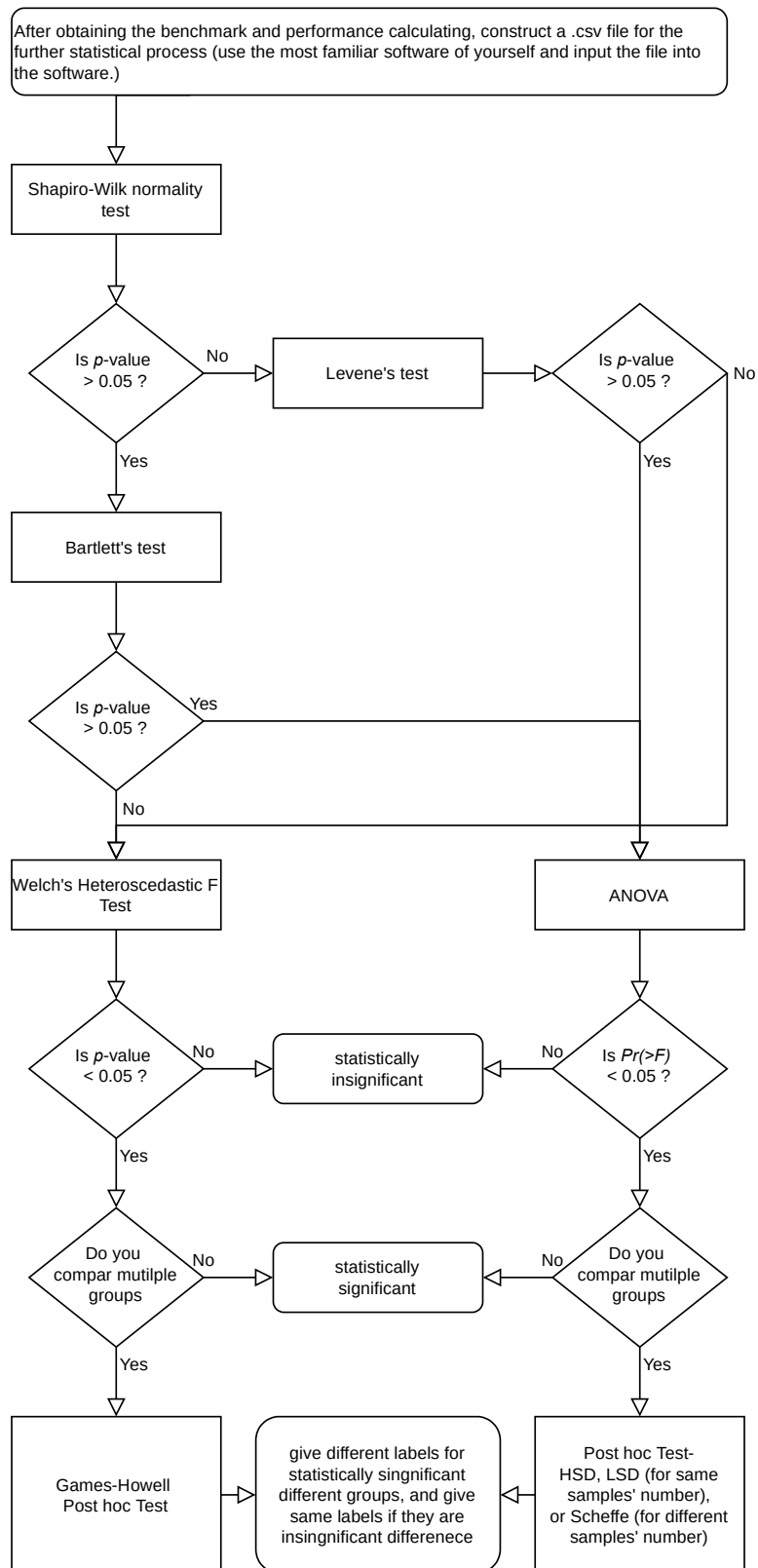


Fig. S7. Flowchart of statistical process of different methods.

No	Q_1	Q_2	Q_t		Q_m
P_1					
P_2		NA			NA
P_3					
... ..	NA				
P_n				NA	
H_1		NA	NA	NA	NA
H_2		NA	NA	NA	NA
... ..		NA	NA	NA	NA
H_n		NA	NA	NA	NA

where Q_t means target variables

where H means different observations collected from other peer-reviewed articles and does not involve the building process of KS approach for holdout testing

In this study, I removed nearly all values for holdout testing and just remained a few traits such as DBH and mean annual temperature that policymakers often have.

Fig. S8. Frameworks of holdout dataset for testing the generalisation ability of the KS approach.

Table S1. Pseudocode of the WTF sub-algorithm in the KS approach.

Algorithm the WTF sub-algorithm in the KS approach	
1: initialise $\mathbf{X}^s \sim \mathbf{X}^{(0)}$	
2: for duplicate $s = 0, 1, 2, \dots$ do	
1: initialise $\mathbf{X}^{(0)} \sim q(\mathbf{X})$	
2: for iteration $i = 0, 1, 2, \dots$ do	
1. $\mathbf{X}_1^{(i)} \sim p(\mathbf{X}_1 = x_1 \mathbf{X}_2 = x_2^{(i-1)}, \mathbf{X}_3 = x_3^{(i-1)}, \dots, \mathbf{X}_D = x_D^{(i-1)})$	
1. procedure random forest	
2. for 1 to T do	
3. Bootstrap $\mathbf{D}$ for forming a new $\mathbf{D}_i$	
[The rest of points from $\mathbf{D}$ for out of bag (OOB) test]	
4. Build full classification and regression trees on $\mathbf{D}_i$	
5. end for	
6. ensemble all T trees	
7. end procedure	
2. $\mathbf{X}_2^{(i)} \sim p(\mathbf{X}_2 = x_2 \mathbf{X}_1 = x_1^{(i)}, \mathbf{X}_3 = x_3^{(i-1)}, \dots, \mathbf{X}_D = x_D^{(i-1)})$	
1. procedure random forest	
2. for 1 to T do	
3. Bootstrap $\mathbf{D}$ for forming a new $\mathbf{D}_i$	
[The rest of points from $\mathbf{D}$ for out of bag (OOB) test]	
4. Build full classification and regression trees on $\mathbf{D}_i$	
5. end for	
6. ensemble all T trees	
7. end procedure	
3.	
4. $\mathbf{X}_D^{(i)} \sim p(\mathbf{X}_D = x_D \mathbf{X}_1 = x_1^{(i)}, \mathbf{X}_2 = x_2^{(i)}, \dots, \mathbf{X}_{D-1} = x_{D-1}^{(i)})$	
1. procedure random forest	
2. for 1 to T do	
3. Bootstrap $\mathbf{D}$ for forming a new $\mathbf{D}_i$	
[The rest of points from $\mathbf{D}$ for out of bag (OOB) test]	
4. Build full classification and regression trees on $\mathbf{D}_i$	
5. end for	
6. ensemble all T trees	
7. end procedure	
3. end for	
3: average all $\mathbf{X}_s$	
4: end for	

Table S2. Common terms and definitions of variables associated with photosynthesis, respiration, and carbon cycles in this these studied *P. edulis* forests.

category of factors	Variables	Units	Definition
Time	Beginning year	yr	The recorded starting year of each study.
	Finished year	yr	The recorded ending year of each study.
Management	Managed	[0,1]	A dummy IV which indicates <i>P. edulis</i> forests have been managed or not.
	Fertilised	[0,1]	A dummy IV which indicates <i>P. edulis</i> forests have been fertilised or not.
	Weeded and selective cut	[0,1]	A dummy IV which indicates common management methods operated in <i>P. edulis</i> forests.
	Shoot dug	[0,1]	A dummy IV which indicates <i>P. edulis</i> forests have been dug bamboo shoots or not.
	Clear cutting	[0,1]	A dummy IV which indicates <i>P. edulis</i> forests have been clear cut or not.
	Mixed with other forests	[0,1]	A dummy IV which indicates <i>P. edulis</i> forests have mixed with other forests or not.
Environment	Mean annual temperature	°C	Average of mean daily temperatures in a year. (Some of research use the Average of mean monthly temperatures in a year depending on their definition of weather stations.)
	Warmth index		Warmth Index (WI) proposed by Kira (1945) is known as an index which is closely related to the distribution of vegetation. Warmth Index is calculated by the following equation. WI=Sum of (Tm-5) when Tm is above 5 °C (Tm: Monthly Mean Temperature)
	Annual rainfall	mm yr ⁻¹	The sum of daily rainfall in a year.
	Snow	mm yr ⁻¹	The sum of daily snowfall in a year.
	Elevation	m	The height above sea level.
	Relative humidity	%	A ratio of the amount of atmospheric moisture present relative to the amount that would be present if the air were saturated.
	Sunshine duration	hr yr ⁻¹	The sum of the time for which the direct solar irradiance exceeds 120 W m ⁻² .
	Wind speed	m s ⁻¹	The rate at which air is moving in a particular area.
	Water content (soil)	kg kg ⁻¹	A measure of the amount of water (mass) contained in a unit mass of soil.
	pH (soil-KCl)		A measure of the acidity or basicity (alkalinity) of a soil by a diluted solution of potassium chloride (KCl) used in the

			analysis, instead of water (H ₂ O), which gives the pH (soil-KCl) value.
	pH (soil-H ₂ O)		A measure of the acidity or basicity (alkalinity) of a soil by a diluted solution of water (H ₂ O) which gives the pH (soil-H ₂ O) value.
	Total N (soil)	g kg ⁻¹	The fraction of nitrogen in soils.
	Total P (soil)	g kg ⁻²	The fraction of phosphorus in soils.
	Total K (soil)	g kg ⁻³	The fraction of potassium in soils.
	Available P ₂ O ₅ (soil)	mg kg ⁻¹	The available phosphorus pentoxide (P ₂ O ₅) for absorption by roots in soils.
	Available SiO ₂ (soil)	mg kg ⁻¹	The available silicon dioxide (SiO ₂) for absorption by roots in soils.
	Cation exchange capacity, CEC (soil)	cmole kg ⁻¹	The total capacity of soils to hold exchangeable cations.
	K ⁺ (soil)	me 100g ⁻¹	The amount of available potassium ions in soils.
	Ca ²⁺ (soil)	me 100g ⁻¹	The amount of available calcium ions in soils.
	Mg ²⁺ (soil)	me 100g ⁻¹	The amount of available magnesium ions in a soil.
	Sand (soil)	%	The fraction of soil particles that the size range between 2 and 0.063 mm.
	Silt (soil)	%	The fraction of soil particles that the size range between 0.063 and 0.002 mm.
	Clay (soil)	%	The fraction of soil particles that the size is smaller than 0.002 mm.
	N (litter)	kg ha ⁻¹	The amount of nitrogen in litter layers per hectare.
	Ca (litter)	kg ha ⁻¹	The amount of calcium in a litter layers per hectare.
	K (litter)	kg ha ⁻¹	The amount of potassium in a litter layers per hectare.
	Mg (litter)	kg ha ⁻¹	The amount of magnesium in a litter layers per hectare.
	P (litter)	kg ha ⁻¹	The amount of phosphorus in a litter layers per hectare.
	Si (storage in aboveground biomass)	kg ha ⁻¹	The amount of silicon stored in aboveground biomass per hectare.
	Si (storage in belowground biomass)	kg ha ⁻¹	The amount of silicon stored in belowground biomass per hectare.
	Si (storage in soil)	kg ha ⁻¹	The amount of silicon stored in soils per hectare.

	Si (primary sink in Plant annually)	kg ha ⁻¹ yr ⁻¹	The total amount of silicon sinked in <i>P. edulis</i> forests per hectare annually.
	Si (net sink in Plant annually)	kg ha ⁻¹ yr ⁻¹	The primary sink of silicon in <i>P. edulis</i> forests minus silicon which returns to soils
	Si (return to soil)	kg ha ⁻¹ yr ⁻¹	The primary sink of silicon minus the net sink in <i>P. edulis</i> forests per hectare annually.
	Relative luminosity	%	The relative brightness of spaces in <i>P. edulis</i> forests, normalised to 0 for darkest space in a black chamber and 1 for the highest luminosity in the open space outside the <i>P. edulis</i> forest.
Growth Trait	Culm density	culm ha ⁻¹	The number of living culms per hectare.
	Mean DBH	cm	The mean of living culms' DBH.
	Mean height	m	The mean of living culms' height.
	BA	m ² ha ⁻¹	The average amount of an area (usually a hectare) occupied by BA of culms.
	LAI (Fisheye lens)	m ² m ⁻²	The one-sided leaf area per unit ground surface area estimated by fish eye lens in <i>P. edulis</i> forests. [leaf area index (LAI) = leaf area / ground area]
	LAI (leaf area scanner)	m ² m ⁻²	The one-sided leaf area per unit ground surface area estimated by scanner for cauculating leaf area in <i>P. edulis</i> forests. [leaf area index (LAI) = leaf area / ground area]
	Leaves C	%	The carbon fraction of leaf dry matters.
	Branches C	%	The carbon fraction of branch dry matters.
	Culms C	%	The carbon fraction of culm dry matters.
	Fine roots C	%	The carbon fraction of fine root dry matters.
	Coarse root C	%	The carbon fraction of coarse root dry matters.
	Rhizomes C	%	The carbon fraction of rhizome dry matters.
	Stump C	%	The carbon fraction of stump dry matters.
	Soil C (0-10cm)	%	The carbon fraction of dry soil from the depth of 0 cm to 10 cm.
	Soil C (10-30cm)	%	The carbon fraction of dry soil from the depth of 10 cm to 30 cm.
	Foliage	Mg C ha ⁻¹	The leaf biomass (carbon equivalent).
	Branches	Mg C ha ⁻¹	The branch biomass (carbon equivalent).
	Culms	Mg C ha ⁻¹	The culm biomass (carbon equivalent).
	AGC	Mg C ha ⁻¹	The aboveground biomass (carbon equivalent).
	Root_shoot ratio		The ratio of underground divided by aboveground biomass (carbon equivalent).
	Roots	Mg C ha ⁻¹	The root biomass (carbon equivalent).

	Rhizomes	Mg C ha ⁻¹	The rhizome biomass (carbon equivalent).
	Stumps	Mg C ha ⁻¹	The stump biomass (carbon equivalent).
	BGC	Mg C ha ⁻¹	The belowground biomass (carbon equivalent).
	TC	Mg C ha ⁻¹	The total biomass (carbon equivalent).
	Litter layer	Mg C ha ⁻¹	The litter mass (carbon equivalent).
	SC	Mg C ha ⁻¹	The soil carbon storage.
	Undergrowth	Mg C ha ⁻¹	The undergrowth biomass (carbon equivalent).
	TEC	Mg C ha ⁻¹	The total ecosystem carbon which includes above- and below-ground live components, dead biomass, and SC.
	LNP	Mg C ha ⁻¹ yr ⁻¹	The leaf net primary productivity.
	BNP	Mg C ha ⁻¹ yr ⁻¹	The branch net primary productivity.
	CNP	Mg C ha ⁻¹ yr ⁻¹	The culm net primary productivity.
	litterfall	Mg C ha ⁻¹ yr ⁻¹	The litter fall per hectare
	ANPP	Mg C ha ⁻¹ yr ⁻¹	The aboveground net primary productivity. (equal to LNP+BNP+CNP+ litter fall)
	RoNP	Mg C ha ⁻¹ yr ⁻¹	The root net primary productivity.
	RhNP	Mg C ha ⁻¹ yr ⁻¹	The rhizome net primary productivity.
	SNP	Mg C ha ⁻¹ yr ⁻¹	The stump net primary productivity.
	BNPP	Mg C ha ⁻¹ yr ⁻¹	The belowground net primary productivity. (equal to RoNP+RhNP+SNP)
	TNPP	Mg C ha ⁻¹ yr ⁻¹	The total net primary productivity (equal to ANPP+BNPP)
	Soil respiration (SR)	Mg C ha ⁻¹ yr ⁻¹	The respiration of soils. (equal to belowground autotrophic respiration + heterotrophic respiration)
	Heterotrophic respiration (HR)	Mg C ha ⁻¹ yr ⁻¹	Total respiration excluded autotrophic respiration. (also equal to TNPP-NEP)
	Net ecosystem production (NEP)	Mg C ha ⁻¹ yr ⁻¹	The difference between organic carbon fixed by photosynthesis and total ecosystem respiration including autotrophic and heterotrophic respiration. (equal to gross primary production minus ecosystem respiration, or to TNPP minus heterotrophic respiration)

Table S3. Results of Pearson's correlation test between dependent and IVs. BA is the average amount of an area (usually a hectare) occupied by culms of *P. edulis* forests. Weeded and selective cutting are common management methods operated in *P. edulis* forests. Managed is a dummy IV which indicates *P. edulis* forests have been managed or not. Root_Shoot ratio is defined as the dry weight of belowground biomass divided by the dry weight of aboveground biomass in this study. Sunshine duration is defined as the sum of the time for which the direct solar irradiance exceeds 120 W m⁻². Shoots dug is a dummy IV which indicates *P. edulis* has been dug bamboo shoots or not. Total K in the soil means the amount of Potassium in the soil of *P. edulis* forests. RoNP is the roots' net primary productivity in *P. edulis* forests. LNP is the leaves' net primary productivity in *P. edulis* forests. Total N in the soil means the amount of nitrogen in the soil of *P. edulis* forests. BNP is the branches' net primary productivity in *P. edulis* forests. CNP is the culms' net primary productivity in *P. edulis* forests. TNPP is the total net primary productivity in *P. edulis* forests. HR is heterotrophic respiration in *P. edulis* forests. TEC is the total ecosystem carbon which includes above- and below-ground live components, dead biomass, and SC. SR is soil respiration in *P. edulis* forests. Y means that the IV is used in layer 2; X means the IV is not used in layer 2.

Dependent (target) variable	IVs	Pearson's product-moment correlation	degree of freedom (df)	p-value	Use in layer 2
AGC storages	BA	0.917	80	2.20×10 ⁻¹⁶	Y
	Culms	0.940	59	2.20×10 ⁻¹⁶	Y
	Culm height	0.676	43	2.78×10 ⁻⁸	X
	Culm DBH	0.417	67	1.08×10 ⁻⁷	X
	TC	0.539	45	1.57×10 ⁻⁶	X
	Branches	0.403	56	3.74×10 ⁻⁶	X
	Weeded and selective cutting	0.141	80	0.001105	X
	Relative humidity	0.187	25	0.002506	X
	Managed	0.095	80	0.006891	X
	Root_Shoot ratio	0.198	45	0.007030	X
	Culm density	0.106	71	0.007747	X
	Sunshine duration	0.135	58	0.008224	X
	Branches carbon ratio	0.400	26	0.016190	X
	Wind speed	0.102	58	0.019650	X
	Leaves carbon ratio	0.371	26	0.020940	X
	Fine roots carbon ratio	0.789	19	0.028130	X
	Shoots dug ⁶	0.057	80	0.030010	X
	Annual rainfall	0.048	79	0.044040	X
	Total K in soil	0.985	17	0.056050	X
	Snow	0.060	57	0.059760	X
	BNPP	0.580	19	0.083690	X
	RoNP	0.554	19	0.092330	X

BGC storages	Root_Shoot Ratio	0.546	45	1.27×10^{-6}	X
	TC	0.532	45	1.95×10^{-6}	X
	Rhizomes	0.623	38	2.73×10^{-6}	X
	Stumps	0.530	30	0.000840	X
	Culm height	0.652	25	0.001624	X
	LNP	0.550	23	0.013460	X
	Rhizomes carbon ratio	0.866	19	0.013960	X
	weeded and selective cutting	0.138	45	0.022430	X
	Litterfall	0.294	26	0.039950	X
	Annual rainfall	0.104	44	0.045810	X
Total carbon storages (TC storages)	BA	0.655	29	0.000064	X
	Culm height	0.832	9	0.001477	X
	Branches	0.407	23	0.043220	X
	Culms	0.687	23	0.000147	X
	Rhizomes	0.548	22	0.005542	X
	Stumps	0.719	14	0.001681	X
SC storages	Annual rainfall	0.445	13	0.096170	X
	TEC	0.704	9	0.015710	X
	Stumps	-0.954	2	0.045570	X
	Rhizomes	-0.622	7	0.073520	X
Aboveground net primary productivity (ANPP)	Snow	0.419	18	0.065830	X
	Total N in soil	0.957	2	0.042860	X
	Leaves carbon ratio	0.991	3	0.001060	X
	Branches carbon ratio	0.991	3	0.001060	X
	Culms carbon ratio	0.991	3	0.001062	X
	LNP	0.591	13	0.020350	X
	BNP	0.946	16	3.18×10^{-9}	X
	CNP	0.944	18	4.14×10^{-10}	X
	TNPP	0.936	4	0.005946	X
Belowground net primary productivity (BNPP)	Culm DBH	0.900	3	0.037490	X
	Culms	0.839	3	0.075700	X
	Roots	0.961	2	0.038670	X
	LNP	0.927	3	0.023590	X
	RoNP	0.918	4	0.009756	X

Table S4. Ratio of below-ground biomass to above-ground biomass (R) in *P. edulis* forests. Values are calculated from Dataset S1.

Domain	Ecological zone	Above-ground biomass	R [Mg root • (Mg shoot) ⁻¹]
Subtropical	Subtropical humid forest	above-ground biomass < 80 Mg ha ⁻¹	0.27 (0.13-0.34)
		above-ground biomass > 80 Mg ha ⁻¹	0.51 (0.34-0.63)
	Subtropical mountain systems		0.80 (0.24-1.98)
Temperate	Temperate oceanic forest	above-ground biomass < 80 Mg ha ⁻¹	1.32 (0.44-2.18)
		above-ground biomass > 80 Mg ha ⁻¹	0.48 (0.28-0.66)
	Temperate continental forest		0.37 (0.09-0.92)

Table S5. Tier 1 estimated biomass values of *P. edulis* referred to the method from table 4.7-4.10 (values are approximate; use only for Tier 1) in 2006 IPCC Guidelines for National Greenhouse Gas Inventories. Values are calculated from Dataset S1.

Climate domain	Ecological zone	Above-ground biomass in Moso bamboo forests (Mg MD ha ⁻¹)	Above-ground biomass in forests plantation of Moso bamboo (Mg MD ha ⁻¹)	Above-ground biomass growth in nature forests (Mg MD ha ⁻¹ yr ⁻¹)	Above-ground biomass growth in forests plantation of Moso bamboo (Mg MD ha ⁻¹ yr ⁻¹)
Subtropical	Subtropical humid forest	110.13 (28.58-221.35)	69.00 (16.00-185.65)	6.60 (1.09-12.14)	10.30 (5.71-18.80)
	Subtropical mountain systems	90.63 (43.30-162.27)	71.47 (44.78-105.11)	13.04 (7.50-17.92)	17.09 (6.90-46.71)
Temperate	Temperate continental forest	57.77	73.37 (52.66-97.80)	-	-
	Temperate oceanic forest	125.02 (23.31-296.50)	114.07 (45.20-217.30)	12.60 (2.6-23.6)	27.59 (15.72-51.00)

Table S6. Performance of k-fold cross-validation between benchmark (IPCC Tier 1) and KS approach.

AGC (n=77), BGC (n=39), total carbon (TC) (n=40), ANPP (n=33), BNPP (n=11), and TNPP (n=11). Average with different significant codes that the performance of the KS approach is significantly better than the benchmark by one-way ANOVA or Welch's Heteroscedastic F Test depending on whether observations are homogeneity or heterogeneity. The old black colour presents lower values (higher performance).

		IPCC Tier 1		LR only		KS approach	
		Mean	SD	Mean	SD	Mean	SD
mean absolute percentage error (MAPE)	AGC ¹ (%)	39.75%	23.90%	24.06%	25.32%	23.29%***	24.74%
	BGC (%)	122.12%	140.90%	86.25%	92.51%	41.84%**	25.15%
	TC (%)	34.32%	23.70%	38.25%	31.68%	26.22%`	15.61%
	ANPP (%)	172.38%	250.90%	104.55%	175.90%	48.92%**	38.53%
	BNPP (%)	505.34%	635.06%	68.09%	76.96%	59.30%*	65.84%
	TNPP (%)	328.44%	390.94%	117.75%	190.50%	52.35%*	51.45%
root mean square error (RMSE)	AGC (Mg C ha ⁻¹)	31.84	32.45	13.81	18.92	14.37***	19.75
	BGC (Mg C ha ⁻¹)	18.67	22.14	16.78	18.98	12.96*	13.50
	TC (Mg C ha ⁻¹)	30.86	33.53	23.40	24.53	20.03**	24.21
	ANPP (Mg C ha ⁻¹)	6.64	10.91	6.66	10.49	4.48	6.88
	BNPP (Mg C ha ⁻¹)	4.16	5.7	3.10	3.67	3.12	4.22
	TNPP (Mg C ha ⁻¹)	13.41	16.97	11.11	15.18	8.15	10.36
mean absolute error (MAE)	AGC (Mg C ha ⁻¹)	27.92	18.83	190.71	358.06	11.50***	13.61
	BGC (Mg C ha ⁻¹)	13.9	12.63	281.61	360.41	11.03***	6.89
	TC (Mg C ha ⁻¹)	25.08	18.21	547.52	601.95	16.87*	10.93
	ANPP (Mg C ha ⁻¹)	4.11	5.29	44.36	110.13	3.16	3.22
	BNPP (Mg C ha ⁻¹)	3.05	2.97	9.60	13.49	2.25	2.27
	TNPP (Mg C ha ⁻¹)	9.63	9.79	123.45	230.44	5.79	6.02
mean error (ME)	AGC (Mg C ha ⁻¹)	-19.39	27.62	9.89	9.70	-2.40***	17.70
	BGC (Mg C ha ⁻¹)	12.02	14.47	13.50	10.10	-0.45***	13.12
	TC (Mg C ha ⁻¹)	-11.05	29.18	19.52	13.07	-3.85	19.91
	ANPP (Mg C ha ⁻¹)	3.28	5.86	4.08	5.35	0.38*	4.53
	BNPP (Mg C ha ⁻¹)	3.02	3.00	2.34	2.12	-0.02*	3.27
	TNPP (Mg C ha ⁻¹)	9.63	9.79	6.99	9.06	3.46	7.74