# A Machine Learning Approach for Resource Mapping Analysis of Greenhouse Gas Removal Technologies

**Article** *in* Energy and Climate Change · July 2023

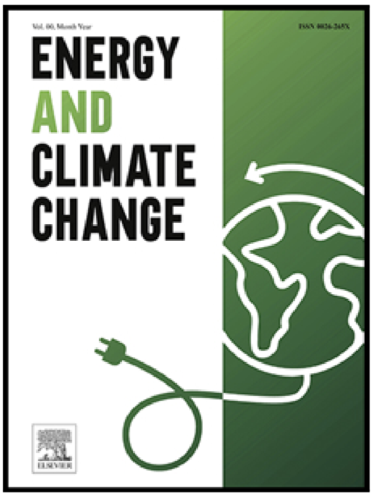
DOI: 10.1016/j.egycc.2023.100112

Jude Asibor
University of Benin
**13** PUBLICATIONS **283** CITATIONS

Peter Clough
Cranfield University
**64** PUBLICATIONS **3,075** CITATIONS

Ali Nabavi
Cranfield University
**112** PUBLICATIONS **2,900** CITATIONS

Vasilije Manovic
Cranfield University
**231** PUBLICATIONS **11,663** CITATIONS



# A Machine Learning Approach for Resource Mapping Analysis of Greenhouse Gas Removal Technologies

Jude O. Asibor , Peter T. Clough , Seyed Ali Nabavi , Vasilije Manovic

PII: S2666-2787(23)00019-3
DOI: https://doi.org/10.1016/j.egycc.2023.100112
Reference: EGYCC 100112

To appear in: *Energy and Climate Change*

Received date: 8 November 2022
Revised date: 15 June 2023
Accepted date: 17 July 2023

# A Machine Learning Approach for Resource Mapping Analysis of Greenhouse Gas Removal Technologies

Jude O. Asibor*, Peter T. Clough*, Seyed Ali Nabavi, Vasilije Manovic
Energy and Power Theme, School of Water, Energy and Environment, Cranfield University,
Bedfordshire, MK43 0AL, UK
* Corresponding author email: jude-odianosen.asibor@cranfield.ac.uk,
P.T.Clough@cranfield.ac.uk

## Highlights

- Locational suitability for GGR deployment was assessed using Machine Learning.
- Performance accuracy of the ML models was typically high.
- Models can be applied for various territorial scales of GGR deployment.

**Abstract**
In this study, machine learning (ML) was applied to investigate the suitability of a location to deploy five greenhouse gas removal (GGR) methods within a global context, based on a location's bio-geophysical and techno-economic characteristics. The GGR methods considered are forestation, enhanced weathering (EW), direct air carbon capture and storage (DACCS), bioenergy with carbon capture and storage (BECCS) and biochar. An unsupervised ML (hierarchical clustering) technique was applied to label the dataset. Seven supervised ML algorithms were applied in training and testing the labelled dataset with the k-Nearest neighbour (k-NN), Artificial Neural Network (ANN) and Random Forest algorithms having the highest performance accuracies of 96%, 98% and 100% respectively. A case study of Scotland's suitability to deploy these GGR methods was carried out with obtained results indicating a high correlation between the ML model results and information in the available literature. While the performance accuracy of the ML models was typically high (76 – 100%), an assessment of its decision-making logic (model interpretation) revealed some limitations regarding the impact of the various input variables on the outputs.

**Keywords**: Machine learning; climate change mitigation; carbon capture and storage; negative emission technologies; random forest; BECCS.

## Abbreviations

| | |
|---|---|
| AgLu | Agricultural Land use |
| BA | Biomass Availability |
| FLu | Forest Land use |
| GNI | Gross National Income Per Capita |
| GSP | Geological Storage Potential |
| LCEA | Low Carbon Energy Availability |
| Q1P, Q2P, Q3P, Q4P | Quarterly Mean Precipitation |

| | |
|---|---|
| Q1T, Q2T, Q3T, Q4T | Quarterly Mean Temperature |
| WA | Water Availability |

## 1 Introduction

Limiting the average global temperature rise to  1.5 °C as stated in the Paris Agreement requires a combination of a proactive emission reduction technological approach and an indispensable option to remove greenhouse gases (especially carbon dioxide) currently present in the atmosphere [1–3]. A number of greenhouse gas removal (GGR) methods have been proposed and assessed for the potential of large-scale deployment [4–8]. These methods include forestation, enhanced weathering (EW), soil carbon sequestration (SCS), biochar, direct air carbon capture and storage (DACCS), bioenergy with carbon capture and storage (BECCS), ocean fertilisation and wetland restoration among others. With the performance of these methods very much dependent on a complex interaction of bio-geophysical, techno-economic and socio-political factors, there is need to evaluate the GGR method(s) that are most suitable for deployment in a given location based on the location’s characteristics. Such a system is expected to provide unbiased information that will guide and enhance the decision-making process of concerned stakeholders on the choice of GGR methods that can be sustainably and efficiently deployed in a given location. Accordingly,  this study aims to utilise machine learning to provide an unbiased prediction of suitable GGR technology deployment worldwide.

Efforts to ascertain the suitability of deployment of GGR methods at local/national levels have been reported in the literature. While some addressed portfolio of GGR methods by analytically assessing the local availability of required resources [5,9–11], others employed various forms of optimisation modelling and decision-making techniques. Donnison et al. [12] applied a land-use spatial optimisation algorithm to determine the favourability of six potential UK locations to deploy BECCS from local biomass resources. By assessing indicators such as bioenergy crop yield, agricultural output, soil organic carbon, flood management, water stress level and land availability, the algorithm generated land-use scenarios for domestic bioenergy crop resources and compared the socio-environmental implications at each location quantitatively. In another study, Forster et al. [13] developed a traffic light system-based framework for feasibility assessment of GGR options within a national context. While it offers a flexible tool to guide participatory, multi-disciplinary, iterative decision making, it is however limited by its susceptibility to subjective interpretation with consequent likelihood of biased outcomes.

While these models and frameworks are largely localised and case-specific, Fajardy and Mac Dowell [14], developed a linear optimisation model, so-called  MONET framework which is applicable in multiple locations though with focus on evaluating optimal deployment conditions for BECCS. This spatio-temporal explicit model has now been updated  (known as the MONET-EU) to include other GGR methods, specifically forestation, DACCS and EW. Based on a variety of sustainability and bio-geophysical constraints, it aims to provide whole-system analysis of least-cost portfolio of GGR pathways, in ten-year time-steps [15]. Its application is however limited to regional and country-level assessment of the European Union (EU). Thus far, within the limits of the assessed literature, no effort has yet been made to apply Machine Learning (ML) in assessing the suitability of deployment of GGR methods for any location. Herein lies the novelty in this study. The use of ML minimises human bias, offers flexibility to adapt to the ever-changing bio-geophysical and techno-economic factors while ensuring that the model is applicable globally.

Machine learning (ML) which is a subset of artificial intelligence (AI) involves the study of statistical computer algorithms that allow computer programs to automatically improve through experience [16,17]. With advantages including ease of trends and pattern identification, minimal human intervention (automation), ability to improve continuously as well as high efficiency in the handling of multi-dimensional and multi-variety data, ML finds wide application in several fields of human endeavours [16,18]. Its application is however limited by factors such as ethics, data availability and quality, misapplication as well as interpretability [19].

With regards to climate change mitigation, ML is expected to play a key role in diverse areas ranging from smart grids to disaster management [20]. This has also been corroborated by Yan et al. [21] who reviewed and highlighted areas of ML deployment and prospects in carbon capture, utilisation and storage (CCUS). Efforts at ML application in the advancement of other GGR methods, though limited, have also been reported in the literature. This includes prediction of biochar yield and carbon content from the pyrolysis of various materials such as lignocellulosic biomass [22], organic waste [23] and plastics [24]. ML has also been applied for reforestation planning [25] and afforestation site selection based on particular local tree species [26,27].

In this study, we apply ML to predict the suitability of deployment of 5 GGR methods (forestation, EW, DACCS, BECCS and biochar) based on a location's biophysical and techno-economic characteristics. The study will involve the application of unsupervised ML for data labelling and supervised ML for predictive analysis. In addition, a case study analysis to test and validate the ML predictive outcome will also be carried out.

## 2 Methods

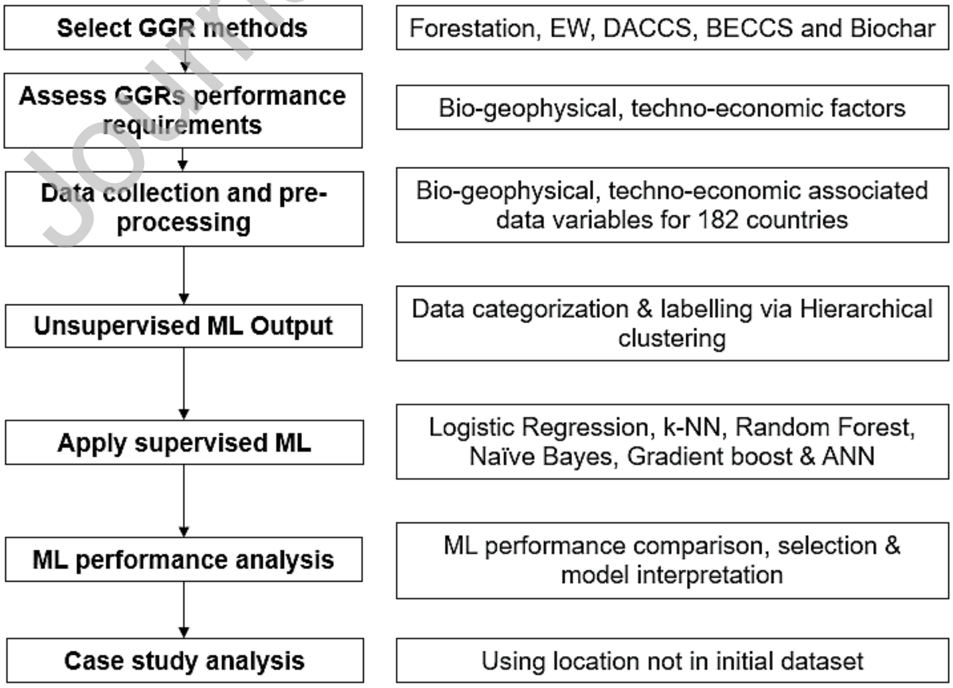

Select GGR methods
Forestation, EW, DACCS, BECCS and Biochar
Assess GGRs performance requirements
Bio-geophysical, techno-economic factors
Data collection and pre-processing
Bio-geophysical, techno-economic associated data variables for 182 countries
Unsupervised ML Output
Data categorization & labelling via Hierarchical clustering
Apply supervised ML
Logistic Regression, k-NN, Random Forest, Naïve Bayes, Gradient boost & ANN
ML performance analysis
ML performance comparison, selection & model interpretation
Case study analysis
Using location not in initial dataset

**Fig. 1**. Method framework

Fig. 1. presents the framework of the method applied in this study. The five GGR methods considered in this study, namely forestation, EW, DACCS, BECCS and biochar, were selected on the basis of their carbon dioxide removal potential and technological readiness levels [4,10]. Major biophysical and techno-economic factors necessary for the sustainable and efficient performance of these GGR methods were identified from the literature [28]. This include climatic factor (temperature and precipitation), land availability, water availability (WA), Low carbon energy availability (LCEA), geological storage potential (GSP), biomass availability (BA) and financial capability represented by the gross national income per capita (GNI).

Associated input variables of these factors which will be applied in the ML modelling were then identified as presented in Table 1. While the climate variable was explicitly indicated for forestation, it wasn't explicitly indicated for BECCS and biochar. This was done to avoid double-counting as its impact on biomass availability is already captured in the indicator applied for assessing biomass availability (the net primary production). However, in the case of biochar, there was need to also assess the impact of weather on the applied biochar. Hence the additional inclusion of climate variables for this GGR method.

**Table 1**. Data variables for the GGR methods

| GGR method | Deployment requirement | |
|---|---|---|
| | Major | Minor |
| Forestation | Climatic factor (average quarterly temperature and precipitation*), Forest land use (FLu) | WA |
| EW | LCEA, GNI | Average quarterly temperature and precipitation*, Agricultural land use (AgLu). |
| DACCS | GSP, LCEA, GNI | WA |
| BECCS | GSP, BA | WA, GNI |
| Biochar | BA | Average quarterly temperature and precipitation*, GNI, AgLu |

*This consists of a total of 8 variables which comprise data for each quarter in a year for temperature (Q1T, Q2T, Q3T, Q4T) and precipitation (Q1P, Q2P, Q3P, Q4P).

## 2.1 Data collection and pre-processing

Most recent data for these identified variables were obtained for 182 countries from diverse sources as follows; for climatic factor, quarterly mean temperature and precipitation covering a 30-year period from 1991 to 2020 sourced from the World Bank climate knowledge portal was used [29]. For water availability, water stress level data soured from the Food and Agriculture Organisation was used [30]. This indicator measures the amount of freshwater used in relation to the total amount of accessible renewable freshwater resources. Geological $CO_2$ storage potential was obtained from the IEAGHG world geological suitability map [31,32]. The share of forest land use and agricultural land use sourced from the CIA's World factbook [33] was applied for assessing land availability. Gross national income per capita (GNI) sourced from the World Bank was adopted to measure financial capability of countries

[34]. For energy availability, data for the percentage share of low carbon energy in each countries' energy mix sourced from the International Energy Agency was adopted [35,36]. Biomass availability data, measured in terms of the net primary production (tC/ha/yr) was sourced from the International Renewable Energy Agency [37]. The net primary production, a measure of the biomass productivity of the various countries was multiplied by their respective forest and agricultural (arable) land use to obtain the current biomass potential (tC/yr). It is important to note that other potential biomass sources such as imports and municipal wastes were not considered on account of data availability.

The raw data was pre-processed into a usable form by the numerical encoding of string type data and analytical determination of incomplete data. Data normalisation was also carried out between 0 and 1 for the entire dataset to eliminate differences in scale/range of values among the variables. In addition to eliminating bias of scale that exist across variables, normalisation also ensured that each country's data was better situated within a global context. This is very important given the global nature of the mitigation situation and the need to comparatively assess the role, scale and potential of nations as well as collaborative opportunities.

The development or training of a ML predictive model or supervised ML requires the availability of labelled dataset (input and corresponding output data). Given that these GGR methods are yet to be deployed on a large scale globally, the obtained locational dataset was unlabelled in regard to the suitability of these countries to deploy GGR methods. In order to develop the ML predictive model, the modelling in this study was thus, done in two stages. Firstly, the obtained pre-processed input dataset was labelled (derive the output – determine the deployment suitability category of the 182 countries based on their respective variables). This was done using the hierarchical clustering technique which is an unsupervised ML method (section 2.2). In the second stage, supervised ML was then applied to train the now labelled dataset comprising of the same input variables and the outputs obtained in stage 1 (section 2.3). This second stage is the focus of this paper. It is aimed at exploring the extent of ML as an innovative approach in assessing the suitability of GGR methods, which is free of human bias.

## 2.2 Hierarchical clustering

The hierarchical clustering technique is a bottom to top unsupervised machine learning method that clusters datasets into a specified number of groups (clusters) based on the similarity of the variables of each data [38,39]. In this case, for each GGR method, the 182 countries considered were clustered into the optimal possible number of clusters ($k$) for the particular dataset. The optimal number of clusters for a particular dataset is the number of clusters that ensures that members of the same cluster are similar to one another and uniquely different from members of other clusters. For each GGR method, the applied ML algorithm (Hierarchical clustering) assessed all the 182 datapoints simultaneously and clusters them based on the similarity of their characteristics. It is important to state that the clustering was wholly done by the ML model based on the data characteristics, and no specific human-defined cut-off levels was set for each variable as threshold for the categorisation. The code was written in Python using the *Agglomerative Clustering* library imported from the *sklearn.cluster* module. The linkage was set to 'ward' while the Euclidean distance method was adopted.

The obtained clusters were then sorted into four levels of deployment suitability (or categories); *highly suitable* (HS), *suitable* (S), *possible* (P) and *unsuitable* (U). A description of these suitability levels is presented in Table 2 with reference to the classification of deployment requirements in Table 1. This sorting of clusters into their respective suitability categories was based on a comparative assessment of their ML-generated cluster characteristics (mean cluster value for each variable) against the requirements highlighted in Tables 1 and 2. The higher the mean cluster value for a variable, the higher the availability of that factor in the countries that make up that cluster. For example, considering BECCS, with GSP and BA as major requirements and WA and GNI as minor requirements (Table 1), the cluster with the highest mean values for  all requirements will be sorted into the Highly Suitable category (Table 2).

**Table 2.** Description of deployment suitability categories

| Suitability category | Description |
|---|---|
| Highly suitable | Satisfies major and minor requirements for deployment |
| Suitable | Satisfies some major and minor requirements for deployment |
| Possible | Satisfies core geo-locational requirement which are not likely to change in a long time (climatic factors and GSP), but currently lacks the highly dynamic requirements such as GNI, LCEA, BA.<br>Given the presence of required geo-locational indicators, the suitability status of nations in this category are expected to improve with improvement in their dynamic indicators. This improvement could be because of factors ranging from economic development to international collaborations. |
| Unsuitable | Satisfies neither major and nor minor requirements for deployment |

## 2.3 Supervised ML modelling

The labelled dataset for each GGR method obtained from the hierarchical clustering consist of the data for the input variables presented in Table 1 and the corresponding output labels (deployment suitability level) for the 182 countries. These datasets were then trained using supervised ML algorithms. The supervised ML, the most commonly applied type of ML involves training the data to learn the relationship between the given inputs and associated output values [40,41]. Since the ML model is expected to predict the deployment suitability category of a given location (a classification-type problem), seven of the most commonly applied ML classification algorithm was employed in this study. These ML algorithms including logistic regression, Decision Trees, k-NN, naïve bayes, Random Forest, gradient boost and ANN, are briefly discussed in Table 3.

**Table 3.** Applied supervised ML algorithms [40,42,43]

| Algorithm name | Description |
|---|---|
| Logistic regression | A classification algorithm that predicts the likelihood of a dependent variable (usually binary) belonging to a category. |

| | |
|---|---|
| Decision trees | This interpretable algorithm performs by splitting values of data features into branches at decision nodes until a final decision output is established. |
| K-Nearest neighbour | KNN is a non-parametric algorithm requiring little to no training that makes its selection based on the proximity to other data points regardless of what feature the numerical values represent. |
| Naïve Bayes | This algorithm is based on the Bayesian theorem which updates the prior knowledge of an event with the independent probability of each feature that can affect the event. |
| Random forest | The algorithm is an ensemble of decision trees characterised by improved accuracy. It operates by generating a multitude of decision trees and uses either the modal vote or average prediction for classification or regression tasks, respectively. |
| Gradient Boost | This ensemble algorithm combines multiple weak algorithms to obtain an improved output. |
| Artificial Neural Network | ANN inspired by the biological neural networks of the human brain, is made up of input, hidden and output layers, as well as a number of parallel-interconnected neurons in each layer. It can be trained to recognise patterns, classify data, and predict future events. |

For all algorithms considered, the dataset was split into training and testing groups in a ratio of 70:30 respectively and randomly. This was done to minimise training and testing bias, obtain best results and ensure that the accuracy of the model is neither overestimated nor underestimated [44,45]. The code was written in Python using the corresponding ML algorithm library imported from their respective *sklearn* module. For each algorithm, the code was run 30 times to minimise bias.

## 3 Results and discussion

### 3.1 Supervised ML modelling

The results of the performance accuracy ($R^2$) of the applied seven ML algorithms in the prediction of the suitability of deployment of the five GGR methods are presented in Fig. 2. The overall performance accuracy of all the ML algorithms for the five GGR methods was high ranging from 0.76 – 1.00. Being a classification kind of problem in which the ML models were expected to predict one out of four categories, the performance accuracy obtained for multiple runs of the various ML algorithms were typically constant with minor variations. Overall, a 99% confidence interval of ±0.01 was evaluated for the performance accuracies. The Random Forest, k-NN and ANN had the highest performance accuracies of 96%, 98% and 100% respectively while the Naïve Bayes had the lowest accuracy (76%). Given the black box nature of ML process, it is difficult to specify the particular reason for the variation in performance of the algorithms across the various GGR methods. This discrepancy in performance by the various ML algorithms is linked to multiple factors ranging from the inherent characteristics of each algorithm (operating principle, strengths and weaknesses) as well as the data characteristics (quantity, shape, dimension and distribution).

The average performance accuracy of all the ML algorithms was highest for DACCS deployment suitability prediction, because it had the least number of input variables.

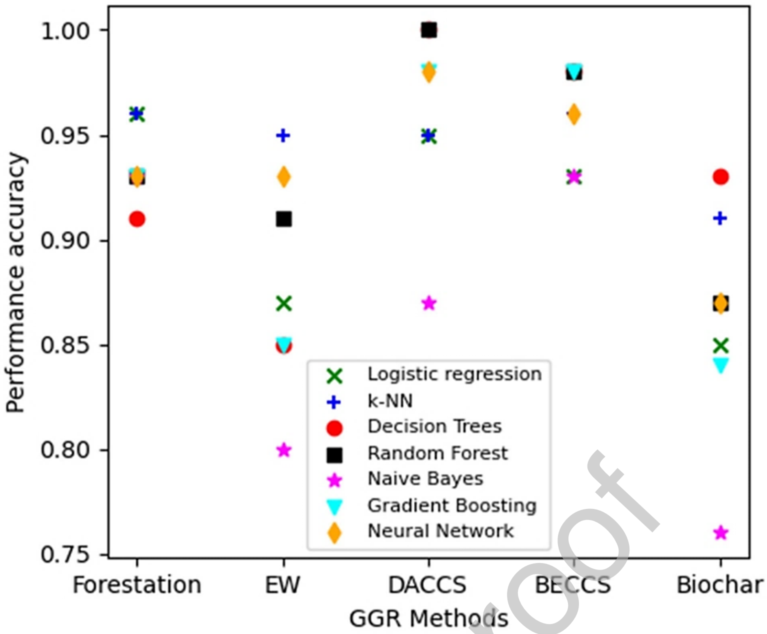


**Fig. 2.** Comparison of ML algorithms performance for the suitability categorisation of 5 GGR methods

### 3.2 Modelling case study – Scotland

In order to validate the performance accuracy of the ML model, data for Scotland, a case study location which was not part of the initial training and testing dataset was used. Data for the input variables were sourced as follows. 10-year mean (2011-2020) weather data (mean quarterly temperature and rainfall) was obtained for Braemar station, Scotland from the Met office historical data [46]. Agricultural land use data were obtained from the Scottish Agricultural Census report [47]. Forest land use data from the Scotland Forestry Strategy report [48]. GDP per capita data was obtained from Scottish Quarterly GDP National Accounts report [49], water availability [30], geological $CO_2$ storage potential [31,32], energy availability [50] and biomass availability [37].

The Random Forest ML model was applied to the obtained Scotland normalised dataset. Fig. 3 presents the model results showing the suitability of Scotland to deploy the five GGR methods. While the ML model predicted forestation as *possible* (P), the other four were found to be *highly suitable* (HS). This ML model outcome is in good agreement with the existing literature reports concerning the probable deployment pathways for GGR methods in Scotland [9,10].

The deployment of forestation in Scotland will involve reforestation which is already underway [10] and afforestation which is expected to deliver the smallest NET contribution (1.24 $MtCO_2$/yr) of all the GGR methods on account of land use constraints [9]. The removal potential of this GGR method is estimated to range from 2.5 – 5.3 $MtCO_2$/yr if the Scottish Government strategy on tree planting is sustained into the 2040s [10]. In the case of EW, the literature indicates high suitability of deployment (HS), which is in total agreement with the ML predicted outcome. This high suitability stems from Scotland's natural abundance of volcanic rocks and alkaline materials. The feasible removal potential of this GGR method is expected to range from 5-10 $MtCO_2/yr$, with the application rate expected to be dependent on the quality of the land [10].

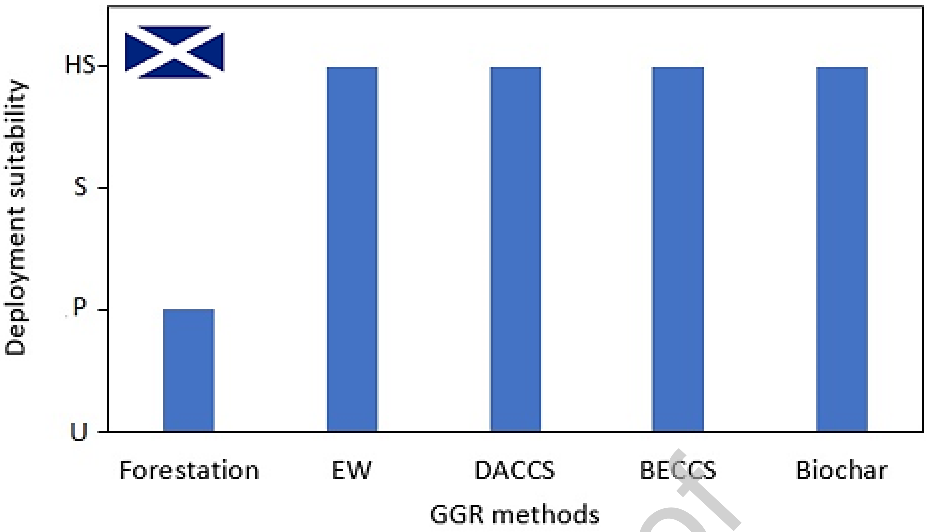


**Fig. 3**. Suitability of Scotland to deploy GGR methods

With a surplus capacity of accessible $CO_2$ storage well beyond its needs, coupled with abundant renewable energy options, Scotland stands tall as a highly suitable location for deploying DACCS [9,10]. This agrees with the obtained ML model output for DACCS. Feasible deployment potential of this GGR method has been estimated to result in removal of 1-10 $MtCO_2/yr$ based on input energy and storage availability [10]. Though specific locations within Scotland are yet to be identified for deploying this technology, the need to consider cost optimal sites that will minimise $CO_2$ transport and energy costs cannot be overemphasised.

The negative emissions potential for BECCS and biochar if implemented on the 0.52 Mha of highly suitable land in Scotland, has been estimated to range from 5.7 – 23 $MtCO_2/yr$ and 2.2 - 14.3 MtCO2 per year respectively [9]. With an initial BECCS deployment potential estimate of 3.59 $MtCO_2/yr$ using existing bioenergy installations with no additional land requirements [51] coupled with a projected 80% increase in available biomass by 2030 [52], Scotland qualifies as a highly suitable location for the deployment of these GGR methods. This suitability categorisation agrees with the obtained ML model output detailed in Fig. 3.

Overall, with Scotland showing significant potential as a very suitable location for deploying these five GGR methods based on the ML model output and the existing literature reports, a major challenge in the particular choice of GGR methods to be eventually developed and deployed still exists. While Alcalde et al. [9] reported that a combination of the maximum potential of BECCS, soil carbon sequestration (SCS) and EW coupled with a small contribution from DACCS could abate over 100% of Scotland's annual $CO_2$ emissions, Haszeldine et al. [10] highlighted the need to ensure that chosen GGR methods are deployable at acceptable cost with minimal disruption while enhancing the development of new skills, jobs and value in Scottish supply chains.

### 3.3 Model interpretation and limitation

Model interpretation was carried out to understand the logic employed by the ML model in its predictive decision process. This understanding assists in verifying the correctness of the model regarding to how well it correlates with literature expectations on the relationship between input variables (bio-geophysical and technoeconomic factors) and outputs (GGR deployment suitability). In order to do this, the impact of the associated ML input variables (represented by the feature importance score) of the five GGR methods on their respective suitability of deployment (output) was evaluated. This was done using the *feature_importances_* tool in the Random Forest library imported from the *sklearn* module. The feature importance score is a numerical representation of how much each input variable contributes or impacts on the ML model output. The sum of the importance scores of all the input variables in a model is usually equal to 1. The obtained result of this analysis showing the impact of the ML input variables on the deployment suitability of the GGR methods is presented in Fig. 4.

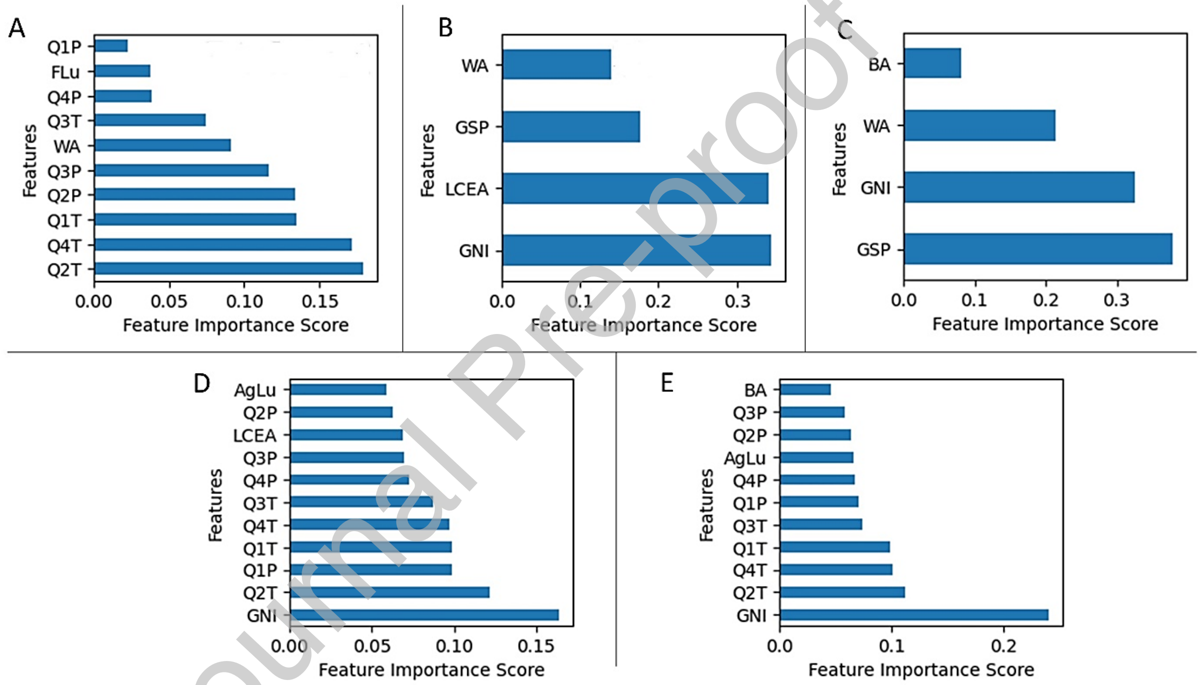


**Fig. 4.** Impact of the ML input variables on the deployment suitability of the GGR methods. A) Forestation, B) DACCS, C) BECCS, D) EW and E) Biochar.

For forestation (Fig. 4A), climatic factors (quarterly temperature and precipitation) had the highest feature importance score. Thus, indicating that these variables have the highest impact in the ML model's decision of determining the suitability of deploying this GGR method. This highly correlates with the work of Asibor et al. [28] that ranks climatic factors as the highest impacting variable for forestation deployment. The tropical region which is characterised by high temperatures and precipitation has been generally identified as the most suitable location for forestation [53,54]. The relative importance of land availability (FLu) on the other hand, has been underestimated by the ML model in its decision making. This outcome poorly agrees with the literature, which generally identifies land availability as a major and the next important factor (after climate) for the deployment of this GGR method [4,5].

In the case of DACCS (Fig. 4B), the obtained feature importance score indicates that LCEA and GNI had the highest impact in the ML modelling outcome. Though these two factors are no doubt major requirements for deployment of DACCS, this importance ranking is however not in agreement with literature expectations. This is because geological storage potential (GSP) has been generally identified as the most important factor for DACCS deployment [28,55,56]. The ML prediction for DACCS deployment is thus, majorly influenced by the financial capability and availability of low-carbon energy in a location with relatively lesser consideration given to the (GSP). On the other hand, water availability (WA) requirement had the lowest feature score which correlates with the literature as this requirement is dependent on the type of DAC technology employed.

For EW (Fig. 4D), similar outcome of a high correlation between the ML variable impact ranking and literature expectations was also obtained for GNI and AgLu. This is because of the high capital requirement and low land requirement associated with the deployment of this GGR method [8,57,58]. Given that EW is also energy intensive, the impact of LCEA was however underestimated by the ML model.

The importance of the availability of biomass in the deployment of BECCS (Fig. 4C) and biochar (Fig. 4E) was largely underestimated by the ML model and does not correlate with literature expectations [3,4]. These outcomes can be linked to the characteristics of the available data such as the density or distribution of the input data. Variables with dense data distribution (fewer differences in data such as the GNI of countries) were more dominant than variables with sparse distribution, such as the biomass availability. The ML model, however, has a fairly good correlation for the impact of water availability, geological storage potential and climatic condition (in the case of biochar).

While the performance accuracy of the ML model is high, coupled with a reasonable correlation with literature expectations on the impact of its input variables, its overall performance is however limited by quantitative and qualitative factors associated with the available data. This is especially so given that these GGR methods are at various levels of technological readiness [5,10] and are mostly yet to be deployed at a large scale. Thus, it is expected that the large-scale deployment of these technologies in the near future will enhance the availability of labelled data. This will eliminate the need to apply unsupervised ML (hierarchical clustering algorithm) and its attendant data-density dependency limitations in labelling the data.

Another aspect closely linked to the limitation of data availability is the use of single data point (national data) to represent the characteristics of countries. While this is a valid approach for country-level assessments, it however fails to capture the spatial variations in characteristics that exists, especially in large countries. While it is recommended that multiple data points at regional or local levels be explored when globally available, it is important to add that this factor does not impact on the correctness of the categorisation. This is because the concept of suitability in this study does not indicate that every part of a country (irrespective of size) is suitable or not. Instead, it indicates the relative possibilities/ potentials of countries (or a location) to deploy the GGR technologies with respect to other countries (or locations) within a global context, based on their characteristics.

## 4 Conclusions

In this study, ML has been applied in determining the suitability of a location to deploy five GGR methods within a global context. The use of ML minimises human bias and ensures that the outcomes are based on applicable data. The GGR methods considered are forestation, EW, DACCS, BECCS and biochar. An unsupervised ML (hierarchical clustering) technique was applied to label the dataset consisting of bio-geophysical and technoeconomic variables for 182 countries. Seven supervised ML algorithms were applied in training and testing the labelled dataset with the Artificial Neural Network (ANN), k-Nearest neighbour (k-NN) and Random Forest algorithms having the highest performance accuracy. The performance accuracy of the ML model was validated using a case study assessment of Scotland's suitability to deploy these GGR methods. While the performance accuracy of the ML models was typically high (76 – 100%), an assessment of its decision-making logic (model interpretation) revealed some limitations regarding the impact of the various input variables on the outputs. This is in addition to the use of single data point for country-level input data, which does not fully represent the spatial variations in characteristics that exists, especially in large countries. The availability of labelled data brought about by expected large-scale deployment of these GGR technologies in the near future will greatly enhance the performance and role of ML in achieving the 1.5 °C warming target.

**Acknowledgement**

Jude O. Asibor is grateful to the Petroleum Technology Development Fund (PTDF), Nigeria, for doctoral study scholarship, award number: PTDF/ED/OSS/PHD/JOA/077/19 and the University of Benin, Benin City, Nigeria.

**CRediT authorship contribution statement**

**J.O. Asibor:** Conceptualisation, Methodology, Software, Formal Analysis, Investigation, Writing - Original Draft. **P.T. Clough:** Conceptualisation, Writing - Review & Editing, Supervision, Project administration. **S.A. Nabavi:** Conceptualisation, Writing - Review & Editing, Supervision. **V. Manovic:** Conceptualisation, Writing - Review & Editing, Supervision.

**Declaration of competing interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Data availability statement**

Data and codes underlying this study can be accessed through the Cranfield University repository (CORD) at: https://figshare.com/s/32e11d6ce96378bc45c1. Data are available under the terms of the [Creative Commons Attribution 4.0 International (CC BY 4.0)].

## References

[1] IPCC. Summary for Policymakers. In: Shukla PR, Skea J, Slade R, Al Khourdajie A, van Diemen R, McCollum D, et al., editors. Clim. Chang. 2022 Mitig. Clim. Chang., Cambridge, UK: Cambridge University Press; 2022, p. 1–30.

[2] Rogelj J, Popp A, Calvin K V., Luderer G, Emmerling J, Gernaat D, et al. Scenarios

towards limiting global mean temperature increase below 1.5 °c. Nat Clim Chang 2018;8:325–32. https://doi.org/10.1038/s41558-018-0091-3.

[3] Smith P. Soil carbon sequestration and biochar as negative emission technologies. Glob Chang Biol 2016;22:1315–24. https://doi.org/10.1111/gcb.13178.

[4] Fuss S, Lamb WF, Callaghan MW, Hilaire J, Creutzig F, Amann T, et al. Negative emissions - Part 2: Costs, potentials and side effects. Environ Res Lett 2018;13. https://doi.org/10.1088/1748-9326/aabf9f.

[5] The Royal Society. Greenhouse Gas Removal. 2018.

[6] Nemet GF, Callaghan MW, Creutzig F, Fuss S, Hartmann J, Hilaire J, et al. Negative emissions - Part 3: Innovation and upscaling. Environ Res Lett 2018;13. https://doi.org/10.1088/1748-9326/aabff4.

[7] Minx JC, Lamb WF, Callaghan MW, Fuss S, Hilaire J, Creutzig F, et al. Negative emissions - Part 1: Research landscape and synthesis. Environ Res Lett 2018;13. https://doi.org/10.1088/1748-9326/aabf9b.

[8] Smith P, Davis SJ, Creutzig F, Fuss S, Minx J, Gabrielle B, et al. Biophysical and economic limits to negative CO2 emissions. Nat Clim Chang 2015;6:42–50. https://doi.org/10.1038/nclimate2870.

[9] Alcalde J, Smith P, Haszeldine RS, Bond CE. The potential for implementation of Negative Emission Technologies in Scotland. Int J Greenh Gas Control 2018;76:85–91. https://doi.org/10.1016/j.ijggc.2018.06.021.

[10] Haszeldine S, Cavanagh A, Scott V, Sohi S, Mašek O, Renforth P. Greenhouse Gas Removal Technologies-approaches and implementation pathways in Scotland Executive summary. ClimateXChange 2019.

[11] Mulligan J, Rudee A, Lebling K, Levin K, Anderson J, Christensen B. CarbonShot: Federal Policy Options for Carbon Removal in the United States. 2020.

[12] IPCC. Global warming of 1.5°C An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change,. 2018.

[13] Förster J, Beck S, Borchers M, Gawel E, Korte K, Markus T, et al. Framework for Assessing the Feasibility of Carbon Dioxide Removal Options Within the National Context of Germany. Front Clim 2022;4. https://doi.org/10.3389/fclim.2022.758628.

[14] Fajardy M, Mac Dowell N. Can BECCS deliver sustainable and resource efficient negative emissions? Energy Environ Sci 2017;10:1389–426. https://doi.org/10.1039/c7ee00465f.

[15] NEGEM Project. A software tool to generate Negative Emissions deployment pathways. NEGEM 2022. https://www.negemproject.eu/news/a-software-tool-to-generate-negative-emissions-deployment-pathways/ (accessed October 26, 2022).

[16] Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science (80- ) 2015;349:255–60.

[17] Garbade MJ. Clearing the Confusion: AI vs Machine Learning vs Deep Learning

Differences. Towar Data Sci 2018. https://towardsdatascience.com/clearing-the-confusion-ai-vs-machine-learning-vs-deep-learning-differences-fce69b21d5eb (accessed February 8, 2021).

[18] Kumar S. Advantages and Disadvantages of Artificial Intelligence. Towar Data Sci 2019. https://towardsdatascience.com/advantages-and-disadvantages-of-artificial-intelligence-182a5ef6588c (accessed February 8, 2021).

[19] Stewart M. The Limitations of Machine Learning. Towar Data Sci 2019. https://towardsdatascience.com/the-limitations-of-machine-learning-a00e0c3040c6 (accessed February 8, 2021).

[20] Rolnick D, Donti PL, Kaack LH, Kochanski K, Lacoste A, Sankaran K, et al. Tackling Climate Change with Machine Learning. ACM Comput Surv 2022;55. https://doi.org/10.1145/3485128.

[21] Yan Y, Borhani TN, Subraveti SG, Pai KN, Prasad V, Rajendran A, et al. Harnessing the power of machine learning for carbon capture, utilisation, and storage (CCUS)-a state-of-the-art review. Energy Environ Sci 2021;14:6122–57. https://doi.org/10.1039/d1ee02395k.

[22] Zhu X, Li Y, Wang X. Machine learning prediction of biochar yield and carbon contents in biochar based on biomass characteristics and pyrolysis conditions. Bioresour Technol 2019;288:121527. https://doi.org/10.1016/j.biortech.2019.121527.

[23] Li Y, Gupta R, You S. Machine learning assisted prediction of biochar yield and composition via pyrolysis of biomass. Bioresour Technol 2022;359:127511. https://doi.org/10.1016/j.biortech.2022.127511.

[24] Alabdrabalnabi A, Gautam R, Mani Sarathy S. Machine learning to predict biochar and bio-oil yields from co-pyrolysis of biomass and plastics. Fuel 2022;328:125303. https://doi.org/10.1016/j.fuel.2022.125303.

[25] Ordóñez Galán C, Matías JM, Rivas T, Bastante FG. Reforestation planning using Bayesian networks. Environ Model Softw 2009;24:1285–92. https://doi.org/10.1016/j.envsoft.2009.05.009.

[26] Chen Y, Wu B, Chen D, Qi Y. Using machine learning to assess site suitability for afforestation with particular species. Forests 2019;10:1–22. https://doi.org/10.3390/f10090739.

[27] Yousefi S, Avand M, Yariyan P, Jahanbazi Goujani H, Costache R, Tavangar S, et al. Identification of the most suitable afforestation sites by Juniperus excels specie using machine learning models: Firuzkuh semi-arid region, Iran. Ecol Inform 2021;65:101427. https://doi.org/10.1016/j.ecoinf.2021.101427.

[28] Asibor JO, Clough PT, Nabavi SA, Manovic V. Assessment of optimal conditions for the performance of greenhouse gas removal methods. J Environ Manage 2021;294:113039. https://doi.org/10.1016/j.jenvman.2021.113039.

[29] World Bank. World Bank Climate Change Knowledge Portal 2022. https://climateknowledgeportal.worldbank.org/ (accessed July 21, 2020).

[30] FAO. FAO-AQUASTAT database 2022. http://www.fao.org/nr/water/aquastat/data/query/results.html (accessed July 15, 2020).

[31] IEAGHG. Global Storage Resources Gap Analysis for Policy Makers 2011.

[32] Kearns J, Teletzke G, Palmer J, Thomann H, Kheshgi H, Chen H, et al. Developing a Consistent Database for Regional Geologic CO2 Storage Capacity Worldwide. Energy Procedia 2017;114:4697–709. https://doi.org/10.1016/j.egypro.2017.03.1603.

[33] CIA. The World Factbook 2022 2022. https://www.cia.gov/library/publications/resources/the-world-factbook/index.html (accessed June 10, 2020).

[34] World Bank. GNI per capita, Atlas method (current US$) | Data 2021. https://data.worldbank.org/indicator/NY.GNP.PCAP.CD (accessed October 9, 2021).

[35] IEA. IEA Energy Atlas 2020. http://energyatlas.iea.org/#!/tellmap/-1118783123/1 (accessed September 15, 2020).

[36] Ritchie H, Roser M, Rosado P. Renewable Energy - Our World in Data 2020. https://ourworldindata.org/renewable-energy (accessed June 14, 2022).

[37] IRENA. Statistical Profiles 2021. https://www.irena.org/Statistics/Statistical-Profiles (accessed December 10, 2020).

[38] Murtagh F, Contreras P. Algorithms for hierarchical clustering : an overview. Wiley Interdiscip Rev Data Min Knowl Discov 2012;2:86–97. https://doi.org/10.1002/widm.53.

[39] Usama M, Qadir J, Raza A, Arif H, Yau KA, Elkhatib Y, et al. Unsupervised Machine Learning for Networking : Techniques , Applications and Research Challenges. IEEE Access 2019;7:1–37.

[40] Singh A, Thakur N, Sharma A. A review of supervised machine learning algorithms. 2016 3rd Int. Conf. Comput. Sustain. Glob. Dev., Bharati Vidyapeeth, New Delhi as the Organizer of INDIACom - 2016; 2016, p. 1310–5.

[41] Wilson A. A Brief Introduction to Supervised Learning. Towar Data Sci 2019. https://towardsdatascience.com/a-brief-introduction-to-supervised-learning-54a3e3932590 (accessed May 3, 2021).

[42] Pareek P. Machine learning Algorithms and where they are used? Samurai's Sakura 2019. https://techsavvypriya.wordpress.com/2019/12/19/machine-learning-algorithms-and-where-they-are-used/ (accessed February 8, 2021).

[43] Dey A. Machine Learning Algorithms: A Review. Int J Comput Sci Inf Technol 2016;7:1174–9.

[44] Güleç F, Pekaslan D, Williams O, Lester E. Predictability of higher heating value of biomass feedstocks via proximate and ultimate analyses – A comprehensive study of artificial neural network applications. Fuel 2022;320:123944. https://doi.org/10.1016/j.fuel.2022.123944.

[45] Gholamy A, Kreinovich V, Kosheleva O. Why 70/30 or 80/20 Relation Between Training and Testing Sets : A Pedagogical Explanation. Dep Tech Reports 2018:1–6.

[46] Met Office. Climate data for Braemar station. Met Off 2020. https://www.metoffice.gov.uk/pub/data/weather/uk/climate/stationdata/braemardata.txt (accessed May 3, 2021).

[47] Scottish Government. Scottish Agricultural Census 2020. https://www.gov.scot/collections/june-scottish-agricultural-census/ (accessed May 3, 2021).

[48] Scottish Government. Scotland’s Forestry Strategy 2019–2029 2019. https://www.gov.scot/publications/scotlands-forestry-strategy-20192029/pages/4/ (accessed May 3, 2021).

[49] Scottish Government. GDP Quarterly National Accounts for Scotland: 2019 Q4 2020. https://www.gov.scot/publications/gdp-quarterly-national-accounts-for-scotland-2019-q4/ (accessed May 4, 2021).

[50] BEIS. Electricity generation and supply in Scotland, Wales, Northern Ireland and England, 2016 to 2020. BEIS Energy Trends 2021. https://www.gov.uk/government/statistics/energy-trends-december-2021-special-feature-article-electricity-generation-and-supply-in-scotland-wales-northern-ireland-and-england-2016-to-20 (accessed June 30, 2022).

[51] Brownsort P. Negative Emission Technology in Scotland: carbon capture and storage for biogenic CO2 emissions. Scottish Carbon Capture & Storage 2018;44:1–54.

[52] Ricardo. The potential contribution of bioenergy to Scotland’s energy system. ClimateXChange 2019.

[53] Favero A, Sohngen B, Huang Y, Jin Y. Global cost estimates of forest climate mitigation with albedo: A new integrative policy approach. Environ Res Lett 2018;13. https://doi.org/10.1088/1748-9326/aaeaa2.

[54] Raihan A, Begum RA, Mohd Said MN, Abdullah SMS. A Review of Emission Reduction Potential and Cost Savings through Forest Carbon Sequestration. Asian J Water, Environ Pollut 2019;16:1–7. https://doi.org/10.3233/AJW190027.

[55] Viebahn P, Scholz A, Zelt O. German Energy Research Program — Results of a Multi-Dimensional Analysis. Energies 2019.

[56] Fasihi M, Efimova O, Breyer C. Techno-economic assessment of CO 2 direct air capture plants. J Clean Prod 2019;224:957–80. https://doi.org/10.1016/j.jclepro.2019.03.086.

[57] Renforth P. The potential of enhanced weathering in the UK. Int J Greenh Gas Control 2012;10:229–43. https://doi.org/10.1016/j.ijggc.2012.06.011.

[58] Strefler J, Amann T, Bauer N, Kriegler E, Hartmann J. Potential and costs of carbon dioxide removal by enhanced weathering of rocks. Environ Res Lett 2018;13. https://doi.org/10.1088/1748-9326/aaa9c4.

**Declaration of interests**

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

**Dr Peter Clough**
**Senior Lecturer in Energy Engineering**
Building 52
Cranfield University
Cranfield, UK
MK43 0AL

E: p.t.clough@cranfield.ac.uk
T: +44 (0)1234 75 4873
www.cranfield.ac.uk

7 November 2022

***Energy and Climate Change***

Dear Editorial Board,

We are pleased to submit a manuscript entitled "**A Machine Learning Approach for Resource Mapping Analysis of Greenhouse Gas Removal Technologies"**, which we wish to be considered for publication in **Energy and Climate Change**.

In meeting the warming target, the need to include greenhouse gas removal (GGR) technologies in addition to the emission reduction efforts cannot be overemphasised. Given the need for urgency and with the performance of these technologies greatly dependent on bio-geophysical and techno-economic factors amongst others, there is great need to identify suitable locations where these GGR technologies can be sustainably and efficiently deployed.

In this study, Machine Learning (ML) was applied to investigate the suitability of a location to deploy five greenhouse gas removal (GGR) methods based on a location's bio-geophysical and techno-economic characteristics. The GGR methods considered are forestation, enhanced weathering (EW), direct air carbon capture and storage (DACCS), bioenergy with carbon capture and storage (BECCS) and biochar. An unsupervised ML (hierarchical clustering) technique was applied to label the dataset. Seven supervised ML algorithms were applied in training and testing the labelled dataset. The performance accuracy of the ML models was typically high ranging from 76 – 100%. This was validated using a case study assessment of Scotland's suitability to deploy these GGR methods. Model interpretation to assess the decision-making logic of the ML model was also carried out.

This novel application of Machine Learning is significant given that many countries are yet to assess their national potentials for GGR deployment and prioritise its inclusion in their nationally determined contributions (NDCs). The ML model thus provides a means for objective feasibility assessment of GGR technologies that is wholly based on the available data while eliminating the likelihood of human bias.

Given that this study falls within the area of climate change mitigation using negative emission technologies or GGRs, we believe that our manuscript is well suited to Energy and Climate Change journal and would be of great interest to your esteemed readers.

The authors note that none of the material in the paper has been published or is under consideration for publication elsewhere.

The authors have no conflict of interest to disclose.

Thank you for your consideration.

Yours sincerely,

**Dr Peter Clough**

**Senior Lecturer in Energy Engineering**