A Deeper Look: The development of global peat depth datasets and subsequent carbon stock estimates

by

Jade Erin Skye
B.Sc., University of Victoria, 2022

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the School of Earth and Ocean Sciences, University of Victoria

We acknowledge and respect the lək̓ʷəŋən (Songhees and Xʷsepsəm/Esquimalt) Peoples on whose territory the university stands, and the lək̓ʷəŋən and W̱SÁNEĆ Peoples whose historical relationships with the land continue to this day.

A Deeper Look: The development of global peat depth datasets and subsequent carbon stock estimates

by

Jade Erin Skye
B.Sc., University of Victoria, 2022

Supervisory Committee

Dr. J.R. Melton, Supervisor
Environment and Climate Change Canada

Dr. C. Goldblatt, Supervisor
School of Earth and Ocean Sciences, University of Victoria

Dr. M. Costa, Committee Member
Faculty of Geography, University of Victoria

Dr. M. Garneau, Committee Member
Department of Geography, Université du Québec à Montréal

ABSTRACT

Peatlands are important carbon stores which are being destabilised by anthropogenic activity and are sensitive to climate change. To faithfully assess the carbon stored in peatlands and to model their responses to future climate scenarios, it is essential to have accurate information on peat depth. Presently, however, observations of peat depth are insufficient for conducting these tasks at the global scale. Thus, the goal of my thesis is to accurately generate a global distribution of peatland depth and use that distribution to estimate how much carbon is stored within them. The first step was to create Peat-DBase, the largest database of harmonised peat depth measurements at the global scale. Peat-DBase was then used as the basis of training and testing data for PeatDepth-ML, a machine learning-based modelling framework designed to predict peat depths globally. I created PeatDepth-ML by adapting an existing modelling framework that was designed to predict peatland spatial extents by including new datasets of environmental variables that may drive or indicate peat formation, updating the cross-validation procedures used for model testing, and adding a custom scoring metric to the model to assist in predicting deeper peat depths. I then used PeatDepth-ML to produce a spatially continuous global map of peatland depths. Inspection of Peat-DBase revealed regional data gaps, such as in the Tropics, and potential sampling biases in peat depth measurements, e.g. the collecting of a single peat core to represent the depth of an entire peatland wherein depth could be varying significantly or the presence of multiple peat cores with highly varying depths over small spatial scales. The impact of Peat-DBases's regional biases on PeatDepth-ML's predictions was assessed by calculating a metric describing the predictions area of applicability. To test the sensitivity of PeatDepth-ML to some aspects of sampling bias, a bootstrapping method was developed to create multiple training datasets from Peat-DBase. Running PeatDepth-ML on the bootstrapped datasets showed that model behaviour could vary significantly in response to changes in the training data, particularly at the regional scale. When compared to other estimates in the literature, PeatDepth-ML achieved a similar or improved level of performance and is of better overall quality because of its global reach and continuous representation of peat and non-peat regions without the use of an independent peatland extent map. However, PeatDepth-ML demonstrated a tendency to predict towards the mean peat depth of its training data, which was relatively shallow possibly due to the inclusion of non-peat data, which was included to allow the model to predict over all regions.

Performing simple carbon stock calculations using PeatDepth-ML's results produced estimates that are in line with those previously published. Collectively, Peat-DBase and PeatDepth-ML are cohesive global datasets of peat depth that can aid future peatland research and policy endeavors.

Contents

List of Tables

List of Figures

List of Acronyms

AOA	area of applicability
BHO	Bayesian hyper-parameter optimization
BLOOCV	blocked-leave-one-out cross-validation
C	carbon
CSV	comma-separated values
DJF	December–February
GEE	Google Earth Engine
JJA	June–August
MAM	March–May
MBE	mean bias error
MI	Moran's I
ML	machine learning
NME	normalised mean error
PD-ML	PeatDepth-ML
RF	random forest
RFE	recursive feature elimination
RMSE	root mean square error
SON	September–November
SWE	snow water equivalent
SWIR3	short wavelength infrared (2225-2275 nm)
TSV	tab-separated values
VIF	variance inflation factor

VPD	vapour pressure deficit
WoSIS	World Soil Information Service

ACKNOWLEDGEMENTS

The data assembled for this project touches many lands with which local and indigenous communities have deep connections. I am grateful to these many peoples for their stewardship of the environments we have come to call peatlands.

I would like to thank:

My supervisors and committee members, Joe Melton, Colin Goldblatt, Michelle Garneau, and Maycira Costa for their mentoring, support, encouragement, and patience.

Environment and Climate Change Canada, for funding this project.

My family and friends, Jessica, Kristine, Karlee, Jude, Char, Wyatt, Nathan, and Brad for supporting me in the low moments.

In the darkest times, hope is something you give yourself. That is the meaning of inner strength.

Uncle Iroh, *Avatar: The Last Airbender*

DEDICATION

This work is dedicated to students everywhere. I believe in us!

Chapter 1

Introduction

Peatlands are significant terrestrial carbon (C) stores. They cover nearly 3% of Earth's land area (Xu et al., 2018; Melton et al., 2022) and store approximately one third of global soil C (Environment, 2022; Ofiti et al., 2023; Joosten and Clarke, 2002; Turunen et al., 2002; Ruppel et al., 2013). Peatlands are a type of wetland and can be found in all but the driest landscapes of the world (Joosten and Clarke, 2002; Koster and Favier, 2005). However, they are primarily concentrated in the boreal, temperate, and tropical biomes as seen in Figure 1.1 (Melton et al., 2022). In addition to C sequestration and subsequent climate regulation, peatlands also fulfill many other important ecological roles, including water retention and regulation, acting as a buffer between saltwater and freshwater systems, and providing a habitat for unique plant and animal species (Ribeiro et al., 2021; Ratnayake, 2020).

Peatlands have historically been C sinks, but are at increasing risk of becoming sources instead. Waterlogging within a peatland, which has resulted from impeded drainage, allows for C sequestration as once oxygen is depleted within the high water table by aerobic decay, an anaerobic environment results and significantly limits further decomposition of submerged organic material deposited by local vegetation. The partially decayed vegetation can then accumulate as peat on timescales of hundreds to thousands of years (Koster and Favier, 2005; Zinck, 2011). Large areas of peatlands are being degraded through human land use change, which results in the emission of sequestered C (Joosten and Clarke, 2002; Koster and Favier, 2005; Ratnayake, 2020). Peatlands are drained for forestry, agriculture, and historically for use as fuel (Fluet-Chouinard et al., 2023). Drainage lowers the water table in a peatland, causing it to dry. The slower anaerobic decomposition of organic matter within waterlogged peat is then replaced by faster aerobic decomposition as drying occurs (Page and Baird,

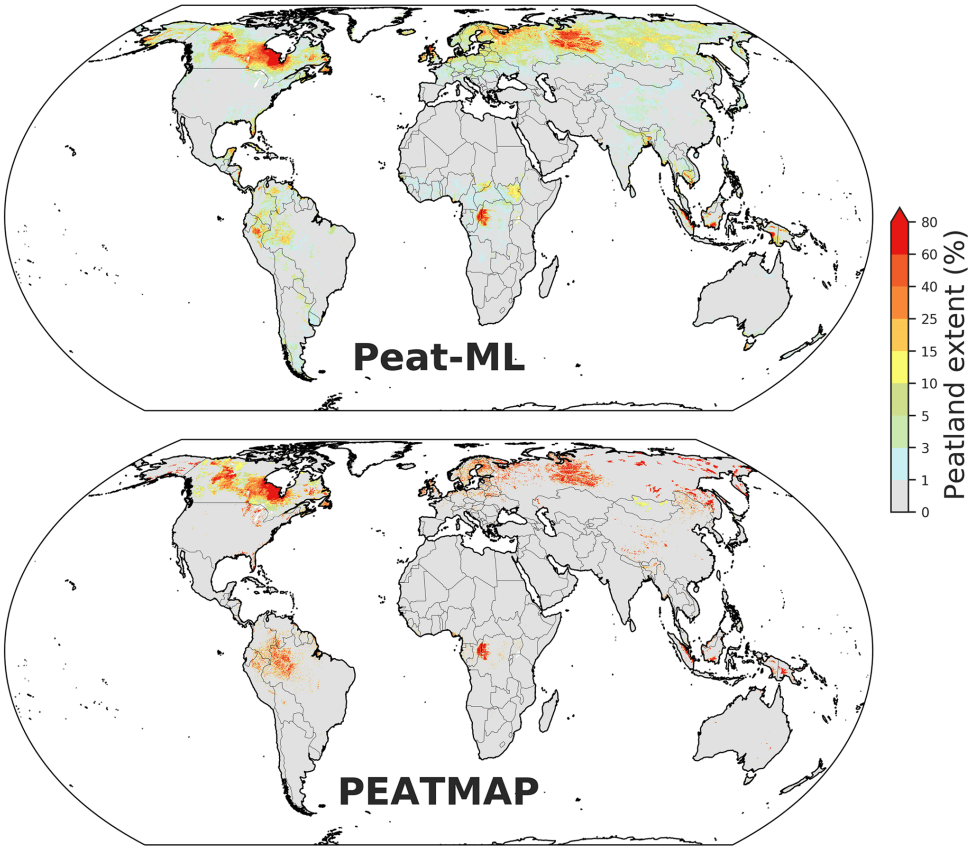

Figure 1.1: Estimates of global peatland fractional coverage as predicted by Peat-ML and PEATMAP from Melton et al. (2022) and Xu et al. (2018).

2016; Turetsky and St. Louis, 2006; Koster and Favier, 2005). Additionally, peatlands may dry out under the warmer and lower soil moisture conditions resulting from climate change in boreal regions. Increased wildfire activity in peatlands is also a risk under these hotter and drier conditions (Canadell et al., 2021; Helbig et al., 2020). Peatlands, which are permafrost-affected, can also degrade due to global warming induced permafrost thaw (Hugelius et al., 2020).

While it is known that peatlands store a significant amount of C, C stock estimates remain highly variable. Recent estimates of peatland C stock values range from 113 to 1029 Pg C (Minasny et al., 2019; Widyastuti et al., 2024). Minasny et al. (2019) describe two methods of calculating the C stock of peat. The first method is based on the age of a peatland and the rate of C accumulation across the peatland area. The second method uses the organic C content, bulk density, depth, and area of a peatland (e.g. Hugelius et al. (2020)). In each of these approaches, observational data on peatlands from field surveys and lab work (e.g. radiocarbon dating for peat age) is required. However, such observations are not always available, particularly at larger regional and global scales. Additionally, previous global peatland C stock estimates have been calculated using different peat definitions and assumed values for one or more of the input variables (e.g. using an average peat depth or organic C content value) and the results can therefore vary depending on these assumptions (Minasny et al., 2019).

There is a growing interest in modelling peatland processes. Land surface models can include terrestrial elements of the C cycle, of which peatlands are a part, and act as components to broader Earth system models and their climate simulations. Land surface models emulate C, water, and energy fluxes of the land surface and can estimate the value of different geophysical variables at different times and locations with varying degrees of certainty (Heyvaert et al., 2024). The C cycle and climate are closely tied through short and long term feedback loops (Canadell et al., 2021). As part of the land surface, peatlands have acted as C sinks and stores over hundreds to thousands of years, thereby having a cooling influence on the climate. However the future climate impact of peatlands under a warming climate and anthropogenic disturbance is uncertain (Canadell et al., 2021; Helbig et al., 2020; Bechtold et al., 2019). By including the C, water, and energy dynamics of peatlands in land surface models, and Earth system models by extension, we can learn more about their past and future effects on the climate through the influence of these added fluxes in climate simulations (Bechtold et al., 2019; Wu et al., 2016; Chadburn et al., 2022).

Current peatland renditions in land surface models can face a variety of limitations. Many of the model frameworks that include peatlands are designed to operate at site-specific or regional scales and over short time spans (Bechtold et al., 2019). Meanwhile, other larger scale land surface model peatland implementations do not yet support dynamic parameterisation of peatland properties in response to changes within the peatland (e.g. as more peat decomposes, its hydrology should change from what it was initialised as) (Chadburn et al., 2022; Bechtold et al., 2019; Müller and Joos, 2020). Site level modelling often uses existing observational data from the peatland at the given location for model input or tuning. However these approaches were not easily implemented globally as such datasets were not previously available at a global scale and sufficient resolution (Bechtold et al., 2019). Regional and global land surface models use a grid-based representation of the land surface, therefore any peatland initialisation data is needed on the same model grids (Melton et al., 2022; Bechtold et al., 2019; Chadburn et al., 2022). For example, when including peatland processes, the Canadian Land Surface Scheme Including Biogeochemical Cycles (CLASSIC) takes peat depth and fractional coverage within each of their grid cells as input variables, among other data (Wu et al., 2016; Melton et al., 2020; Seiler et al., 2020).

Peatlands have historically been understudied, and thus underrepresented, in landscape inventories at a global scale (Krankina et al., 2008; Hugelius et al., 2020; Melton et al., 2022). Traditional approaches to mapping peatland depth involve field surveys or proximal sensing techniques (Minasny et al., 2019; Jowsey, 1966). Field surveys typically involve going to a known or suspected peatland and collecting peat soil cores or probing with metal poles (Crezee et al., 2022; Lähteenoja et al., 2012). The cores are analysed to establish the depth at which the peat transitions to a mineral layer, with this bottom layer of peat often being referred to as basal peat (Lähteenoja et al., 2012; Ruwaimana et al., 2020; Quik et al., 2022). Peat depth can also be sensed with proximal geophysical processes such as ground penetrating radar, electrical resistivity imaging, and electromagnetic induction. Such proximal sensors work near to the ground and measure the electrical properties of a peatland to produce 2-D depth surveys along transects, or complete 3-D surveys in some cases. Although they may cover a more significant area faster than coring, researcher presence in the peatland is still required, and some of these procedures need calibration based on field survey data from the site for the best results (Minasny et al., 2019; Altdorff et al., 2016; Comas et al., 2015; Rosa et al., 2009). Additionally, these proximal methods

are inferring peat depth from the relationship between the electrical properties they are based on and the field survey data they are calibrated to, rather than directly observing peat depth as you would with coring (Minasny et al., 2019). Both proximal geophysical processes and field survey methods become less practical as the study area increases in size (Gatis et al., 2019). Peatlands are often underexplored since conducting such surveys is challenging in the waterlogged and often remote conditions of these environments, as well as being potentially very costly (Minasny et al., 2019; Rudiyanto et al., 2016). Tropical peatlands are often more poorly mapped than boreal and temperate peatlands (Page et al., 2007; Zinck, 2011; Ruwaimana et al., 2020).

In response to the difficulty in mapping peatlands and the subsequently limited amount of observational data, digital soil mapping has developed as an auxiliary method. Digital soil maps predict the values of a target variable over a given area by running a statistical model on training data and predictor data (Minasny et al., 2019; Rudiyanto et al., 2016; McBratney et al., 2003). When applying digital soil mapping techniques to peatland analysis, field survey data of the target peatland characteristic is used for training data. In the context of peatlands, 'predictors' are environmental variables that may have some relationship with the target peatland characteristic (Minasny et al., 2019; McBratney et al., 2003). Peatland predictors are established from the indicators of peatland presence, drivers of peatland initiation and development, and the sensors capable of observing these features (Minasny et al., 2019).

Peat-ML is a spatially continuous global map of peatland fractional coverage developed using a digital soil mapping workflow (Melton et al., 2022) (Figure 1.1). Existing regional peatland fractional coverage maps and predictors of peatland presence were used to train a machine learning (ML) model, which then predicted the fraction of peatland coverage within a grid cell globally on a five arc minute grid (grid cell widths of ca. 0.0833° which corresponds to 9.26 km at the Equator and 4.63 km at 60°N). Peat-ML was evaluated against training data that was iteratively withheld from the model, as well as other peatland coverage maps that had been derived through various techniques, and was found to do well in these inspections. The cross-validated Peat-ML output had a mean bias error (MBE) of -0.29% and a root mean square error (RMSE) of 9.11%, which indicate that the model was able to replicate its peatland test data relatively closely. For the Boreal Plains region of Canada, where a more detailed inter-map comparison was conducted, the cross-validated Peat-ML out-

put performed nearly as well as a more labour-intensive traditionally mapped product based on soil surveys and air-photo interpretation (Melton et al., 2022). This region was chosen for analysis due to the significant amount of observation data available for comparison through Ducks Unlimited Canada (Smith et al., 2007). Hugelius et al. (2020) first assembled a database of peat cores for training data, then used an ML approach to predict peat depth over the northern latitudes. However, they also produced a separate peatland extent map to mask out non-peat regions after prediction rather than training their ML model to account for areas with a peat depth of zero cm. Hugelius et al. (2020)'s depth map initially underestimated deep peat values and achieved an RMSE of 142.2 cm, prior to bias correction. After applying bias correction, their results more closely resembled their observed peat depth data. Thus, ML has shown to be a reasonable strategy for predicting peatland extent and capable of serving as a basis for estimating peat depth, given there is peat data available for training.

Melton et al. (2022) were motivated to develop Peat-ML as previous peatland fractional coverage maps were of insufficient quality and scale for use in land surface models. Similar circumstances are encountered when working with peat depth, mainly because the primary means of measuring depth require researchers to be present in a peatland to apply techniques such as probing or ground penetrating radar, which become increasingly labour intensive as mapping extent grows (Gatis et al., 2019). Digital soil mapping strategies may provide an effective method of mapping peat depth for large areas. A review of digital peatland mapping efforts by Minasny et al. (2019) indicate that various digital soil mapping processes have been applied to assessing peat depth. However, these peat depth datasets (e.g. Beilman et al., 2008; Rudiyanto et al., 2016; Akumu and McLaughlin, 2014) have varying grid resolutions, have not reached a global scale, and some have not undergone validation or uncertainty quantification (see Table 4 in Minasny et al. (2019) for complete list), which would leave a combination of these products unsuitable for use directly in a land surface model. To my knowledge PEATGRIDS (Widyastuti et al., 2024, currently in preprint), which is also generated using ML methods, is the only global map of peat depth and C stock. Nonetheless, PEATGRIDS is not continuous over the entire land surface, as depth values were primarily produced for grid cells considered to be 'peat dominated' by the Global Peat Map (United Nations Environment Programme, 2021).

The global distribution of peatland depths and subsequent peatland C stocks

therefore remains uncertain. Thus, the objective of my thesis is to address two main questions:

1. What is the global distribution of peatland depths?

2. How much C is stored within peatlands?

With the goal of filling these knowledge gaps, I have developed a spatially continuous global map of peatland depths and used these results to produce preliminary C stock estimates. The global peat depth map was created by adapting the ML framework constructed for Peat-ML to predict peat depth instead of fractional coverage. Hereafter, the original ML workflow of Peat-ML will be referred to as the Peat-ML Framework and the newly developed peat depth version will be called PeatDepth-ML (PD-ML). To facilitate the implementation of a global peat depth digital soil mapping method, substantial amounts of observed peat depth data is required for model training and testing. To this end, I formed Peat-DBase version 1.0, the largest database of peat depth measurements with global coverage that I am aware of and an important contribution to peatland science. As such, Chapter 2 is dedicated to the documentation of Peat-DBase and its uncertainties, with the intention of being published as a database paper providing public access to the data. Chapter 3 details the development of PD-ML and its use to predict a global peat depth map as a means of finding the global distribution of peatland depths. The use of PD-ML's output to estimate peatland C stocks is also discussed in Chapter 3. I compare my peat depth and C stock results to other studies, particularly Hugelius et al. (2020) and PEATGRIDS, and demonstrate the importance of quantifying the uncertainty in peat depth modeling that results from potential bias in the observational peat depth inputs. Chapter 3 is also intended for future publication, thus there is some repetition of information across Chapters 1, 2, and 3. Chapter 4 addresses the overall conclusions that can be drawn from my global peat depth map and C stock estimates.

1.1 Terminology and Background

1.1.1 Peatland Definitions

Table 1.1 describes the different definitions used throughout this thesis and they are discussed in more detail through this Chapter. Following Joosten and Clarke (2002) and Lourenco et al. (2022), I define peatlands as a wetland ecosystem with or without

vegetation in which a layer of peat has accumulated at the surface of the landscape (Table 1.1). Peat is a soil containing significant amounts of partly decomposed organic material (Gumbricht et al., 2017; Page et al., 2011; Lourenco et al., 2022). The high level of organic matter marks peat as an organic soil. This is in contrast to mineral soils which make up most of the world's soils and are low in organic matter content, with even the organic surface horizon generally remaining below 5% organic matter (by weight) (Wilson, 2019; Voroney et al., 2024). Mineral soils are found throughout boreal and tropical forests, as well as grasslands, savannahs, and deserts (Voroney et al., 2024). Osman (2013) found that the average amount of organic matter in one m depth of forest soils ranged from 0.87 to 2.50%. While no internationally accepted definition of peat currently exists, Lourenco et al. (2022) show that a minimum of 20% organic matter is among the lowest values presented for a soil to be considered peat over a variety of definitions in previous literature. Given the widespread nature of mineral soils and their prominent organic matter difference from peat, a binary representation of peat and mineral soil is sufficient for this thesis as my work is not intended to provide detailed information on non-peat areas. More detailed soil classification schemes have been developed such as the World Reference Base for Soil Resources and regional schemes like the United States Department of Agriculture Soil Taxonomy system and the Canadian System of Soil Classification (IUSS Working Group WRB, 2022; Soil Classification Working Group, 1998; Soil Survey Staff, 1999). The histosol classification of the World Reference Base is considered peat in several regions, however the name and definition of histosols can vary under other schemes (IUSS Working Group WRB, 2022; Xu et al., 2018; Soil Survey Staff, 1999; Soil Classification Working Group, 1998).

Table 1.1: Definitions of peatlands used throughout this thesis and their applications.

Definition	**Use within thesis**
Peatlands are a wetland ecosystem with or without vegetation in which a layer of peat has accumulated at the surface of the landscape (Joosten and Clarke, 2002; Lourenco et al., 2022).	This definition is used throughout the thesis as the basic description of a peatland.
Continued on next page	

Definition	**Use within thesis**
Peat soil has a minimum of 5% organic C content to a minimum depth of 10 cm (Lourenco et al., 2022).	This thesis applied the 5% organic C content threshold of this definition explicitly to the treatment of soil with less 5% organic C content as mineral soil. This mineral soil delineation is used within Chapter 2 as part of the inclusion of mineral soil in Peat-DBase.
Peat must reach a depth of 30 cm to be considered a peatland. This is a common depth threshold within the literature (Loisel et al., 2017; Melton et al., 2022; Lourenco et al., 2022).	This thesis applied this depth threshold only when plotting peat depth data, e.g. Figures 2.1, 3.2, 3.4, 3.8. This depth threshold is never used to limit the collection or prediction of peat depth data.

Peat is often defined with minimum organic matter or organic C content (by weight) thresholds, together with a minimum depth threshold. Organic C content differs from organic matter in that organic matter refers to the combination of all elements that make up such material (Lourenco et al., 2022). The World Reference Base indicates that 40% organic matter coincides with approximately 20% organic C content (IUSS Working Group WRB, 2022). Lourenco et al. (2022) suggest peat soil be defined as having a minimum of 5% organic C content to a minimum depth of 10 cm. This thesis broadly follows this definition in treating soil with less than 5% organic C content as mineral soil (Table 1.1). No minimum depth requirement is applied to data collection or model predictions within this work, however negative values are not permitted in the output of PD-ML.

While the definitions used within peatland science have been inconsistent since the field's inception, the 5% organic C content boundary and lack of minimum depth applied in this thesis is conservative (Lourenco et al., 2022; Gumbricht et al., 2017; Page et al., 2011; Zinck, 2011; Xu et al., 2018). Lourenco et al. (2022) selected the minimum boundary of 5% organic C content based on the recommendation of the Food and Agriculture Organization (**?**). The Food and Agriculture Organization explains that because organic C content is a dry weight percentage, it captures little information on volumetric C content which represents the amount of soil C that could be emitted should drainage and subsequent exposure to oxygen occur. A pure peat generally has low volumetric C content, but a high dry weight organic C content

percentage. However, a soil with a lower organic C content and higher weight could have a similar volumetric C content and thus the same emissions upon drainage (**?**). Therefore, using a lower bound of 5% organic C content is beneficial from a conservation and C accounting viewpoint, as it can account for more soil C stocks. The lack of a minimum depth requirement is motivated by similar factors. For example, the edges of existing peat bodies or relatively young peatlands could be shallow and risk being excluded by applying a depth threshold (Grand-Clement et al., 2015; Hribljan et al., 2016). Including as many different samples as possible is also practical from the ML spatial mapping perspective as such ML algorithms can only make reasonable predictions over regions that demonstrate the same behaviour as the areas represented within their training data (Meyer and Pebesma, 2021; Meyer et al., 2019). However, for consistency with existing peat research practice, I plot some peat depth results with distinct colours below and above 30 cm (see Table 1.1 and figures of Chapter 3). This 30 cm depth threshold is a common delineation within the literature (see reviews by Loisel et al. (2017) and Lourenco et al. (2022)), with Peat-ML being one example of its use (Melton et al., 2022).

1.1.2 Modelling Definitions

Having an understanding of the general structure of the Peat-ML Framework and the terms used in association with these modelling processes is a useful primer for learning about PD-ML. Figure 1.2 shows the steps within the Peat-ML Framework.

The Peat-ML framework takes in training data and predictor data that have been gridded at the same resolution (five arcminute grid cells, equivalent to approximately 0.0833° or 9.26 km at the Equator and 4.63 km at 60°N). The training data is composed of multiple peatland and non-peatland mapping products of different regions (Figure 1.3). Various terrain, climate, vegetation, and soil datasets were used as predictors. The initial collection of predictors is filtered to remove those with high information overlap, or high multicollinearity, and find those with the greatest predictive power. This predictor filtration reduces the risk of the model becoming overfit. Overfitting occurs when a model has too closely memorised its training data and therefore cannot generalise well when predicting into unseen areas (Hawkins, 2004; Peng and Nagata, 2020; Russell and Norvig, 2020). The variance inflation factor (VIF) is used to eliminate the most multicollinear predictors, and recursive feature elimination (RFE) is used to find the most informative subgroup of these predictors to use

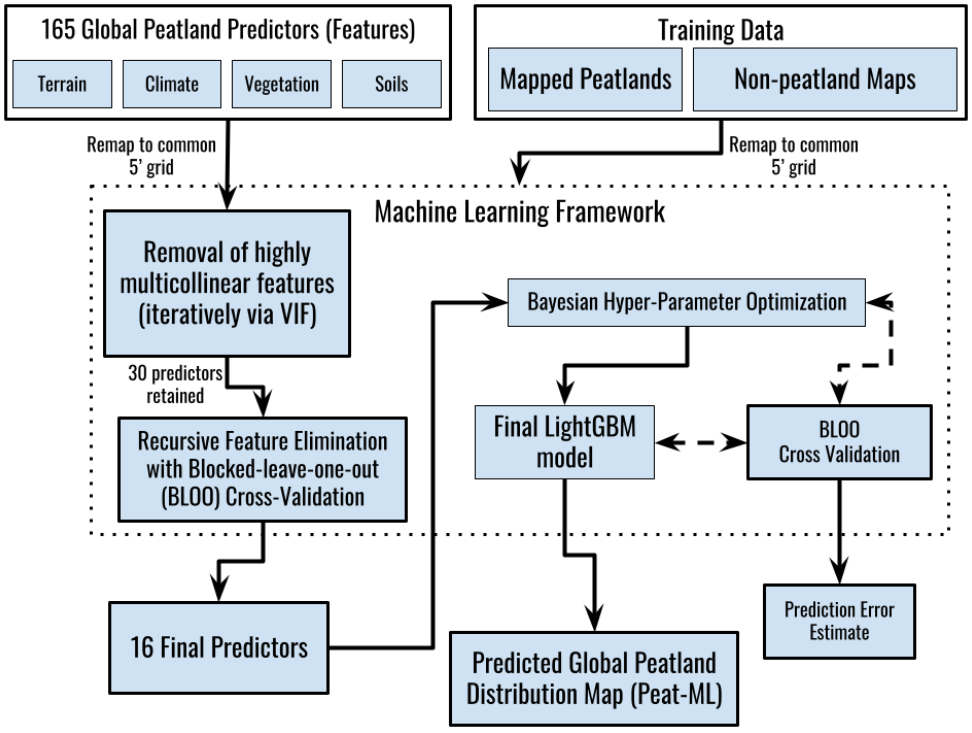

Figure 1.2: Flow chart of the steps within the Peat-ML Framework ML process from Melton et al. (2022). Terms and acronyms within this figure are explained later in the text.

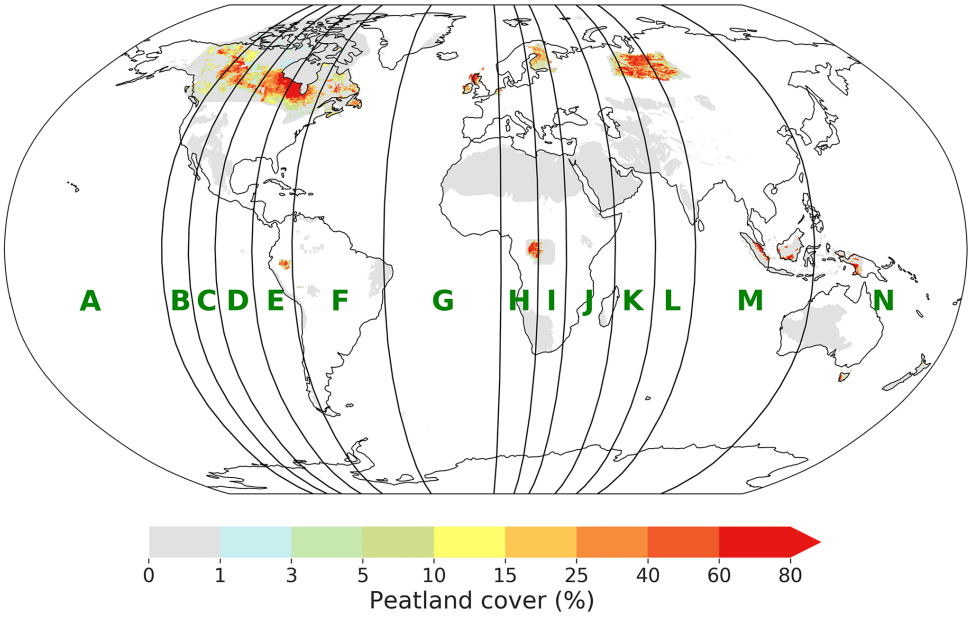

Figure 1.3: The Peat-ML training data from Melton et al. (2022), where white space has no data present. The green lettered blocks are used in the cross-validation process.

in the final prediction. RFE finds this subgroup by iteratively excluding predictors with the lowest importance (quantified by importance criterion calculated within the ML algorithm) and retraining the model until it finds the subgroup that yields the best model performance (Kuhn and Johnson, 2013).

The ML algorithm used in Peat-ML is the Light Gradient Boosting Machine or LightGBM, an efficient gradient boosting decision tree algorithm that was chosen for its computational speed (Ke et al., 2017). Decision trees can be thought of as a hierarchical structure made up of decisions about some data, with each decision breaking the data into smaller groups to approach a solution (Russell and Norvig, 2020). Gradient boosting decision tree algorithms generally operate by constructing multiple decision trees, where each consecutive tree aims to predict the error of the previous tree, then combining the tree predictions together to arrive at a final result (such as by calculating a weighted sum). The first gradient boosting tree attempts to predict the error of some initial prediction, such as the results from a very simple decision tree (Russell and Norvig, 2020; Wade, 2020; Ke et al., 2017).

With the optimal set of predictors established, Bayesian hyper-parameter optimization (BHO) is used to adjust the parameters of the LightGBM algorithm for the best model performance. Such parameters change how the algorithm behaves in some

way and will have certain values that produce the best results for the given task. For example, the `learning_rate` parameter controls the weighting in the weighted sum of decision tree predictions and unless it is adjusted, early trees with higher errors may over contribute to the final result (Wade, 2020). ML algorithms frequently have many parameters or the parameters have continuous values which are challenging to tune by hand, so additional algorithms exist to search the parameter space for the most effective values. BHO treats the parameter values as an input vector and the model error is the output, then tries to find a function which minimises this output. Iterative model runs are conducted to get multiple input-output pairs with which to update the function and thus inform the best selection of parameter values (Russell and Norvig, 2020). With the most optimal parameters chosen, the model is then able to predict the final global peatland distribution map.

Blocked-leave-one-out cross-validation (BLOOCV) is used in the RFE step, the BHO step, and to produce an error estimate for the model. The general idea of cross-validation stems from the need to use some data for model training and some for model validation, without using the same data for both purposes at the same time. If the same data was used for training and validation, the validation results would yield an overly optimistic view of model performance as the model was not assessed on its ability to predict beyond what information it had already seen. To achieve data separation, the original training data can be separated into k groups that are referred to as folds or blocks. For k iterations, the model is trained using all but one of the folds, with the remaining fold then being used for validation. The average performance score over the k validations provides a more reasonable estimate of the model's capabilities (Russell and Norvig, 2020; Meyer and Pebesma, 2022).

Cross-validation is a standard practice in ML, with random sample k-fold cross-validation being a common option wherein the data is split into folds randomly (de Bruin et al., 2022). Randomly separating the data in this way assumes that the data itself was collected via random sampling and is therefore identically distributed and independent. In practice, such an approach to in situ data collection is rarely possible and most measurements in geospatial databases are highly clustered around areas of interest for research (Meyer and Pebesma, 2022). Geospatial data is often spatially autocorrelated, meaning that measurements in close proximity are similar and should be treated more like replicates rather than separate data points while within the distance of such spatial relationships (Ploton et al., 2020). Using a random sample k-fold cross-validation in such scenarios can produce unrealistically high

estimations of model performance since it does not account for spatial dependence within the data (de Bruin et al., 2022; Meyer and Pebesma, 2022; Ploton et al., 2020). For BLOOCV within Peat-ML, the training data is separated into blocks where the size of each block is chosen such that it exceeds the distance of spatial autocorrelation and contains a roughly equal amount of data (Figure 1.3). Using BLOOCV in this way is beneficial when spatial autocorrelation may be present and prediction errors estimated with it more closely resemble real model error (Melton et al., 2022).

Chapter 2

Peat-DBase: A Compiled Database of Global Peat Depth Measurements

Peatlands have been important C sinks and stores throughout history, but are now at risk of becoming C sources (Canadell et al., 2021). The waterlogged nature of peatlands limits decomposition; once oxygen has been depleted in the high water table through aerobic decay, only slow anaerobic decomposition is possible, thereby allowing submerged partially decomposed material to accumulate as peat over hundreds or thousands of years (Koster and Favier, 2005; Page and Baird, 2016; Joosten and Clarke, 2002). Thus, peatlands, as wetland ecosystems, have developed to contain roughly a third of global soil C (Joosten and Clarke, 2002; Jackson et al., 2017) while only making up about 3% of Earth's land cover (Xu et al., 2018; Melton et al., 2022). Peatlands also store about 10% of the world's fresh surface water and are home to unique flora and fauna (Joosten and Clarke, 2002; Page and Baird, 2016). However, anthropogenic activity is disturbing these ecosystems. Humans drain peatlands for agricultural use and forestry, among other purposes (Fluet-Chouinard et al., 2023). Drainage allows the peat to dry and aerobic decay to resume due to the lowered water table, which prompts an increase in C emissions (Page and Baird, 2016; Warren et al., 2017; Koster and Favier, 2005). The rising temperatures and decreasing soil moisture brought about by climate change can also cause peatlands to dry out and experience more wildfires (Canadell et al., 2021; Helbig et al., 2020).

There is a growing need for comprehensive data on peatlands. For example, peat-

land C stocks can be estimated using peatland area, depth, organic C content, and bulk density (Minasny et al., 2019). However, there are often insufficient measurements of these values for detailed regional or global scale C stock inventories, thus such inventories may rely on single best estimates for one or more values (Page et al., 2011; Minasny et al., 2019). Additionally, there is more interest in investigating the historical and future impact of peatlands on the climate through their inclusion in land surface models. One application of land surface models can be to act as the terrestrial component of Earth system models and simulate the energy, C, and water dynamics of the land surface. Earth system models are then capable of performing global climate simulations. Thus, efforts are being made to add peatland C, water, and energy fluxes to land surface models to explore how these fluxes impact the simulation of climate processes in Earth system models (Chadburn et al., 2022; Bechtold et al., 2019; Wu et al., 2016).

Peatland data is relatively limited as they are not well represented in global soil mapping initiatives (Krankina et al., 2008; Minasny et al., 2019). The wet and often remote locations of peatlands make them difficult environments to gather data in (Minasny et al., 2019; Rudiyanto et al., 2016). Mapping peat depth is especially challenging as it generally involves performing field surveys or proximal remote sensing techniques (Minasny et al., 2019; Jowsey, 1966).

To support present modelling initiatives and determine regions in need of further research, we present the Peat Depth Database (Peat-DBase) version 1.0, the largest collection of global scale, harmonised, and quality controlled basal peat depth measurements to date. A version of Peat-DBase with more extensive data on peat-free areas is also produced to further enable future modelling efforts. Mineral soil core data was used to supply this information (Batjes et al., 2020). Chapter 2.1 describes the acquisition, formatting, and processing of peat study data used to form Peat-DBase. The subsequent acquisition, formatting, and processing of mineral soil data is explained in Chapter 2.2. Chapter 2.3 provides an analysis and discussion of the resulting database, including its limitations and areas of future work. Conclusions and data availability are discussed in Chapters 2.4 and 2.5 respectively.

2.1 Peat Study Data

2.1.1 Data Acquisition

Peat study data was accepted into Peat-DBase provided the measurements were taken down to the basal depth indicated by mineral soil or bedrock; otherwise, any coring or sampling method was allowed that was able to accurately determine the distance from peat surface to bedrock or mineral soil. In permafrost regions, Cold Regions Research and Engineering Laboratory corers (Brockett and Lawson, 1985) were often used. In non-permafrost sites, Russian-type corers (Jowsey, 1966), Box corers (Shotyk and Noernberg, 2020; Fenton, 1980), and Jeglum corers (Jeglum et al., 1991) were the primary tools. In cases where pole probing was conducted as part of a coring transect, the probing measurements were also included. Probing involves the use of metal poles which are inserted into the peat until they meet a non-peat layer and cannot go any further (Crezee et al., 2022). The sampling method varied depending on the goals of the researchers. In some cases, transects of varying lengths were chosen with measurements taken at consistent intervals across the transect (Crezee et al., 2022; Kelly et al., 2020). In other instances, unique core sites were chosen across single or multiple peatlands (Cole et al., 2015; Davies et al., 2023b,a; Silvestri et al., 2019b). Some peat study datasets also included some peat-free cores as a result of their sampling procedures (e.g. Crezee et al., 2022; Keys and Henderson, 1987; Thibault, 1992).

Generally, the largest previously existing peat depth datasets were compilations of other datasets themselves. Such compiled datasets were frequently developed for modelling purposes (Hugelius et al., 2020; Treat et al., 2017, 2019). These compiled datasets were added to Peat-DBase under the single citation of the compiling authors, with additional fields being used to track any information the compiling authors provided on their data sources (see Chapter 2.1.2 and Table 2.2). 32 580 peat depth measurements came from 19 published sources (post-error/duplicate assessment in Table 2.1). 2552 measurements came from 5 sources that are currently in preprint or that make their data available upon request (post-error/duplicate assessment in Table 2.1). A publication may have been omitted from Peat-DBase because the data source included a lack of readily accessible and usable data files or an inability to confirm from the publication that the measurements went to basal peat depth. The sources of peat depth measurements used in Peat-DBase are listed in Table 2.1.

Table 2.1: The sources of peat study measurements in Peat-DBase (* indicates sources that are confirmed to be a compilation of other datasets).

Publications and data contributors of peat study data	**Geographic region covered by measurements**	**Number of measurements provided with all necessary data**	**Number of measurements retained after error/duplicate assessment**
Bauer et al. (2024) *	Canada	769	753
Beilman et al. (2009)	West Siberian Lowlands	23	23
Benfield et al. (2021) *	Sierra Nevada del Cocuy (Eastern Colombian Andes)	22	20
Cole et al. (2015)	Sarawak, Malaysian Borneo	3	3
Comas et al. (2015)	West Kalimantan, Indonesia	8	8
Crezee et al. (2022)	Central Congo Basin	1558	1558
Davies et al. (2021)	Southern Hudson Bay Lowlands	1	1
Davies et al. (2023a)	Western Hudson Bay Lowlands	2	2
Davies et al. (2023b)	Western Hudson Bay Lowlands	3	3
Gorham et al. (2012) *	North America	1685	1514
Hribljan et al. (2023)	Colombian, Ecuadorian, Peruvian, and Bolivian Andes	25	25

Continued on next page

Publications and data contributors of peat study data	**Geographic region covered by measurements**	**Number of measurements provided with all necessary data**	**Number of measurements retained after error/duplicate assessment**
Hugelius et al. (2020) *	North of 23°N	7738	7110
Kelly et al. (2020)	Quistococha, Pastaza-Marañón Foreland Basin, Peru	29	29
Keys and Henderson (1987); Thibault (1992) - Digital dataset supplied by E. Prystupa *	New Brunswick, Canada	20 505	20 505
Lab of Dr. Angela Gallego-Sala	North of 60°N and between 5°S and 5°N	230	230
Lawson et al. (2023)	Pastaza-Marañón Basin, Peru	280	280
Manitoba Department of Natural Resources and Northern Development (Available upon request) *	Manitoba, Canada	1709	1598

Continued on next page

Publications and data contributors of peat study data	**Geographic region covered by measurements**	**Number of measurements provided with all necessary data**	**Number of measurements retained after error/duplicate assessment**
Mariusz Lamentowicz, Daria Wochal, Adam Mickiewicz, Sambor Czerwiński, Joanna Landowska, and Jacek Landowski (2024)	Poland	263	263
Silvestri et al. (2019b,a)	Kubu Raya District, West Kalimantan, Indonesia	63	63
Treat et al. (2017, 2019) *	Global	614	504
Sun et al. (2023) *	Tibetan Plateau	146	146
Warren et al. (2012) *	Indonesia	33	33
Warren M. (unpublished)	Indonesia	276	275
Winton et al. (in prep)	Colombia	186	186

2.1.2 Data Formatting

All collected data was processed into a consistent format. Source datasets were first converted to a comma-separated values (CSV) file format. Any measurements that were missing a latitude, longitude, or depth value were dropped. All peat depth values were converted to centimetres and the coordinates of each measurement location were converted to the World Geodetic System 1984 (WGS84 or EPSG:4326) coordinate system where required. When the depth measurement was presented as a range

(this occurred in less than 5 measurements), the median of the boundary values was calculated and used within Peat-DBase. All datasets were combined into one CSV file. The columns of this synthesised database are explained in Table 2.2.

Table 2.2: The column headers within Peat-DBase and their meaning. Note that an 'exact duplicate' means that the lat, lon, and depth_cm values are identical. Note that 'rounded duplicate' means the lat, lon, and depth_cm values are identical when rounded to 2 decimal places.

Peat-DBase version 1.0 column names	**Peat-DBase version 1.0 column meaning**
original_dataset	A citation of the publication or owner of the original dataset that the measurement was retrieved from for Peat-DBase.
original_entry_num	A number representing the location of the measurement in the ordering of its original dataset.
lat	Latitude coordinate of peat depth measurement in decimal degrees.
lon	Longitude coordinate of peat depth measurement in decimal degrees.
depth_cm	The basal peat depth measurement in centimetres.
original_dataset_source_notes	Any citation information provided by the original_dataset publication or owner is retained here. If the original_dataset is confirmed to be a primary field study, this is noted here.
peat_measurement	A flag set to 'True' if the original_dataset is a peat study or source. It is set to 'False' if the data comes from WoSIS (see Chapter 2.2).
Continued on next page	

Peat-DBase version 1.0 column names	**Peat-DBase version 1.0 column meaning**
suspected_duplicate_assessment	A numerical flag that indicates whether the measurement has been observed as a possible duplicate of another data point. '1' - the measurement was confirmed to be unique by the original publication/owner and did not undergo assessment as part of this work. '2' - The measurement was not initially detected as a possible duplicate. '3' - the measurement was found to be the first instance of an exact duplicate. '4' - the measurement was found to be a redundant instance of an exact duplicate. '5' - the measurement was found to be the first instance of a rounded duplicate. '6' - the measurement was found to be the redundant instance of a rounded duplicate, but manual assessment of the data sources did not find reasonable evidence that this was a true duplicate. '7' - the measurement was found to be a redundant instance of a rounded duplicate and manual assessment of the data sources found evidence that this was a true duplicate.

Continued on next page

Peat-DBase version 1.0 column names	**Peat-DBase version 1.0 column meaning**
suspected_error_assessment	A flag set to 'True' if the measurement contains possible errors (such as incorrect coordinates). Otherwise it is set to 'False'. Any notes on the nature of the error are provided in the investigation_notes column.
investigation_notes	Any notes on the nature of the error in suspected_error_assessment.
original_dataset_collision	If a measurement was found to be a redundant instance of an exact or rounded duplicate and given a suspected_duplicate_assessment value of '4', '6', or '7', then this column will be set to the original_dataset value of the measurement that was found to be the first instance of that duplicate.
original_entry_num_collision	If a measurement was found to be a redundant instance of an exact or rounded duplicate and given a suspected_duplicate_assessment value of '4', '6', or '7', then this column will be set to the original_entry_num value of the measurement that was found to be the first instance of that duplicate.

Continued on next page

Peat-DBase version 1.0 column names	**Peat-DBase version 1.0 column meaning**
group_id	A hash code that is generated from the original_dataset and original_entry_num value of measurement during the process of duplicate assessment. Secondary instances of a duplicate will be given the group_id of the corresponding first instances during this process.

2.1.3 Data Processing

Data entry errors must also be handled. As mentioned in Chapter 2.1.2, measurements that were incomplete, e.g. missing a latitude or longitude value, were excluded prior to data processing. However, there are some complete measurements that had obvious errors, e.g. a peat measurement located in the ocean. The clearest cases of possibly erroneous entry were flagged accordingly with the suspected_error_assessment column in Peat-DBase being set to True and investigation_notes recording any information on the error (see Table 2.2).

Due to our inclusion of compiled peat depth datasets that are themselves compilations of previous datasets (see entries marked with an * in Table 2.1, e.g. Treat et al., 2017; Hugelius et al., 2020) within Peat-DBase, some additional processing steps are needed for quality control. Sarracino and Mikucka (2017) investigated the impact of duplicate data entries on regression estimates and possible methods of resolving such an issue. If used for modelling, the presence of duplicates in input data can bias regression estimates, particularly when their distribution is not random, such as when the same peat depth measurements appear across different compiled datasets (Sarracino and Mikucka, 2017). Therefore, as each new source dataset was added to Peat-DBase, the peat depth measurements within the database were reassessed for duplicates. Duplicates were flagged such that one measurement can be kept while superfluous entries can be filtered out as needed by a database user (see Table 2.2 for columns used in duplicate assessment). This removal strategy was one of the most successful at reducing bias among those tested by Sarracino and Mikucka (2017), with other options being to ignore duplicates, drop all instances of a duplicate, or apply a

form of weighting or control term to the duplicates.

There were two phases of duplicate flagging. In the first phase, duplicates that had the same exact depth value and coordinates were found, and all but the first instances of any such duplicates were flagged for removal. For the second phase, the precision of the measurements was intentionally downgraded to detect rounding by previous data sources. This check found measurements that had both the same depth values when rounded to the nearest 0.01 centimetres and the same coordinates when rounded to the nearest 0.01°, then flagged them for manual assessment. These specifications were determined by iteratively checking less precise rounding until the number of duplicates (confirmed via the following assessment) became minimal. We assessed these potentially duplicated measurements, and any citation information provided by the sources we acquired them from, for evidence they may have originated from a common study. A measurement was then flagged for removal if its source was a compiled dataset that possibly contained processed or rounded versions of measurements also contained within another study included in Peat-DBase. For example, Sun et al. (2023) and Treat et al. (2017) both use data from Zhao et al. (2014) and convert the original coordinates from arc minutes to decimal degrees. Treat et al. (2017) often retained less significant figures for the coordinates, so their instances of the duplicated measurements were flagged for removal in Peat-DBase. However, the citation practices of some datasets did not always allow for robust conclusions. If the source dataset was confirmed to result directly from a field study, the measurement was retained. In general, significant caution was used for data removal such that measurements that did not have a clear possibility of being rounded versions of each other were kept. In all cases, one instance of a possible duplicate set was kept. Duplicate flags and a flag to indicate whether a measurement had been previously assessed for duplication were tracked in data fields present for each measurement in the original database (see Table 2.2).

To provide a point of reference for the shape of the non-zero cm depth distribution in version 1.0 of Peat-DBase, a basic assessment of some probability distribution functions available through the Python SciPy package (Virtanen et al., 2020) was performed (i.e. Lognormal, Poisson, Weibull Minimum, Gamma, Beta Prime, Inverse Gaussian, and Pareto). The Weibull Minimum distribution produced the lowest RMSE when compared to the other probability distribution functions tested (Virtanen et al., 2020).

2.2 Non-Peat Study Data

Non-peat study data was added to Peat-DBase to provide a representation of non-peat regions. Given that non-peat areas were not the primary concern of Peat-DBase, detailed soil profiling was not the priority. Rather, the goals in acquiring non-peat data were broad land coverage and mineral soil presence. Measurements in these areas would be assigned a peat depth of zero cm.

2.2.1 Data Acquisition

The World Soil Information Service (WoSIS) maintains a harmonised and quality-controlled database of global soil profiles for digital soil mapping purposes. Existing soil data is submitted by owners for consideration in WoSIS, at which point it is stored, assessed, and standardised through the WoSIS workflow. Iterations of the fully quality assessed and standardised database are released periodically as snapshots. The September-2019 snapshot was acquired for use in Peat-DBase. This snapshot is documented in Batjes et al. (2020) and the data is available through Batjes et al. (2019).

2.2.2 Data Formatting

The non-peat study data was processed to the same format as the peat study data. The WoSIS database is stored across several tab-separated values (TSV) files. Only the wosis_201909_profiles.tsv and wosis_201909_layers_chemical.tsv files were required for the next processing steps, thus these were converted to a CSV file format. These files contain the soil profile coordinates in WGS84 format and the chemical properties of the soil profiles respectively (Batjes et al., 2020). This information was used to determine which cores represented mineral soil cores and therefore non-peat cores.

2.2.3 Data Processing

The acquired soil profile data was filtered to only include mineral soil profiles. This study broadly follows the definition of a peatland suggested by Lourenco et al. (2022) which is an area with a minimum of 5% organic C content to a minimum depth of 10 cm. Here we treat soil with less than 5% organic C content as mineral soil for

the purpose of establishing peat-free locations (see Table 1.1). The WoSIS dataset often contains organic C content measurements in g/kg for multiple layers within a soil profile, although not all soil profiles have organic C content data associated with them (Batjes and Van Oostrum, 2023; Batjes et al., 2020). Thus, the first processing step was to drop all soil profiles with no organic C content measurements since their qualification as mineral soil based on our 5% organic C content boundary could not be quickly determined otherwise. Next, all organic C content measurements were converted from gC/kg of soil mass to a percent of mass. Any profiles that had a layer with an organic C content greater than or equal to 5% were then dropped from the dataset. The coordinates of all remaining soil profiles were collected as a new dataset and these locations were assigned a peat depth of zero cm.

The data derived from WoSIS was not subject to any duplicate assessment within this study. A duplicate assessment was not necessary within the dataset itself, as an assessment and exclusion process was applied prior to the release of the WoSIS snapshot (Batjes et al., 2020). A duplicate assessment was not conducted between the derived mineral soil cores and the peat study database as it was assumed that it would be unlikely for significant duplication to occur between the mineral soil profiles and the peat study data.

2.3 Results and Discussion

2.3.1 General Data Totals

After harmonisation and quality control measures have been applied Peat-DBase version 1.0 contains 35 132 measurements from 24 peat study sources (Table 2.1) and 129 747 measurements when including the non-peat study data from WoSIS (Figure 2.1). The peat study data spans 54.933°S to 82.217°N. This latitudinal range remains the same when considering non-peat study data as well. Figure 2.1a shows that a significant amount of peat study data is located in the northern extratropics and that most peat depth measurements appear to exceed 30 cm of depth. A minimum peat depth of 30 cm is a common threshold for an environment to be called a peatland (Loisel et al., 2017). Within the data from peat studies, 3286 measurements (9.4%) have a peat depth of zero cm, often from sampling schemes that measured across transects to test for peatland presence, e.g. Crezee et al. (2022) and Keys and Henderson (1987). The number of zero cm depth measurements increases to 97 901 when

including the filtered data from WoSIS (Figure 2.1b). However, Figure 2.1b also indicates that there is a limited amount of soil profile data available in desert regions such as the Sahara Desert, as well as for specific countries such as Paraguay. These data gaps are present in the WoSIS database prior to any filtering applied within the data processing steps of Chapter 2.2.3 and are discussed in more detail in Batjes et al. (2020).

2.3.2 Spatial and Depth Distribution of Data

Most major peatland areas are represented in Peat-DBase. The distribution of peat depth measurements within Peat-DBase (Figure 2.1a) can broadly be compared to the peatland fractional coverage presented in PEATMAP (Xu et al., 2018) and predicted in Peat-ML (Melton et al., 2022) (Figure 1.1). PEATMAP is a map of peatland fractional coverage developed by combining the most recent and detailed maps available prior to 2018 for any given area (Figure 1.1). Peat-ML is a peatland fractional coverage map predicted using ML. Peatland fractional coverage maps of various regions and datasets of a variety of environmental variables were used to train a ML model, which then produced a global peatland map (Melton et al., 2022). Generally, Peat-DBase has measurements present in most of the peatland complexes shown in PEATMAP and Peat-ML, such as those across North America, Eurasia, South America, the Congo, and the Malay Archipelago (Figure 2.1a and Figure 1.1). However, Peat-DBase is absent or has a limited number of depth measurements in the Amazon Basin, Indonesia, and Papua New Guinea, where PEATMAP indicates more extensive peatland fractional coverage (Xu et al., 2018). Similarly, Peat-ML also indicates greater coverage in these regions, as well as in Eastern Russia (Figure 1.1). A paleoecological view of peatlands also suggests a greater presence of peatlands in Eastern Russia (see Figure 1 of Yu et al. (2010)), however we are not currently aware of any additional readily available peat depth datasets for the region.

Peat-DBase represents much of our current knowledge on peat depth, however the database also contains implicit biases as a result of the constraints within field research. For example, Figures 2.1, and 2.2 show that the low latitudes have less data overall. Tropical peatlands have historically undergone less mapping in general (Zinck, 2011; Ruwaimana et al., 2020). While the lack of peat data for the tropics contributes to the data distribution within Peat-DBase, efforts are being made to improve our knowledge of the region (e.g Winton et al., (in prep)) and upcoming

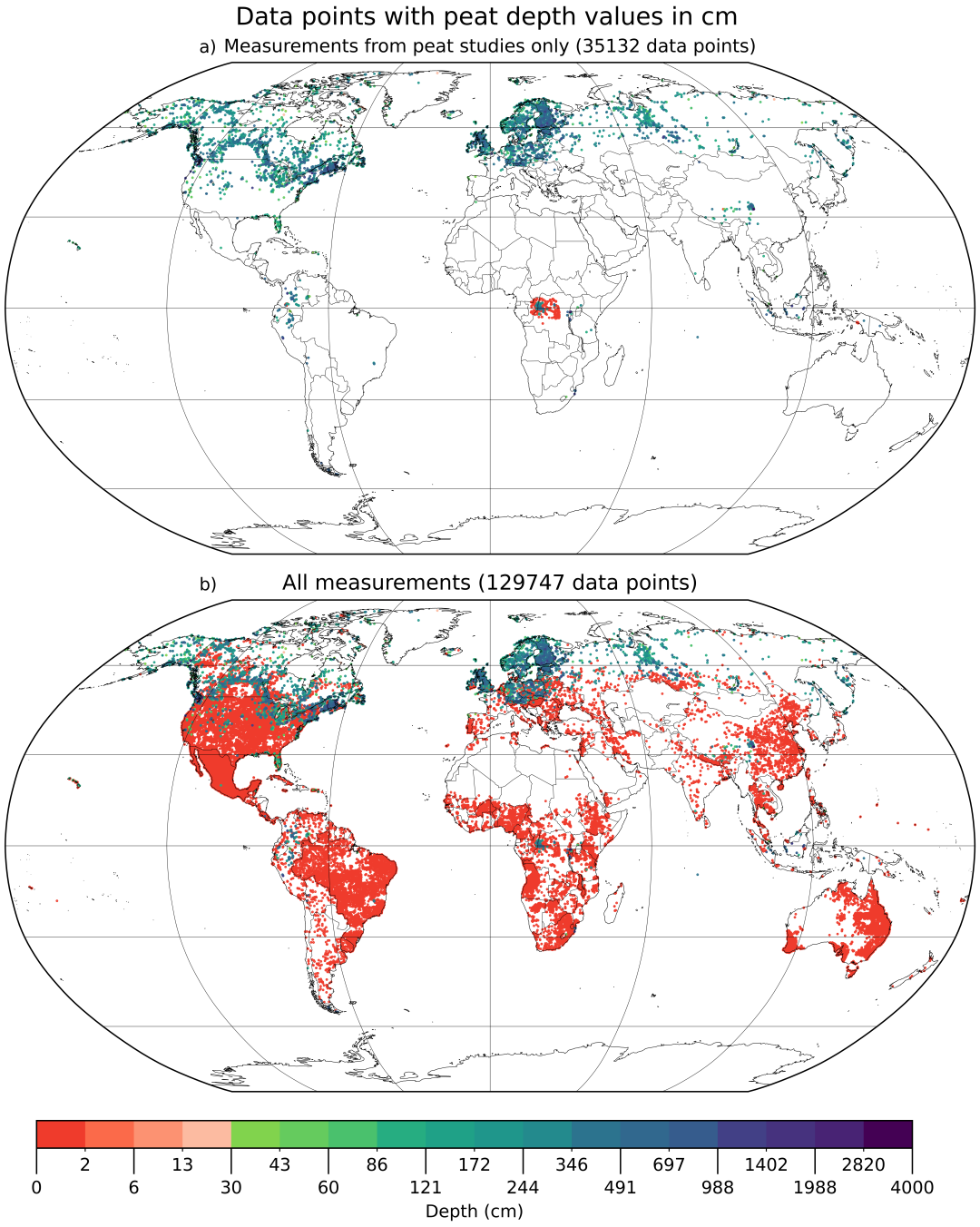

Figure 2.1: Geographical distributions of the data points in Peat-DBase version 1.0, in cm of peat depth. (a) The data points originating from peat studies. (b) All data points, including those filtered from WoSIS. The colour bar is presented in a log scale with a colour break at 30 cm to represent a commonly used threshold when classifying an area as peatland (see Table 1.1) (Loisel et al., 2017).

studies can be added to Peat-DBase over time.

The distribution of peat measurements in Peat-DBase reflect the prominence of peatlands in boreal, temperate, and tropical regions (Xu et al., 2018; Melton et al., 2022; Joosten and Clarke, 2002; Koster and Favier, 2005). The latitudinal and depth distribution of data points shown in Figure 2.2 indicates peat measurements are primarily concentrated in the high latitudes, particularly between 40°N and 50°N, and near the equator. Some other isolated peatland regions can also be observed, such as the grouping around 35°N that corresponds with peatland complexes sampled in Florida and the Tibetan Plateau (Figure 2.1a). The deepest peat measurement of 3527 cm is located in the Tibetan Plateau (Sun et al., 2023). The present day deepest peat measurements often appear in regions that would not have been glaciated during the Last Glacial Maximum or would have been among the first areas to become ice free as the ice sheets retreated and thus have had longer to accumulate, with topography such as flat floodplains or narrow basins formed from riverbeds also playing a role (Figure 2.1, 2.2, and see Figure 1 in Treat et al. (2019)) (Treat et al., 2019; Ruwaimana et al., 2020; Gowan et al., 2021). However, there is not always a strong or simple linear correlation between peat depth and age, as peatlands can experience periods of enhanced growth, diminished growth, or loss due to changing climatic or hydrologic conditions (Ruwaimana et al., 2020; Blaauw and Christen, 2005; van Bellen et al., 2011).

The depth distribution of the non-zero cm measurements in Peat-DBase can be approximated by a Weibull Minimum distribution (red line in Figure 2.3 and see Chapter 2.1.3). Figure 2.3 shows the database is heavily weighted towards zero cm depths, with non-zero cm measurements making up 24.5% of the total database. Given that peatlands cover roughly 3% of Earth's land area, the strong non-peat presence is expected from a global land cover perspective (Xu et al., 2018; Melton et al., 2022). Since WoSIS contributes a large number of data points, it can be useful to observe some qualities of the database when it is excluded. Without the data from WoSIS, non-zero cm depths make up 90.6% of the measurements within the database and the mean peat depth is 231.3 cm.

Most peat field studies are not conducted with spatial scaling in mind, so their distribution is not identical and independent as a result. Instead, the measurements are usually highly clustered around places of interest for research (Hugelius et al., 2020; Meyer and Pebesma, 2022). This clustering means it is challenging to draw direct relationships between the ratio of peat to non-peat data within Peat-DBase

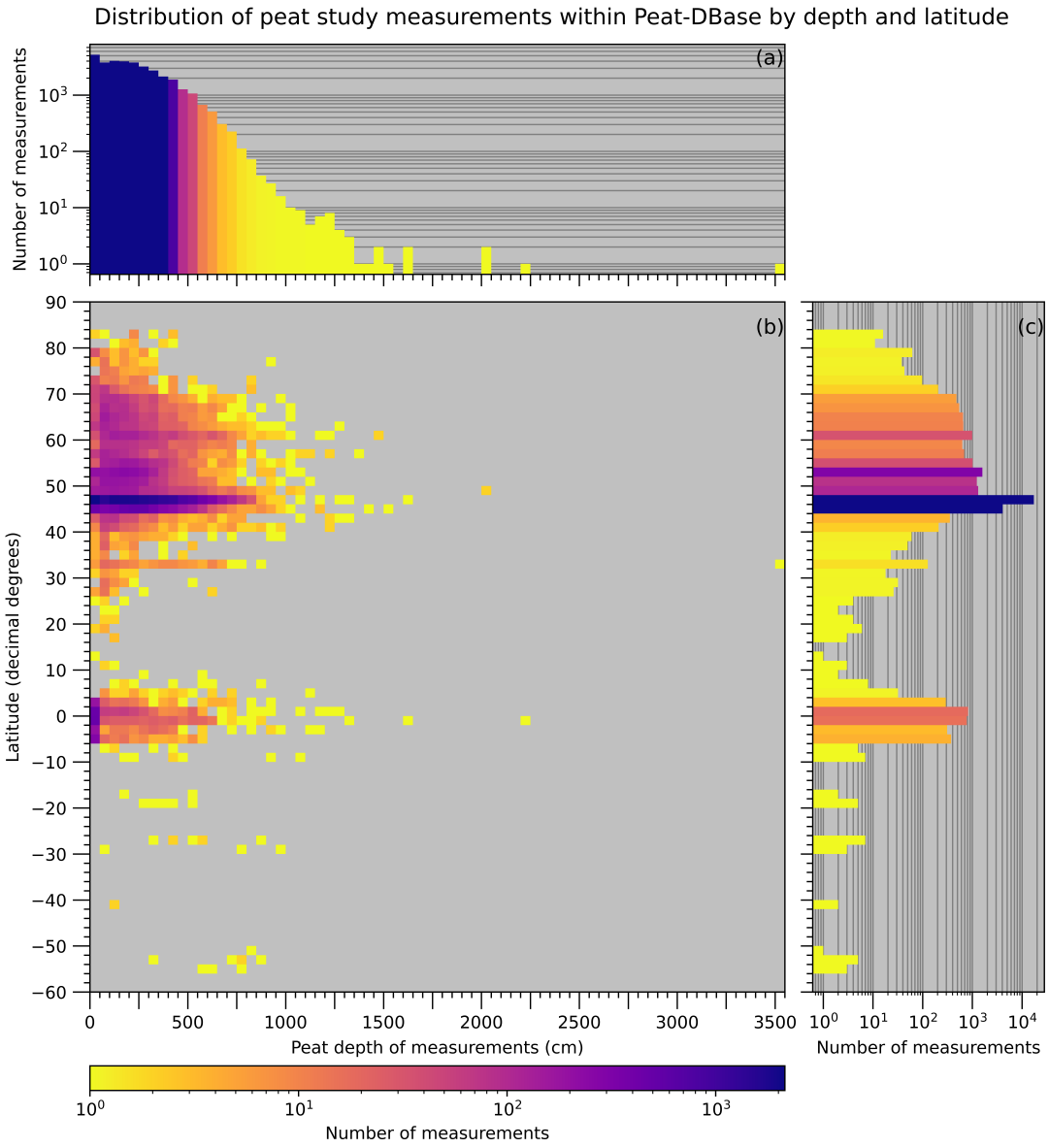

Figure 2.2: (a) Distribution of measurements from peat studies within Peat-DBase by depth in cm (b) Distribution of measurements from peat studies within Peat-DBase by depth in cm and latitude in decimal degrees. (c) Distribution of measurements from peat studies within Peat-DBase by latitude in decimal degrees. WoSIS data is not included within these plots.

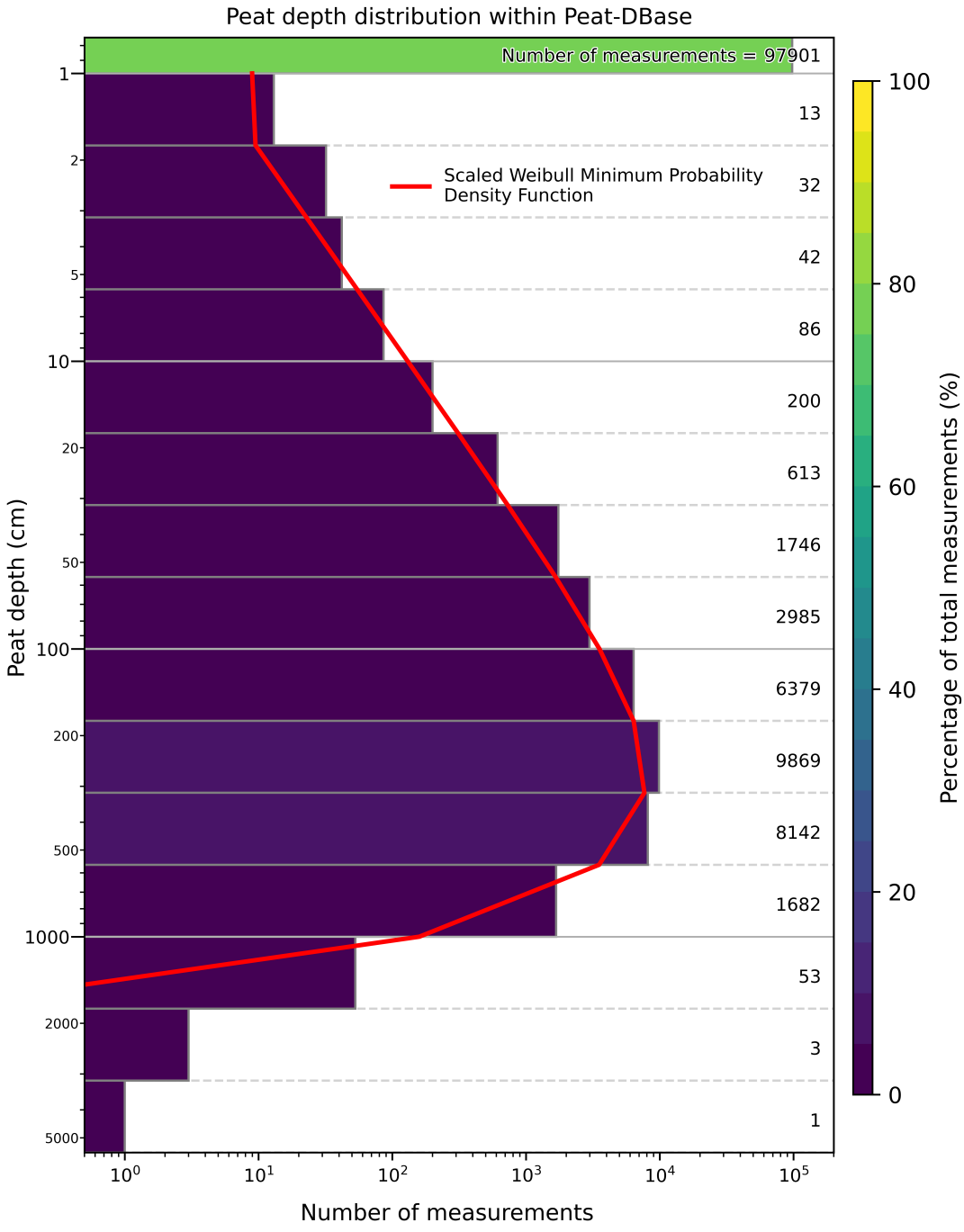

Figure 2.3: Peat depth distribution within Peat-DBase version 1.0 in cm grouped on a log scale of depth, 94615 measurements come from WoSIS (see Chapter 2.2). The red line indicates the Weibull Minimum distribution of the non-zero cm depth measurements within Peat-DBase (Virtanen et al., 2020).

and that of peat land coverage. Figure 2.3 also indicates there is significantly less data within Peat-DBase for the deeper peats. This distribution may be reflective of true peat development patterns, but it is also potentially impacted by sampling bias. The deeper the peat, the more difficult it becomes to acquire and transport the equipment necessary for coring to basal depth. Typical coring equipment can generally handle depths ranging from tens to hundreds of centimetres before adding extensions, depths exceeding 1000 cm begin to require more complex strategies and equipment (Bansal et al., 2023; Shotyk and Noernberg, 2020). Additionally, given that we prioritised the acquisition of large peat depth datasets, smaller datasets or studies using single cores which reached substantial depths may have been missed. The presumed deepest or center-most areas of peatlands are frequently prioritised for coring for the purpose of conducting paleo-reconstructions, estimating peat accumulation rates, or calculating C stocks of a peatland as some examples. In such cases, a limited number of cores may be collected (Hugelius et al., 2013; Hribljan et al., 2016; Loisel et al., 2017), which could also influence the distribution of depths within the database.

The collection of a small number of cores to represent an entire peatland area can influence the depth distribution accuracy of Peat-DBase. When a peatland has developed over a flat mineral basin and has relatively smooth surface formations, a limited number of measurements may be sufficient to capture the full range of peat depths present. If the peatland formed by infilling a more complex topographic region or if it has greater variation in its surface gradient, a minimal number of cores risks being less representative of peat depth (Hugelius et al., 2020; Loisel et al., 2017). For example, van Bellen et al. (2011) demonstrate the variability that can occur in the geometry of a single peatland using probing, coring, and ground penetrating radar. Their study indicated peat surface altitude variations of two to eight metres within peatlands, and that the deepest peats do not always occur at the geographic centre of a peatland due to the underlying basin topography. National inventories and datasets produced for peatland mapping initiatives often utilise transects or other sampling strategies intended to capture a range of peat formations (Hugelius et al., 2020; Crezee et al., 2022; Silvestri et al., 2019b). These broader sampling schemes help to balance peat depth representation within Peat-DBase.

2.3.3 Database Limitations and Future Work

There are uncertainties within Peat-DBase that should be understood prior to use of the database, as well as areas of potential improvement for the future. For example, Peat-DBase does not represent a precise record of current peatland persistence, as no restrictions were placed on the current state of peatlands in order for their depths to be included. Thus, human land use and climate change may have since disturbed or removed peatlands from which measurements were taken previously (Joosten and Clarke, 2002; Koster and Favier, 2005; Ratnayake, 2020; Silvestri et al., 2019b).

Given the varying definitions of peat that appear across the literature, collecting measurements from a range of unique sources introduces uncertainties into Peat-DBase (Lourenco et al., 2022; Gumbricht et al., 2017; Page et al., 2011; Zinck, 2011; Xu et al., 2018). For example, Silvestri et al. (2019b) require that peat consists of at least 30% organic matter, meanwhile, Cole et al. (2015) and Crezee et al. (2022) require an organic matter content of at least 65%. At this time, these unique definitions are not formally tracked within Peat-DBase.

The accuracy of peat depth and location measurements is limited by the technology available to researchers at the time of their fieldwork. For example, several measurements present in the database originate from surveys conducted from the 1950s to 1990s when global positioning system receivers were not invented yet, or less accurate (Treat et al., 2017; Hugelius et al., 2020; Sun et al., 2023; Keys and Henderson, 1987). Data storage capabilities are affected similarly, such that data may be available, but not in a format that is easy to access and process by modern standards (Thibault, 1992). The accuracy of peat coring and probing can also be variable. For example, metal probes may meet resistance prior to the base of the peat layer by coming into contact with buried wood fragments (Parry et al., 2014). Peat can also become compressed when cored, which would influence the resulting depth measurement (Shotyk and Noernberg, 2020).

The WoSIS database faces similar uncertainties in sampling methodologies to the peat study portion of Peat-DBase. Batjes et al. (2020) discuss the variation in accuracy of the geographic coordinates and lab measurements within WoSIS, as well as the corresponding measures of accuracy that they include within the dataset. Since Peat-DBase did not focus on precision in mineral soil presence, we did not account for the stated uncertainties in WoSIS when deriving non-peat locations.

The duplicate assessment described in Chapter 2.1.3 may introduce human error

into Peat-DBase. For example, the assessment is biased based on what we determine to be reasonable evidence of rounding. Additionally, we did not assess rounding to all possible decimal places, thus there could be more or less extreme instances of rounding that were not detected.

We acknowledge that while there are regions where data may not yet be available (see Chapter 2.3.2), there is also data that has not been included in this database. For example, we are aware of additional peat depth data available in the Amazon Basin, such as for the area surrounding the Madre de Dios River (Householder et al., 2012). Some relevant peat sources are also available for areas in Indonesia (Anda et al., 2021). In these cases, the data was not included as it was not available in a readily accessible point-based format. Also, publications with a single or small number of measurements may not have been included as we generally prioritised the collection of datasets with several measurements. Data from publications not previously included in Peat-DBase can be added in the future.

2.4 Conclusion

Taken together, Peat-DBase version 1.0 provides a strong and representative basis of peat depth data at the global scale across the literature. With over thirty-five thousand peat depth measurements from peat studies, it is the largest globally spanning database available to our knowledge and increases in size when considering the non-peat study data. The distribution of peat depth measurements corresponds with maps of peatland coverage to a large extent, and there are avenues of improvement for depth data coverage in areas with less correspondence. While the distribution of peat depth across Peat-DBase is likely affected by sampling bias, it raises awareness of this issue and presents opportunities for future research by demonstrating knowledge gaps.

2.5 Data Availability

Peat-DBase version 1.0 is stored in a CSV file located here `https://doi.org/10.5281/zenodo.15530645`. This dataset does not contain the WoSIS data and excludes any peat study data that has publication restrictions on it at the time of this writing.

Chapter 3

PeatDepth-ML: Using Machine Learning to Predict a Global Map of Peat Depth

Peatlands play a critical role in the C cycle, hydrological systems, and biodiversity in many parts of the world. As a type of wetland, these ecosystems are waterlogged landscapes with soil containing high amounts of organic material due to low oxygen levels and therefore limited decomposition (Rydin and Jeglum, 2013a; Joosten and Clarke, 2002). Peatlands hold about a third of global soil C stocks while only covering roughly 3% of all land area (Turunen et al., 2002; Ruppel et al., 2013; Melton et al., 2022; Joosten and Clarke, 2002). They are important environments for water storage, as they hold approximately 10% of global surface freshwater, and host a wide variety of plant and animal species (Joosten and Clarke, 2002; Rydin and Jeglum, 2013c; Ribeiro et al., 2021). Historically peatlands were C sinks, however they are now under threat of becoming C sources due to anthropogenic pressures such as land use change, deforestation, and drainage (Rydin and Jeglum, 2013c; Joosten and Clarke, 2002; Fluet-Chouinard et al., 2023). Climate change may also impact peatlands, particularly those found in boreal regions where warming and a reduction in soil moisture can lead to increased peat drying (Canadell et al., 2021; Minasny et al., 2019).

Efforts to estimate peatland C stocks and model peatland processes in land surface models have been hindered by a lack of peatland data. Peat depth, surface area, bulk density, and organic C content values can be used to approximate peatland C stocks

(Minasny et al., 2019). Land surface models simulate fluxes of C, energy, and water. When integrated into an earth system model, these fluxes form the bottom boundary conditions for the atmosphere model. The addition of peatland dynamics to land surface models, and earth system models by extension, facilitates the investigation of peatland influence on climate and vice versa (Chadburn et al., 2022; Bechtold et al., 2019; Wu et al., 2016). In these models, the land surface is represented as a grid, thus information on peatland fractional coverage and peat depth is required in the same grid format (Wu et al., 2016; Melton et al., 2022). Traditional peatland mapping methods include air photo interpretation and field surveys (Minasny et al., 2019). However, field surveys can be expensive and labour intensive to undertake given the frequently remote and waterlogged nature of peatland environments (Rudiyanto et al., 2016; Minasny et al., 2019). Because of these limitations, existing observational peat data is often of insufficient accuracy and coverage (see Chapter 2) for application in regional or global C stock estimations and land surface modeling (Minasny et al., 2019; Melton et al., 2022). This is especially the case for peat depth, which cannot be approximated using classical optical remote sensing techniques (Krankina et al., 2008). Instead, peat depth is most commonly measured through probing, coring, and proximal remote sensing methods such as ground penetrating radar (Minasny et al., 2019).

Digital soil mapping techniques, including ML, can be used to estimate peat data at larger scales. Digital soil mapping methods for peatlands generally involve providing observational peat data and data on other environmental variables to some form of statistical model which can then be used to make predictions over an entire area (Minasny et al., 2019; McBratney et al., 2003; Rudiyanto et al., 2016). Environmental variables used in such cases can be referred to as predictors and would likely have been collected by proximal and remote sensing methods themselves. For making predictions of a peat property, predictors should be chosen based on the environmental signals of peat presence, drivers of peatland development, and the sensors capable of measuring such qualities (Minasny et al., 2019; Melton et al., 2022). The statistical model used for digital soil mapping can be simple, such as linear regression, or more complex, such as ML algorithms (Minasny et al., 2019; Rudiyanto et al., 2016; Melton et al., 2022).

ML has been applied to mapping peatland fractional coverage at a global scale and resolution suitable for use in land surface models. Melton et al. (2022) created Peat-ML, a spatially continuous global map of peatland extent on a five arcminute

grid (or about 0.0833°, which corresponds to 4.63 km at 60°N and 9.26 km at the Equator), using a gradient boosting decision tree algorithm called LightGBM (Ke et al., 2017). The algorithm was trained on existing regional maps of peatland extent and predictors of peat presence. When compared to peatland fractional coverage data that was iteratively withheld from training the model, the cross-validated Peat-ML output had a RMSE of 9.11% and a MBE of -0.29%, which demonstrates the model was able to reproduce the peatland test data reasonably well. Peat-ML was also compared to other peatland maps over the Boreal Plains region of Canada, for which an extensive amount of observational data is available through Ducks Unlimited Canada (Smith et al., 2007). Tarnocai et al. (2011), a traditionally mapped product developed through assessment of aerial photography and soil surveys, only slightly outperformed the cross-validated Peat-ML map in this comparison.

Efforts have been made to use ML to map peat depth as well, however further research can be done towards creating a robust global prediction. Hugelius et al. (2020) used a random forest (RF) ML algorithm to predict peat depths north of 23°N. Widyastuti et al. (2024) (currently in preprint) also used an RF approach to produce a global map of peat depth, C content, and bulk density called PEATGRIDS. To our knowledge, PEATGRIDS is the only global map of its kind, but its uncertainty assessment is limited. Here we adapt the Peat-ML modelling workflow, henceforth referred to as the Peat-ML Framework, to predict peat depth rather than fractional coverage and provide an extensive quality and uncertainty assessment of our peat depth results. This new modelling approach is referred to as PD-ML going forward. Chapter 3.1 contains the definition of peatlands we used as well as an explanation of our input datasets and modelling scheme. In Chapter 3.2 we discuss the output of PD-ML, our detailed model assessment, some preliminary C stock estimates made using our peat depth results, and the limitations of our chosen methods. Chapter 3.3 provides our final conclusions and Chapter 3.4 details the data availability of this work.

3.1 Materials and Methods

3.1.1 Definition of Peatlands

There are no internationally accepted definitions for peat and peatlands, however there are some broad explanations that encompass most of the varying criteria that

exist. We follow Joosten and Clarke (2002) and Lourenco et al. (2022) in defining peatlands as ecosystems that have accumulated a layer of peat at the surface, wherein vegetation may or may not be present (see Table 1.1). Peat is a type of soil composed largely of partially decomposed organic material (Lourenco et al., 2022). Lourenco et al. (2022) conservatively proposes that peat soil must have a minimum of 5% organic C content to a minimum depth of 10 cm. The database used for training PD-ML, Peat-DBase (Chapter 2), observed this 5% organic C content limit when establishing non-peat measurements, but did not place any restrictions on peat depth (see Table 1.1 and Chapter 2). This open-ended approach allows for the inclusion of a significant amount of peat depth data from a variety of peatland ecosystems. The variety is beneficial for ML as these algorithms are only capable of making sensible predictions in regions that exhibit similar behaviour to those represented in the data provided to them for training (Meyer and Pebesma, 2021). We did not enforce any specific thresholds on the PD-ML model itself such that it can predict a full range of peat depths from zero cm and deeper. However, when presenting figures of PD-ML output, we often plotted the data with different colours for depths deeper than 30 cm, as 30 cm is a common depth delineation in other peat datasets (see Table 1.1) (Loisel et al., 2017). See Lourenco et al. (2022) and Loisel et al. (2017) for further discussion on the variation and implications of different peat definitions.

3.1.2 Gathering and Preparing Data

While PD-ML follows a similar workflow of data input and model training to Peat-ML, there are notable differences (Figure 3.1). PD-ML operates on the same file format and five arcminute grid as Peat-ML, thus the training data and new predictors added for PD-ML were converted to netCDF files on the same grid (Melton et al., 2022). Most new predictors were first acquired as GeoTiff raster files and were adapted to the desired format using similar tools to Peat-ML (e.g. geospatial data abstraction software library (Rouault et al., 2023), Climate Data Operators (Schulzweida, 2022), NetCDF Operators (Zender, 2008)). Peat-DBase (Chapter 2), the database that formed the basis of the PD-ML training data, was originally in a point-based format within a CSV file and therefore required different processing methods. Additionally, a bootstrapping approach (see Chapter 3.1.2.1) was used to create several different training datasets for PD-ML (Johnson, 2001; Russell and Norvig, 2020; Hesterberg, 2011). These data processing steps are described in Chapters 3.1.2.1 and 3.1.2.2.

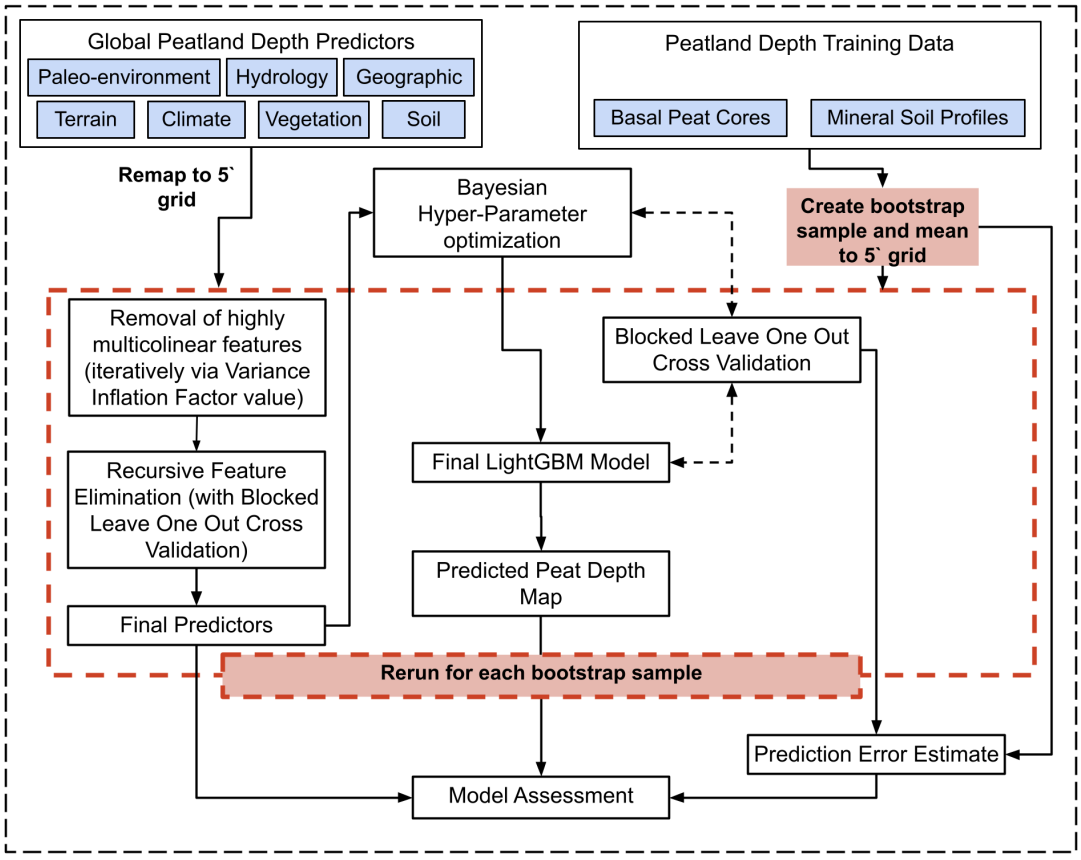

Figure 3.1: Flow chart of the PD-ML modelling process, adapted from the Peat-ML Framework (Melton et al., 2022). The steps and acronyms are explained in more detail in Chapters 1.1.2 and 3.1.

3.1.2.1 Training Data and Bootstrapping

Since the purpose of PD-ML is to predict peatland depths globally, extensive and spatially disparate peat depth data are required for training and testing the model. While ML models are capable of fitting complicated relationships, in a geospatial context their predictions are only reasonable for areas which are similar to those covered by their training data. Thus, when making predictions over new geographic areas that are dissimilar to their training data, the ML models will extrapolate leading to results which may be less trustworthy (Meyer and Pebesma, 2021). Thus, data from a variety of peatland and non-peatland regions is needed for PD-ML to produce reasonable spatially continuous results.

We used Peat-DBase version 1.0 (see Chapter 2) as the foundation of our training and testing data. It is the largest database of peat core and probe measurements with global coverage that we are aware of. Additionally, Peat-DBase includes a large number of non-peat soil cores as well. Nonetheless, there are aspects of Peat-DBase that must be kept in mind when modelling with it. Specifically, there are less peat depth data available for the tropics, non-peat data is missing over desert regions, and sampling biases may impact the distribution of depth within the database (see Chapter 2.3.2). Other notable data gaps include northern and central Eurasia, much of southeastern South America, and New Zealand.

The Peat-DBase point-based data was gridded to match the format expected by the model (Chapter 3.1.2). The measurements within Peat-DBase are represented as single numerical depth values in centimetres at locations given by latitude-longitude coordinates (World Geodetic System 1984 coordinate system). We first filter the database to remove any potential duplicate or erroneous measurements using the flags provided within the dataset (see Chapter 2.1.2 to 2.1.3). The resulting data was converted to the five arcminute grid by calculating the mean of the measurements that occurred in each grid cell. For grid cells that contained measurements with a depth of zero cm and measurements with a depth greater than zero cm, measurements with a depth of zero cm were removed prior to calculating the mean for the grid cell. This was done to ensure that the resulting grid-based dataset represented the mean depth of peatlands within each grid cell, rather than the mean depth of peat across all environments in the grid cell. Figure 3.2 shows the final gridded version of Peat-DBase.

As PD-ML requires a single value per grid cell, we take the mean of peat depth

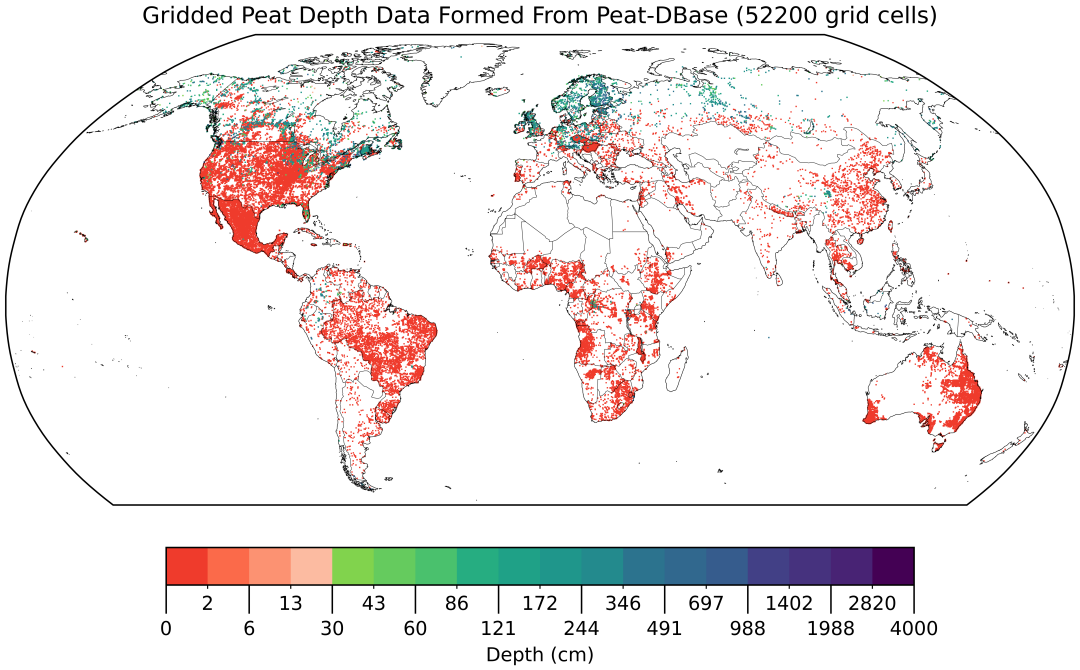

Figure 3.2: Geographical distribution of the gridded Peat-DBase version 1.0. Any area without a coloured grid cell does not contain any data. The removal of zero cm depth measurements from grid cells with non-zero cm depth measurements as described in Chapter 3.1.2.1 means the gridded data represents the mean peat depth of the peatlands within a grid cell, rather than the mean peat depth of the entire grid cell area. Note that the coloured grid cells have been increased in size for easier viewing. The colour bar is presented in a log scale with a colour break at 30 cm to represent a commonly used threshold when classifying an area as peatland (see Table 1.1) (Loisel et al., 2017).

measurements within a grid cell. Figure 3.3 shows that this shifts the values we provide to the model for training compared to what is seen in a database of individual depth measurements. A distinct change from Figure 3.3a to 3.3b is the loss of some shallow peat depth values. By taking the mean of the depth measurements within a grid cell, the shallowest and the deepest depth values are replaced by moderate mean values. However, this distribution will shift across the different training datasets produced under our bootstrapping scheme (Figure 3.3c) which is discussed later in this section. Regardless of the bootstrapping, grid cells with depths of zero cm remain the most numerous values overall. This zero-inflation within the dataset is notable from a modelling perspective, as these values make up the majority of the information the model will draw upon.

Meyer and Pebesma (2021) propose a method for assessing where a model is making reasonable predictions informed by relationships learned from the training data, rather than extrapolating. They refer to these learned regions as the area of applicability (AOA). Their approach uses a 'dissimilarity index' which measures the distance to the nearest training data point within a multidimensional space built from the predictors used by the model. For the most accurate representation of the model AOA, the distance is found to the nearest training data point that is not within the same cross-validation block (see Chapter 1.1.2, 3.1, and Melton et al. (2022) for information on cross-validation). A dissimilarity index threshold is determined by finding the maximum dissimilarity index between training grid cells. Any grid cells with a dissimilarity index exceeding this threshold are considered outside of the AOA and therefore considered beyond the model's knowledge, potentially leading to model extrapolation.

Early tests with PD-ML and AOA demonstrated the importance of including data in the desert regions as the model was otherwise poorly constrained there. Thus, data was added to account for desert and xeric shrubland biomes that would likely have little to no peat presence. The shape files of these ecoregions are from Olson et al. (2001), the same product used in the Peat-ML training dataset (Melton et al., 2022). All grid cells within these biomes are set to a peat depth of zero cm. A random selection of five percent coverage of the biome was found to give a reasonable AOA and not significantly impact model performance outside of these regions.

The final version of gridded observed peat depth data used for training and testing is shown in Figure 3.4 (and the depth distribution in Figure 3.3b). It should be noted that the inclusion of data for the desert and xeric shrubland areas increases the zero-

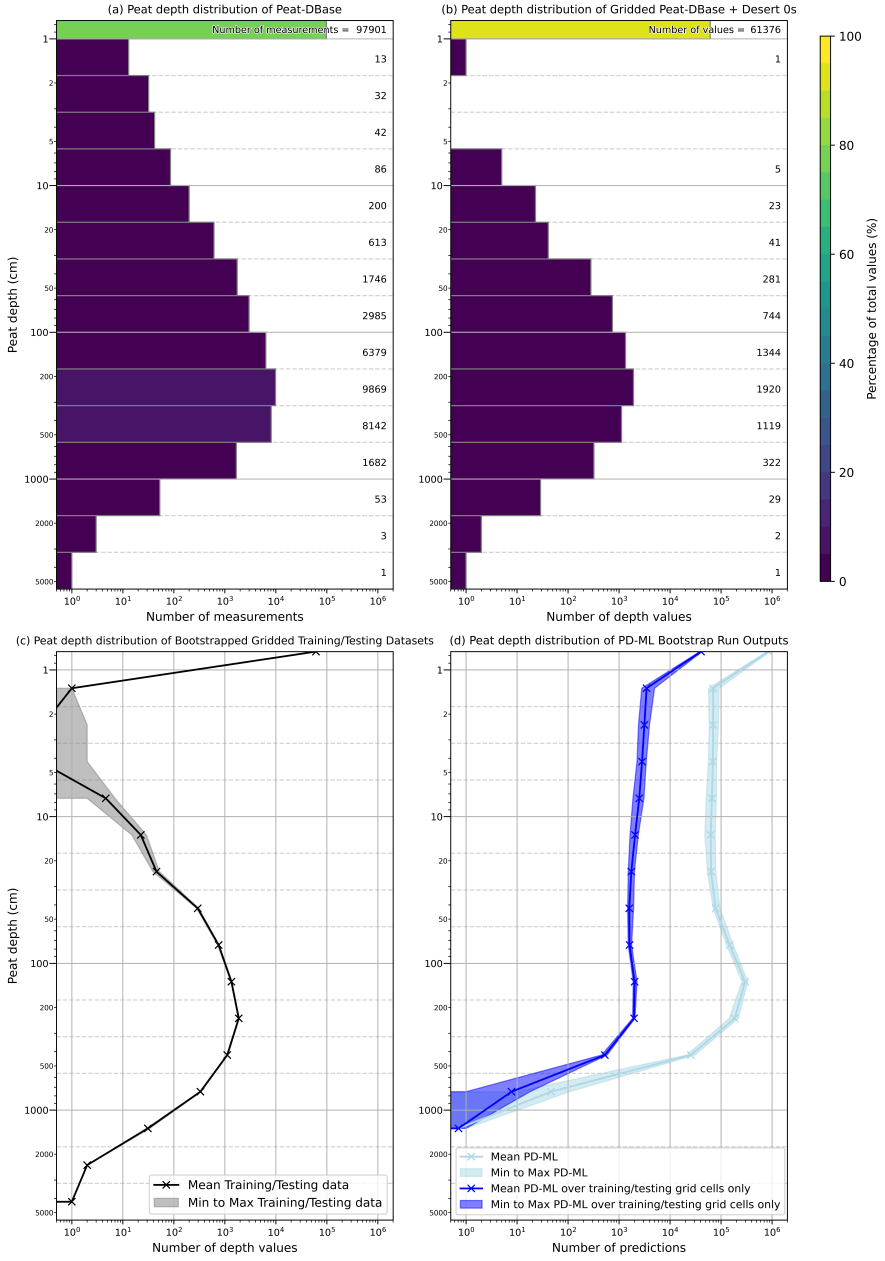

Figure 3.3: Histograms showing peat depth distributions across different iterations of the training datasets. (a) The peat depth distribution of Peat-DBase (see Figure 2.3 in Chapter 2). (b) The peat depth distribution of Peat-DBase when gridded, with the addition of desert data (see Chapter 3.1.2.1). (c) The peat depth distribution of the training and testing datasets made from bootstrapping Peat-DBase (see Chapter 3.1.2.1). (d) The peat depth distribution of the PD-ML bootstrap model run outputs (see Chapter 3.2), where the dark blue lines indicate only the output values for grid cells that contained training and testing data. Panels c and d present the mean distribution as a solid dark line and the minimum to maximum range of the distribution in a lighter band. Note that the axes of all panels are presented in log scales, which can inflate the prominence of smaller values.

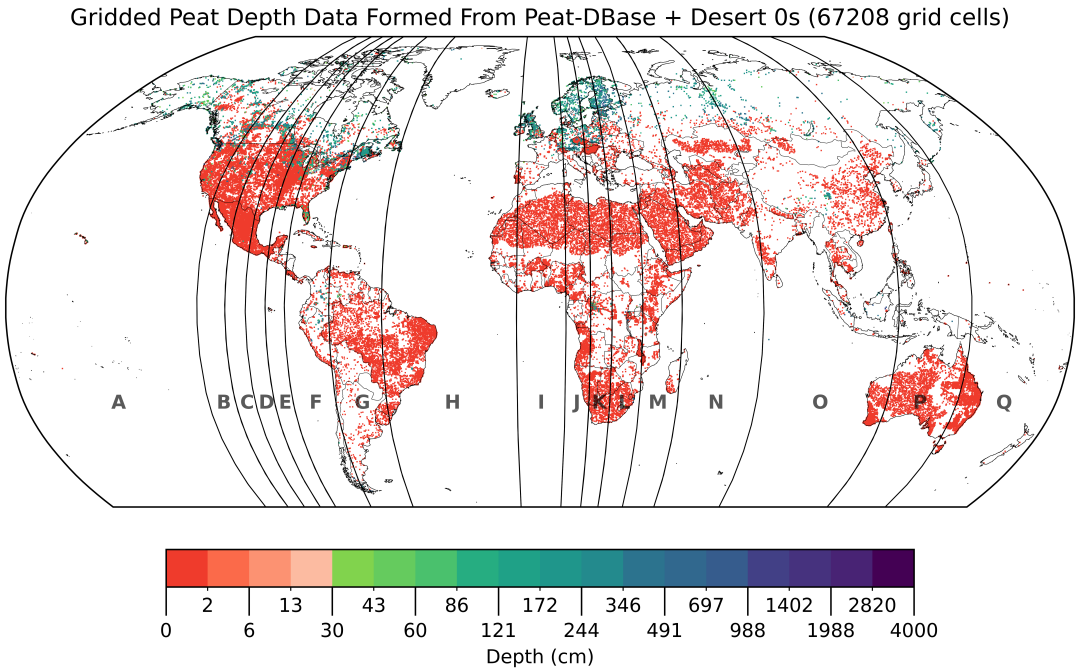

Figure 3.4: The gridded PD-ML training dataset, with desert data processed and combined. Any area without a coloured grid cell does not contain any data. The removal of zero cm depth measurements from grid cells with non-zero cm depth measurements as described in Chapter 3.1.2.1 means the training and testing data represents the mean peat depth of the peatlands within a grid cell, rather than the mean peat depth of the entire grid cell area. Note that the coloured grid cells have been increased in size for easier viewing. The colour bar is presented in a log scale with a colour break at 30 cm to represent a commonly used threshold when classifying an area as peatland (see Table 1.1) (Loisel et al., 2017). The letters indicate the blocks established for use in cross-validation (see Chapter 1.1.2, 3.1.3, and Melton et al. (2022)).

inflation of the overall dataset. Data in a total of 15 008 new grid cells were added, which amounts to 22.3% of the dataset. This seems reasonable, however, as a recent classification of biomes suggests deserts and semi-deserts make up approximately 20% of global land area (Loidi et al., 2023). Grid cells with a non-zero cm depth now make up 8.7% of all grid cells within the training dataset.

Cores collected in a peatland may not accurately capture the depth of the entire peatland, particularly when there is significant variation in the peat surface and underlying basin topography. Taking the mean of core depths and assigning the resulting value to represent the mean peat depth of an entire grid cell perpetuates such uncertainties, especially for grid cells that contain only one core. Additionally it

is possible to have multiple separate peatland complexes within the area of one grid cell, while only having cores from one of the peatlands. To test how such uncertainties in the data would affect PD-ML, we simulate sample variability within a grid cell through a bootstrapping approach. Our use of bootstrapping is somewhat unique as this technique is more commonly used to make subsamples from the entire input dataset, rather than to vary some of the values within the dataset as we are (Galdi and Tagliaferri, 2019; Johnson, 2001; Russell and Norvig, 2020; Hesterberg, 2011).

For each grid cell containing more than one non-zero cm peat depth measurement, we subsampled the measurements with replacement, such that the total number of measurements within the grid cell remained the same but the new collection was a subset of the original data with some being duplicates (see Figure 3 in Galdi and Tagliaferri (2019)). The depth of the peat in the grid cell was set to be the mean of this new subsample. A total of 5829 grid cells had non-zero cm measurements within them, with 3960 of those grid cells only containing one such measurement (Figure 3.5). Thus this bootstrapping approach varies the depth values of about 32% of the non-zero cm grid cells. We performed 400 bootstrapping iterations, for a total of 401 training and testing datasets (when including the gridded dataset in Figure 3.4 where no bootstrapping was applied). This collection of training and testing data is also occasionally referred to as observed data hereafter.

While more bootstraps are generally desirable (Efron and Tibshirani, 1994; Hesterberg, 2011), we were limited in the number of times we could run PD-ML due to computational cost. Additionally, over half of the grid cells being bootstrapped had less than five measurements within them and therefore have repeating subsamples within the 401 iterations (Figure 3.5). The latitudinal distribution of the grid cells with the most measurements is roughly even across high and low latitudes.

3.1.2.2 Predictor Data

We used the predictors originally collected for Peat-ML, while also providing new datasets to PD-ML (Table 3.1). The majority of the Peat-ML predictors are categorised as climate, soils, terrain, and vegetation. The environmental variables that act as drivers or indicators of peat occurrence may be similar, but do not necessarily have the same predictive power for peat depth. For example, particular kinds of vegetation which are specialised for surviving in wet, low-nutrient environments would be strong indicators of peatland presence, but could not inherently reveal as much about

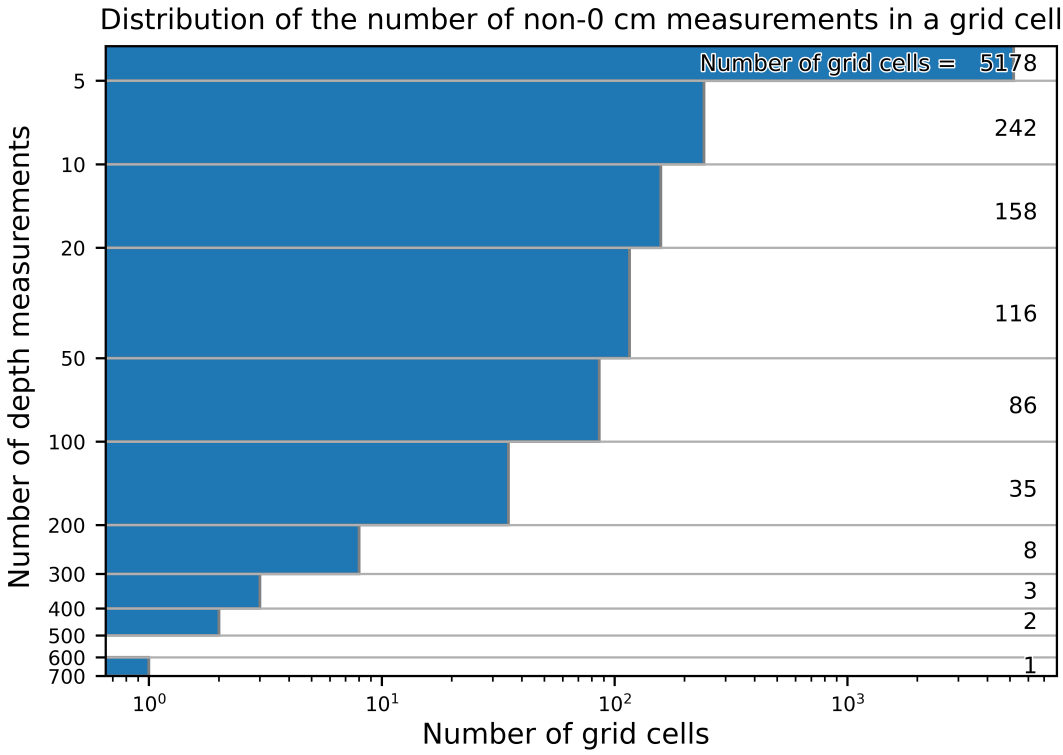

Figure 3.5: The distribution of the number of non-zero cm depth measurements from Peat-DBase that are within the same grid cell. There are 3960 grid cells that only contain one such measurement.

peat depth (Minasny et al., 2019). While peatlands are widely varied in their geometry and vegetation, they all developed because organic matter deposition exceeded decomposition rates for a significant time period, usually in waterlogged conditions (Gorham, 1957; Moore, 1989). Thus, we sought to include hydrological predictors, i.e. water table depth, surface water recurrence, and surface water seasonality, that may indicate an area is consistently saturated enough to routinely enable peat accretion. We also provided paleo-environmental information, i.e. years since exposed from ice and sea, although this was more challenging to facilitate since this modelling framework does not support temporal datasets. It should be noted that the predictor datasets have their own uncertainties associated with them which would impact PD-ML (Meyer and Pebesma, 2022). These influences may be broadly captured by the bootstrapping (see Chapter 3.1.2.1 and 3.2.1), but the predictor uncertainties are not immediately separable from the training data uncertainty.

All datasets chosen for use as predictors which originally had a temporal dimension needed to be aggregated in some way. Melton et al. (2022) found that using the seasonal means of September–November (SON), December–February (DJF), March–May (MAM), June–August (JJA) for the climate predictors did not yield strong differences in their results compared to using annual minimums and maximums, despite these seasons potentially being suboptimal representations of tropical regions. We elected to include both of these aggregation types as well as some additional climate predictors which may have more universal applicability (e.g. growing degree days) or more specified applicability for the tropics (e.g. monsoon intensity) (Wang and Ding, 2008).

We chose to exclude the organic C content and bulk density predictors from PD-ML. These predictors are derived from products made using ML algorithms which use similar predictors to those collected for Peat-ML (Hengl et al., 2017; Hengl and MacMillan, 2019; Melton et al., 2022). Thus, using these products as input to PD-ML may be circular in nature, as a model would be making predictions based on a model output of a similar target variable.

Table 3.1 outlines the predictors provided to PD-ML and indicates which ones were new additions for our framework. We note the original formats and calculations for the predictors we gathered below, refer to Melton et al. (2022) for explanations of the others.

Table 3.1: The environmental variable datasets provided as possible predictors of peat depth to PD-ML. The italicised variables are the new predictors collected for PD-ML. The bold variables are the predictors that were selected by any of the bootstrap model runs (see Chapter 3.2.1). See Table 1 in Melton et al. (2022) for more details on predictors originally collected for Peat-ML.

Type	**Source and original resolution (time period)**	**Predictors**
Climate	TerraClimate (Abatzoglou et al., 2018) 1/24° (1985–2015)	Actual evapotranspiration, **climate water deficit, soil water**, potential evapotranspiration (Penman–Monteith), **precipitation accumulated, downward surface shortwave radiation, snow water equivalent, runoff, Palmer Drought Severity Index (PDSI), minimum temperature**, maximum temperature, **vapour pressure, vapour pressure deficit, 10 m wind speed**
	CHELSA-BIOCLIM+ (Brun et al., 2022) 1 km (1981–2010)	*Climate moisture index, growing season length, growing degree days (5℃),* ***growing season precipitation****, growing season temperature*

Continued on next page

Type	**Source and original resolution (time period)**	**Predictors**
	WorldClim 2 Bioclimatic Variables (Fick and Hijmans, 2017) 1 km (1970–2000)	*Isothermality, temperature seasonality, mean temperature of driest quarter, temperature of warmest quarter,* ***precipitation seasonality****, precipitation of driest quarter, precipitation of warmest quarter*
	Calculated from TerraClimate (Abatzoglou et al., 2018) 1/24° (1985–2015) precipitation data based on the methodology of Zeng and Zhang (2020) and Wang and Ding (2008)	***Monsoon intensity****, monsoon intensity masked by monsoon domain*
Soils	Open Land Maps (Hengl, 2018) 250 m (–)	**clay content, sand content**, soil water content, at field capacity (33 kPa)
	SMAP (ONeill et al., 2021) 9 km (2015-2023)	***Soil moisture mean****, soil moisture ratio below 50% (calculated),* ***soil moisture ratio above 50% (calculated)***

Continued on next page

Type	**Source and original resolution (time period)**	**Predictors**
Terrain	Geomorpho90m (Amatulli et al., 2020) 250 m (–)	**Slope, aspect, eastness, northness, convergence index**, compound topographic index (topographic wetness index), **stream power index, first and second directional derivatives (east–west, north–south), profile curvature, tangential curvature**, elevation standard deviation, **geomorphology landform, roughness indices, topographic position index, maximum elevation deviation**
Vegetation	PALSAR/PALSAR2 (Shimada et al., 2014) 25 m (2007–2010)	**Horizontal transmit and Horizontal receive and Horizontal transmit and Vertical receive polarisation backscattering coefficients**
	MOD17A3 V055 (Running et al., 2011) 1 km (2000–2015)	**Net primary productivity**

Continued on next page

Type	**Source and original resolution (time period)**	**Predictors**
	S-NPP VIIRS vegetation indices (VNP13A1) (Didan and Barreto, 2018) 500 m (2012–2019)	**Three-band Enhanced vegetation index (EVI), Two-band EVI (using only red and NIR band), near-infrared radiation (NIR)**, shortwave infrared radiation reflectance (SWIR) 1 (1230-1250 nm), **SWIR2 (1580-1640 nm), SWIR3 (2225-2275 nm), normalised difference vegetation index (NDVI)**, NIR reflectance, **green reflectance, blue reflectance, red reflectance**
	MODIS Global Vegetation Phenology (MCD12Q2 V6 Land Cover Dynamics) (Friedl et al., 2019) 500 m (2001–2018)	**Dormancy**, EVI_Amplitude, **EVI_Area**, EVI_Minimum, **Greenup**, Maturity, **MidGreendown, MidGreenup, Peak, Senescence**
	MODIS Terra+Aqua (Wang, 2021) 0.05° (2002-2023)	***Photosynthetically active radiation***

Continued on next page

Type	**Source and original resolution (time period)**	**Predictors**
	SMAP (ONeill et al., 2021) 9 km (2015-2023)	***Vegetation Water Content***
Hydrology	*2020 update to Fan et al. (2013) 30 arc-seconds (about 1 km) (2004-2014)*	***Water table depth***
	(Pekel et al., 2016) 30 m	***Surface water recurrence (1984-2021), surface water seasonality (2021)***
Geographic	Calculated	Length of the longest day of the year in hours
Paleo-environment	*Calculated from PaleoMIST 1.0 (Gowan et al., 2021) 1/4° (past 80,000 years)*	***Years since exposed from ice and sea***

Our additional climate predictors were sourced from CHELSA-BIOCLIM+ (Brun et al., 2022, last access: 30 May 2023), WorldClim 2 (Fick and Hijmans, 2017, last access: 25 July 2023), and TerraClimate (Abatzoglou et al., 2018, last access: 11 September 2023 via Google Earth Engine (GEE) (Gorelick et al., 2017)). CHELSA-BIOCLIM+ (climatologies at high resolution for the Earth's land surface areas – bioclimatic variables plus) is a collection of climate-related time series at a one kilometre resolution developed by applying statistical downscaling and a variety of calculations to climatologies primarily from CHELSA V2.1 (Karger et al., 2017, 2021) and ERA5 (Hersbach et al., 2020). WorldClim 2 is a set of one kilometre resolution monthly climate datasets created from weather station data that underwent interpolation informed by Moderate Resolution Imaging Spectroradiometer (MODIS) satellite data. The WorldClim 2 Bioclimatic variables are annual trends calculated from the monthly rainfall and temperature related WorldClim 2 datasets (Fick and Hijmans, 2017). The TerraClimate data is summarised in Melton et al. (2022), however for PD-ML we reacquired the original TerraClimate precipitation dataset to calculate monsoon intensity predictors. We calculate monsoon intensity using the method described in Zeng and

Zhang (2020) but take the absolute value to allow consistency over the equator. We made an additional version of the monsoon intensity predictor which is masked by the global monsoon precipitation domain, an area delineated using the definition set out by Wang and Ding (2008).

The new soil predictors were developed from Soil Moisture Active Passive (SMAP) (ONeill et al., 2021, last access: 20 September via GEE) data. SMAP is a nine kilometre resolution daily composite soil moisture product that is calculated from interpolated observations collected by the SMAP L-Band radiometer. We also calculated the ratio of daily soil moisture measurements with a value below 50% of the soil moisture mean and the ratio above 50% of the mean as additional predictors. It should be noted that SMAP measurements are not taken for areas that are covered in water or frozen, so the northern hemisphere has significant gaps for its winter months in the SMAP product. Given that we took the mean of all measurements to remove the temporal dimension, these gaps do not persist in the SMAP datasets provided to the model. However, it should be acknowledged that the resulting datasets may be less certain in the affected regions.

We added new vegetation predictors derived from MODIS Terra+Aqua Photosynthetically Active Radiation (PAR) (Wang (2021), last access: 25 June 2023 via GEE) and SMAP data. MODIS Terra+Aqua PAR is a three hourly 0.05° resolution dataset. Wang (2021) calculate PAR from surface reflectance, which was determined from multi-temporal MODIS signatures, and top-of-atmosphere radiance and reflectance values. The original SMAP product includes a vegetation water content variable which serves as an input to the soil moisture calculations. Vegetation water content is calculated primarily from a MODIS Normalised Difference Vegetation Index (NDVI) product (ONeill et al., 2021).

Our hydrology predictors were processed from the 2020 update to the work of Fan et al. (2013), and Pekel et al. (2016) (last access: 18 July 2023 via GEE). Fan et al. (2013) produced thirty metre resolution annual and monthly water table depth datasets by compiling well site observations and performing gap filling with a groundwater model forced by terrain, sea level, and modern climate. Pekel et al. (2016) developed thirty metre resolution global surface water datasets of varying temporal resolutions by applying evidential reasoning, expert system, and visual analytic techniques to Landsat 5, 7, and 8 satellite images. Melton et al. (2022) also considered using these global surface water products, but elected against it due to the shortcomings of Landsat products for treed and small peatlands. In this case, we have

chosen to include them due to the importance of waterlogging in the development of peatlands (Page and Baird, 2016; Joosten and Clarke, 2002; Rydin and Jeglum, 2013b). The predictor filtration steps of the model framework can then exclude these predictors if they do not provide sufficient information (Chapter 1.1.2 and Melton et al. (2022)).

When searching for paleo-environmental data that may be informative to a peat depth model, ice and sea coverage was of particular interest. The retreating ice sheets of the Last Glacial Maximum paced the northward expansion of what is now many modern day peatlands in regions such as Canada and the West Siberian Lowlands. Meanwhile tropical peatlands were more impacted by sea level change and subsequent alterations to regional hydrology (Treat et al., 2019). Thus, the time at which an area was exposed from the ice and sea may act as a boundary condition that could be informative to the model. To facilitate this, we derived our paleo-environmental predictor from PaleoMIST 1.0 Gowan et al. (2021, last access: 29 May 2023). Paleo margins, ice sheets, and topography (PaleoMIST 1.0) is a 0.25° resolution product going back 80 000 years in 2500 year time steps which was developed using reconstructed sea levels and a three-layered Earth model. From the PaleoMIST 1.0 ice thickness and paleotopography variables we find the time steps at which each grid cell is ice and sea free. For areas that have been exposed for the entire duration of the product, we assign a value of -82 500 to be one step older than the maximum age and for areas still covered, we assign a value of 2500 to be one step younger than the minimum age.

3.1.3 Adjustments to Peat-ML Framework

The ML algorithm, parameter optimisation, cross-validation method, and predictor selection process have been established in the Peat-ML Framework, however adaptations were needed for this modelling approach to be applicable to peat depth (see Chapter 1.1.2 for an explanation of the modelling workflow). The first step was to allow the model to produce values of zero and up, rather than capping results at a maximum of 100. Next, the geographic blocks used in the cross-validation process must be updated for the peat depth dataset now being used as model input. Like Melton et al. (2022), we established our minimum block sizes by finding the distance at which spatial autocorrelation ended, as determined by calculating the Moran's I (MI) of our model residuals at different lag distances (see Chapter 1.1.2 for an explanation

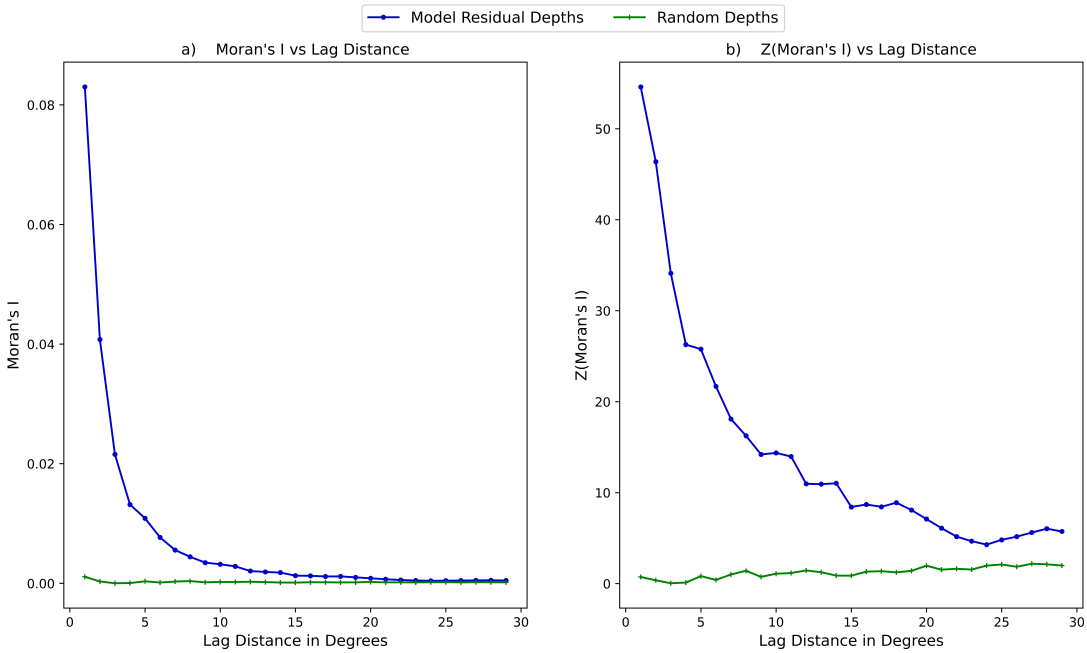

Figure 3.6: (a) MI as a function of increasing distance in degrees. (b) Z-score of MI as a function of increasing distance in degrees. In both cases, calculations are performed for the PD-ML model residuals and a generated dataset of random numbers (included to demonstrate how a dataset with no spatial autocorrelation would behave).

of spatial autocorrelation). MI describes how similar a variable at one location is to neighbouring locations. Positive MI values indicate that similar variable values are clustered together while a negative MI indicates that dissimilar values are clustered together. The closer the MI value is to zero, the less clustering, and therefore spatial autocorrelation, is present (Getis, 2010). We also calculated the z-score of our MI results to establish their statistical significance, with values closer to zero indicating low significance (Getis and Ord, 1992). Figure 3.6 shows that by a distance of about 6°, the MI of the model residuals is nearing zero and the z-score is also decreasing sharply. Both the MI and z-score decrease further for distances over 6°, however 6° was chosen as the minimum block size to help balance the lack of uniformity in the training data distribution such that as many blocks as possible have access to a wide distribution of peatland types (i.e. boreal and tropical). Additionally, while the model residual z-scores do not approach the same lows as a sample dataset of random numbers, they correspond to low model MI values and are therefore deemed not of significant concern.

Our result of 6° is lower than the 10° determined for Peat-ML. However, when we

tested our MI algorithm on residuals of Peat-ML calculated using a version of Peat-ML training data provided to us by Melton et al. (2022), the output was similar to that of our model residuals. We therefore hypothesise that the difference may be due to variations in how the function used for Peat-ML calculates MI or due to the Peat-ML MI calculations using earlier iterations of their data. We further corroborated our results for PD-ML using Geary's C and G statistics, which are also used for assessing spatial autocorrelation (Getis, 2010; Getis and Ord, 1992; Ord and Getis, 1995). Using 6° as a minimum block size then distributing the gridded observed data evenly across each block produces the 17 blocks shown in Figure 3.4. These blocks are then used in the cross-validation steps to delineate which grid cells are used for training and which ones are used for testing (see Chapter 1.1.2 for more on cross-validation).

Early tests of PD-ML showed the zero-inflated nature of the peat depth training data was likely impacting the model's ability to predict deeper peats, thus we tested a variety of methods to address this. By default, the loss or error function used by LightGBM is mean square error (Ke et al., 2017), with RMSE being used for the cross-validation process (Melton et al., 2022). We adapted this to a custom scoring method wherein the model error for zero cm depth test grid cells are weighted less than the error of non-zero cm depth grid cells in the cross-validation process. This is a strategy seen in binary classification models when confronted with skewed datasets (Krawczyk, 2016; Russell and Norvig, 2020). We used weights of 30% for zero cm depth grid cells and 70% for non-zeros, such that while non-zero cm depth grid cells make up only 8.7% of the observed data, the model error calculated for such locations was weighted more heavily (we previously tested a 50-50 split and only saw a slight difference in model behaviour compared to that of the 30-70 split). The model successfully predicted deeper depths using the custom loss function. However, it was often less than 20 cm different on average when analysing results within known peatland regions. The custom metric also had a negative impact on model performance in grid cell level comparisons, but this was expected when a large amount of the grid cells being tested against have a depth of zero cm which we were now deliberately deprioritizing.

Additionally, we tested data preprocessing techniques that sought to reduce the skew in the training data prior to use in the model, i.e. log transformation (Feng et al., 2014; West, 2022), Box-Cox, and Log-sinh (Huang et al., 2023). We decided against retaining a transformation in the PD-ML framework as they did not result in appreciable improvements and added considerable computational cost to the modeling

process.

To summarise our final modelling approach, we ran our model several times over the bootstrapped training datasets and aggregated the results to produce our final output. As discussed in the previous section, we produced 401 observed dataset versions for training and testing by using bootstrapping to vary the depths in some grid cells. We ran PD-ML for each bootstrapped dataset such that for each bootstrap we had the corresponding model output, cross-validation output, and the list of predictors used by that model run to produce said output. We then took the mean of the depth predicted over all the bootstrap model run outputs in each grid cell to produce a final PD-ML product (discussed in Chapter 3.2.2).

3.2 Results and Discussion

3.2.1 Predictor Importance

Across the 401 PD-ML bootstrap model runs, 123 different predictors were selected by one or more of the runs (see the bold predictors in Table 3.1 for all predictors that were selected in some form). However, only 10 predictors were selected by more than 75% of the runs, indicating that many of these predictors were not consistently of importance to the model. Figure 3.7 shows the top 15 predictors selected by averaged importance across all runs (see Chapter 1.1.2 for an explanation of the predictor selection steps). The predictor with the highest average importance is the average vapour pressure deficit (VPD) over SON. VPD SON is selected for use in 87% of the bootstrap model runs and when it is chosen, it is always among the top two most important predictors in that particular run. The second most important predictor is the average snow water equivalent (SWE) over DJF, it is selected in 62% of the model runs with a more even distribution of importance relative to VPD SON (Figure 3.7). No other predictors approach the same level of importance as seen with VPD SON and SWE DJF, except for the overall mean VPD which, when selected, frequently has an importance greater than 10%, but is only selected in 26% of the runs and therefore has a lower average importance. Beyond these three predictors, most range from having an importance of 0% to about 10%. No predictor was selected for all 401 bootstrap model runs, although the standard deviation of short wavelength infrared (2225-2275 nm) (SWIR3) MAM was close with a presence in 99% of runs, followed by Runoff SON and an indicator of geomorphological landforms (Geomorphon) which

were both selected 96% of the time.

Of the top 15 predictors, 10 relate to climate, two to vegetation, two to terrain, and one is the newly added paleo-environment predictor (see Table 3.1 for predictors and categories). In their analysis of factors driving peat formation, Minasny et al. (2019) list climate as being among the most important at the global scale, along with vegetation and topography. PD-ML therefore is capturing the broad elements of peatland processes, even in the presence of strong uncertainty due to variability in the training data. SWE DJF, Runoff SON, and Geomorphon were among the top 10 predictors for Peat-ML as well. PD-ML and Peat-ML chose several of the same general predictor types (i.e. the season or aggregation type selected may vary) like downward surface shortwave radiation, SWIR3, soil water, and wind speed (see Figure 3 of Melton et al. (2022)). Compared to Peat-ML, PD-ML appears to favour climate predictors as being the most important rather than terrain, thus peat depth and distribution may rely on somewhat different environmental variables and relationships.

The bootstrapping in PD-ML has revealed the sensitivity of the model predictions and predictor selection to the training data. PEATGRIDS (Widyastuti et al., 2024) and Hugelius et al. (2020) chose 19 and 12 environmental variables respectively to use as predictors in their ML peat depth modelling initiatives. Many of the assembled variables for these products are similar to each other and to those used in PD-ML (e.g. aggregations of temperature and precipitation data, and representations of topography - see Table 1 in Widyastuti et al. (2024) and Table S2 in Hugelius et al. (2020)). However, neither PEATGRIDS nor Hugelius et al. (2020) apply any predictor filtration steps as is done in PD-ML, which uses VIF filtration and RFE (see Chapter 1.1.2 and Figure 3.1). Predictor filtration is a common step in ML processes as it helps reduce the chance of the model overfitting to its training data, which would prevent it from being able to make reasonable predictions beyond the areas it has seen in training (Hawkins, 2004; Peng and Nagata, 2020; Russell and Norvig, 2020). Thus, there is a risk of these other peat depth models being overfit and their presentation of a single model instance may not be robust, given the fluctuation in predictor use we have demonstrated in response to changes in the training data. Additionally, by preemptively choosing a specific set of predictors for the model to use, PEATGRIDS and Hugelius et al. (2020) may be using suboptimal predictors. In our approach, we have allowed the model to choose from a large number of potential predictors and therefore optimise which predictors are selected.

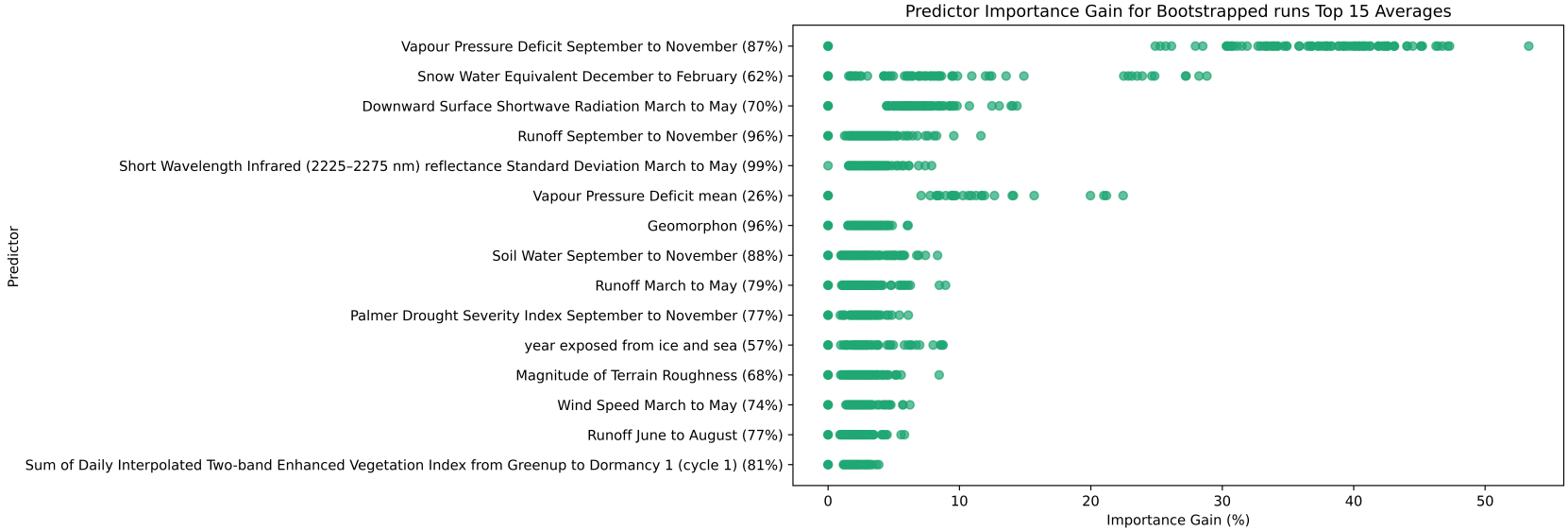

Figure 3.7: Importance of top 15 predictors based on information gain in percent, ordered based on the average importance of the predictors across all bootstrap runs. Each green dot represents the importance of the associated predictor within a single bootstrap run. When a bootstrap run did not select the listed predictor, it is treated as having an importance of 0%. The number listed next to each predictor records the percentage of bootstrap runs that selected that predictor.

Melton et al. (2022) note that it can be enticing to consider whether their modelling approach for Peat-ML can provide new information on the processes of peat development and persistence, but that it is difficult to delineate whether the chosen predictors have a relationship of cause or effect with peatlands. This is also the case for PD-ML. However, we further emphasize that it is precarious to make assumptions about predictor-peat process relationships based on our model due to its demonstrated sensitivity to uncertainties in the observations used for training. Nevertheless, we can still make some broad observations about the overall PD-ML model behaviour. Frequent to continuous waterlogging, whether due to precipitation, groundwater, or some combination of these, is an important factor in the initiation and persistence of peat accumulation (Page and Baird, 2016). Figure 3.7 shows that, with the exception of downward surface shortwave radiation and wind speed, all climate predictors in the top 15 are related to water or moisture presence. PD-ML could be selecting VPD SON in particular (and mean VPD potentially as well) as a means of delineating where shallow or non-peatland areas are located. Given the strong zero-inflation of our training data, the model is likely finding significant success in its cross-validation steps when choosing predictors that are informative for such values, even with our custom scoring metric. SWE DJF may be chosen by PD-ML to separate boreal and temperate peatlands from tropical peatlands, as was suggested for Peat-ML (Melton et al., 2022). Peat-ML also had SWIR3 among its most important predictors and suggested it may be assisting in differentiating between wet and dry earth as well as identifying fens. The standard deviation of SWIR3 MAM is the most selected predictor across the bootstraps in PD-ML. PD-ML may be using this version of SWIR3 to assess for areas that experience too much moisture fluctuation to accumulate deep peats as low moisture conditions can lead to increased respiration which discourages peat growth. PD-ML also frequently selects Runoff SON, which could act as a form of rainfall indicator for coastal and tropical peatlands (Ratnayake, 2020; Page and Baird, 2016). Geomorphon is another common predictor between Peat-ML and PD-ML. In Peat-ML it is suggested that it provides information on the topographical characteristics that are conducive to peat development (Melton et al., 2022). In PD-ML, it may assist in indicating depth as well as location, however the growth of peat could alter the geomorphological aspects of an area, as noted for Peat-ML.

3.2.2 Predicted Peat Depths and Trustworthiness

The global peat depth averaged across all PD-ML bootstrap runs is shown in Figure 3.8, this output is referred to as the PD-ML Mean Product (modelled mean) moving forward. The modelled mean indicates peat depths exceeding 30 cm in most of the major peatland regions predicted in Peat-ML (Figure 1.1) including parts of Canada, the Congo Basin, parts of Scandinavia and the Northern European Plain, the West Siberian Lowlands, and areas in the Malay Archipelago. Of course, PD-ML and Peat-ML are providing datasets of different variables, therefore any comparison between them is indirect at best. There are some differences between the modelled mean and Peat-ML, with one of the most significant being the modelled mean's widespread peat accumulation across the boreal and temperate regions of the Northern Hemisphere where Peat-ML indicates less coverage. We also see that the modelled mean has predicted non-zero cm peat presence across far more of the Malay Archipelago and New Zealand relative to Peat-ML. Both Peat-ML and the modelled mean indicate peat presence in the Amazon basin and the southern tip of South America, however the shape of these complexes differ between the two maps. The modelled mean tends to position more peat along the Andes in particular. The modelled mean indicates peat depths approaching 30 cm in some coastal areas surrounding the Gulf of Mexico while Peat-ML indicates greater coverage there; with this still being an indirect comparison of different peatland variables. The Tibetan Plateau also experiences different peat representations between the two products with the modelled mean predicting deep peat more in the east and along the Himalayas. The modelled mean predicts deeper peat over the Caucasus mountains whereas Peat-ML indicates only a small amount of peat coverage to the south of this region.

Some data gaps are present in PD-ML outputs due to some predictor datasets missing data themselves. Figure 3.8 shows that the modelled mean is missing data for areas in northern Canada such as Victoria Island, parts of the Barren Grounds, and Baffin Island, as well as all of Greenland, and parts of the Novoya Zemlya archipelago in Russia. The model is unable to make a prediction for a grid cell that does not have all provided predictors present. Gaps that persist in the modelled mean are the result of gaps present in the same location throughout every bootstrap run. Of the grid cells with data in the modelled mean, only 0.5% of them were missing data in a fraction of the bootstrap outputs.

Recall that not all areas of the modelled mean (or any single bootstrap run result)

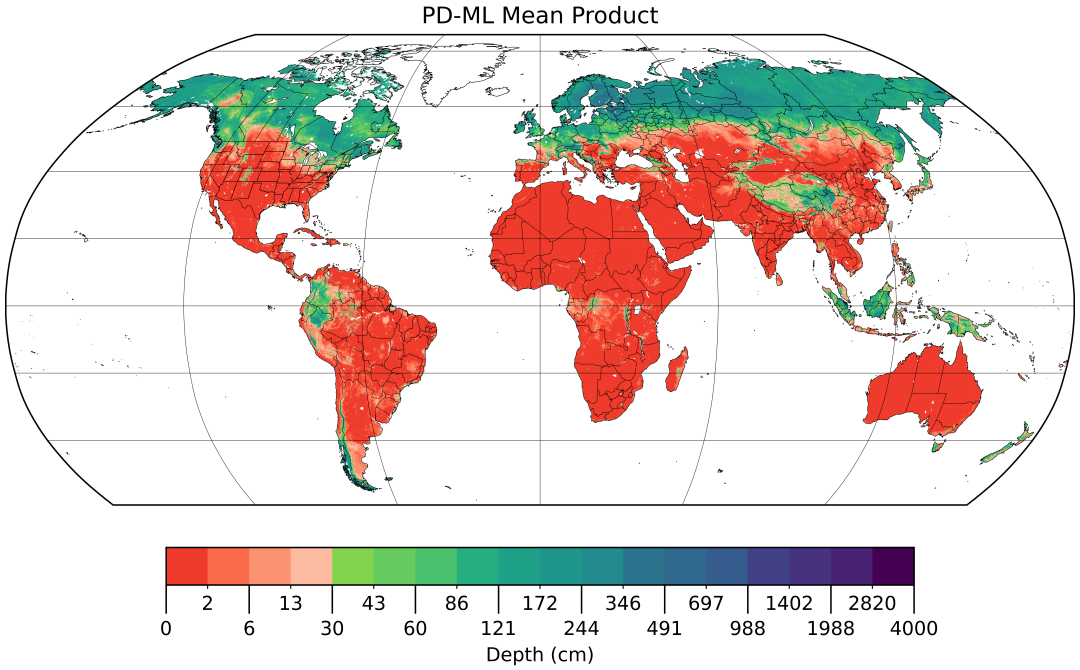

Figure 3.8: The PD-ML Mean Product (modelled mean) made by averaging the results of 401 bootstrap PD-ML model runs. Any area without a coloured grid cell does not contain any data. As the training and testing data represents the mean peat depth of the peatlands within a grid cell rather than the mean peat depth over the entire grid cell (see Chapter 3.1.2.1), PD-ML's results represent the same value. The colour bar is presented in a log scale with a colour break at 30 cm to represent a commonly used threshold when classifying an area as peatland (see Table 1.1) (Loisel et al., 2017).

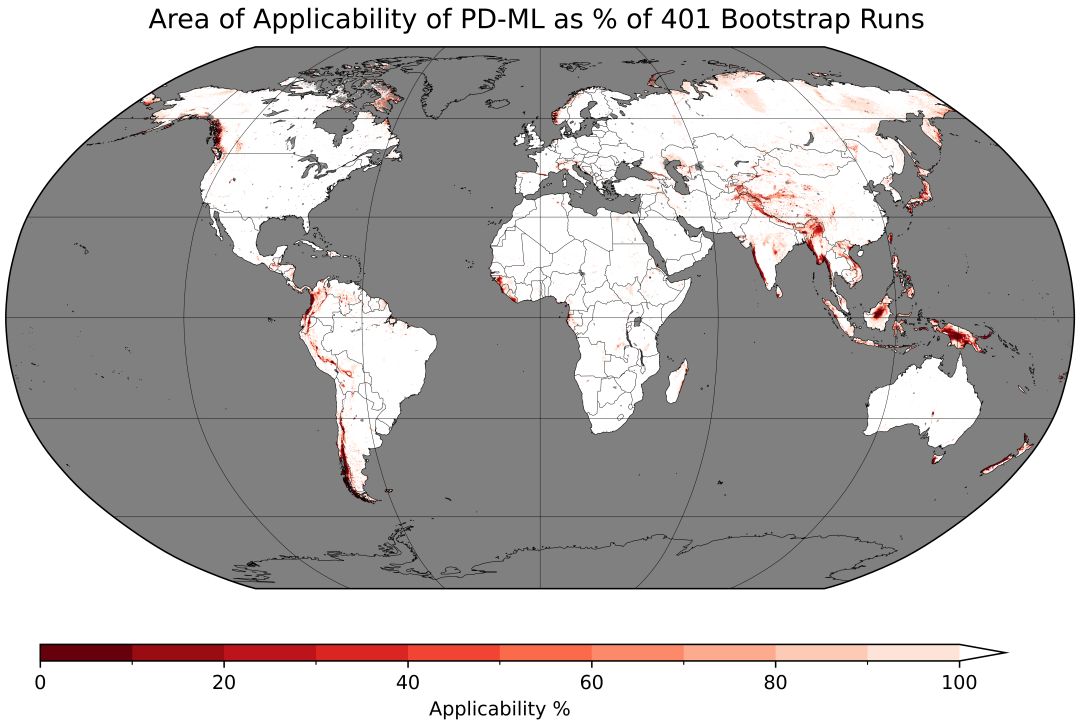

Figure 3.9: The percentage of bootstrap runs that are applicable for any given area. Any grid cell that is applicable across all 401 bootstrap runs is white.

should be considered trustworthy based on the AOA. Figure 3.9 shows the AOA across all bootstrap runs, with the gradient indicating the percentage of runs for which a grid cell was deemed applicable. For locations with the aforementioned missing data in some bootstraps, these are counted as having reduced applicability in accordance with the amount of runs the data was missing for. Of the grid cells that experience any amount of inapplicability, 4.2% are due to occasional data gaps in some bootstrap outputs and the majority of these are located on the northern coasts of the Caspian sea. We consider this AOA map to represent the areas of trustworthiness of the modelled mean and should be carefully noted by any future users of our depth map.

Looking globally, the modelled mean is shown to be least applicable over mountainous regions such as portions of the Southern Alps (New Zealand), Maoke Mountains (Indonesia), Müller Mountains (Papua New Guinea), Western Ghats (India), Himalayas, Caucasus Mountains (Eastern Europe), Scandinavian Mountains (Northern Europe), Guinea Highlands (West Africa), Coast Mountains (Canada), and southern Andes (South America) (Figure 3.9). While we do have some training data in or near mountainous regions, they contain a limited amount of training data, especially in

Indonesia (Figure 3.4). Our lack of observational data for these regions likely contributes to their higher inapplicability. There are regions in the Siberian Plateau, East Siberian Mountains, and in areas stretching south from the Chukchi Peninsula to the Kamchatka Peninsula with an applicability of 80-90%. Once again, these regions have limited training data available. Additionally, the AOA is based on the predictors used by the model for a given run. Therefore, the AOA is influenced by the uncertainty in the training data, as represented by the bootstrapping, and further supports the notion that relying on a single instance of model output is potentially risky.

Figure 3.10 now compares PD-ML to PEATGRIDS (Widyastuti et al., 2024) and Hugelius et al. (2020). However, these comparisons cannot be exact due to their differing resolution and spatial extent. The one km resolution depth predictions in PEATGRIDS are limited to 'peat dominated' areas established within the Global Peat Map version 2.0 (United Nations Environment Programme, 2021) with some alterations in Indonesia based on national peat maps. Hugelius et al. (2020) predicted peat depths north of 23°N at a 0.1° resolution. Additionally, Hugelius et al. (2020) do not include non-peat data in their depth modelling process, as they combine their peat depth results with a peat coverage map of their own making to mask out non-peatlands. By establishing non-peat areas in a pre or post modelling step rather than including those regions in the model itself, the model output and the predictor importance determined by PEATGRIDS and Hugelius et al. (2020) should not always be considered trustworthy beyond the bounds of their chosen peat coverage products. In contrast, PD-ML was intended to achieve a global representation of peat depth independent of other peatland datasets and any possible errors therein.

We can perform a visual analysis of the modelled mean and PEATGRIDS (Widyastuti et al., 2024) in Figure 3.10. In general, the modelled mean has some of its deepest peats in similar areas to which PEATGRIDS was able to predict peat depth. For example, these patterns are visible throughout Indonesia, the West Siberian Lowlands, Hudson Bay Lowlands, Canadian Boreal Plains region, Pastaza-Marañón Foreland Basin, and Congo Basin. However, PEATGRIDS has deep peat covering a larger portion of the Congo Basin than the modelled mean. Conversely, the modelled mean indicates deeper peat than PEATGRIDS in their areas of overlap within southern Chile and Argentina. While both the modelled mean and PEATGRIDS feature peat in Alaska, it is some of the shallowest peat within PEATGRIDS, but among the deeper peat for the modelled mean. The Tibetan Plateau is not within the bounds of

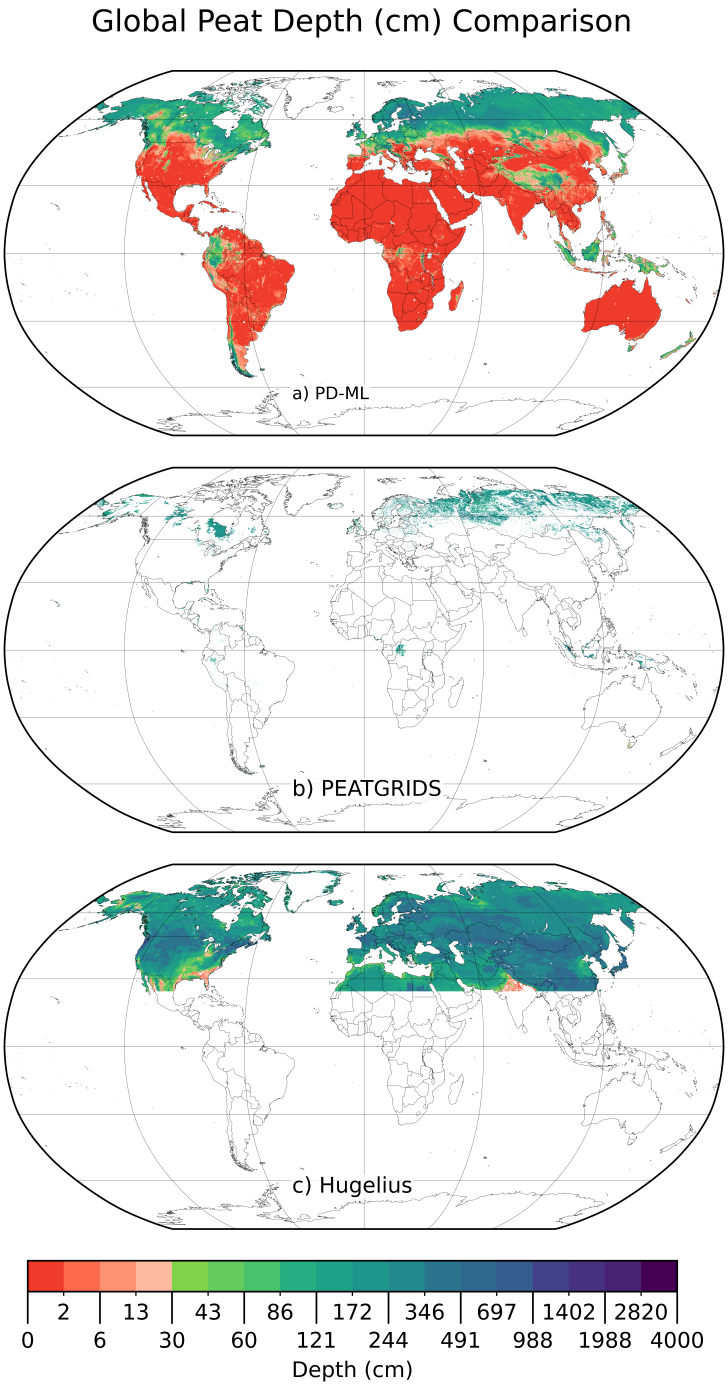

Figure 3.10: (a) Average peat depth predictions of the modelled mean (five arcminute resolution. (b) PEATGRIDS peat depth results (Widyastuti et al., 2024) (one km resolution). (c) Hugelius et al. (2020) peat depth results (0.1° resolution). The colour bar is presented in a log scale with a colour break at 30 cm to represent a commonly used threshold when classifying an area as peatland (see Table 1.1) (Loisel et al., 2017).

prediction for PEATGRIDS, but appears as a significant peat complex for PD-ML.

Hugelius et al. (2020) can also be compared to the modelled mean and PEATGRIDS in the northern latitudes. Hugelius et al. (2020) predict the most widespread and some of the consistently deepest peat compared to the other products, although large portions of it are filtered out by their post processing step (e.g. most of the continental US, large portions of central Eurasia, and the Sahara). Some notable areas of difference are parts of Alaska, the Hudson Bay Lowlands, and West Siberian Lowlands, where Hugelius et al. (2020) predicts shallower peat than the modelled mean and PEATGRIDS. Hugelius et al. (2020) are capable of predicting peat depths shallower than 30 cm, such as in India and the Coastal Plains of the US. However, the scarcity of non-peat regions demonstrates the inconsistent quality of this product when isolated from their peat coverage map and corroborates the need for assessments such as the AOA to accompany ML mapping products (Meyer and Pebesma, 2021).

PD-ML, PEATGRIDS (Widyastuti et al., 2024), and Hugelius et al. (2020) must be harmonised to allow a statistical analysis. The datasets were prepared by reducing the resolution of PEATGRIDS to match that of PD-ML and Hugelius et al. (2020) respectively, then removing all data outside the bounds of these low resolution versions of PEATGRIDS for both PD-ML and Hugelius et al. (2020). The mean peat depth could then be found for each product over roughly the same area, with the mean of each PD-ML bootstrap being calculated then averaged together. We are able to provide uncertainty bounds for PD-ML via the bootstrap outputs as well. We determined the 5th and 95th percentile of the bootstrap means then calculated a symmetrical uncertainty by finding the difference between the 5th percentile mean and 95th percentile mean and dividing the solution by two.

A numerical comparison between PD-ML, PEATGRIDS (Widyastuti et al., 2024), and Hugelius et al. (2020) is shown in Table 3.2 and Figure 3.11. Table 3.2 shows that PD-ML generally has a shallower mean than the other depth products. This is especially the case for the tropics where the mean depth of PEATGRIDS is nearly 100 cm greater than that of PD-ML. In the northern hemisphere, Hugelius et al. (2020) reached the deepest mean depth across all of the products, which is unsurprising given the noticeably deeper peat seen in Figure 3.10. The difference in resolution across the three depth maps may have some impact on the mean depths presented here. Figure 3.11 provides more detail on the distribution of depth values across PD-ML, Hugelius et al. (2020), and PEATGRIDS. Given the difference in how each depth product handled non-peat regions, depths over 30 cm are more directly comparable

(30 cm is a commonly used delineation when classifying an area as peatland; Loisel et al. (2017)). In general, Figure 3.11 shows the three depth products have the same relative distribution of depths, with a peak occurring at roughly 250 cm. PD-ML and Hugelius et al. (2020) are largely in agreement, except Hugelius et al. (2020) reaches greater depths. PEATGRIDS once again achieves deeper depths than PD-ML in the tropics. Because PEATGRIDS is at a higher resolution, it has more data points overall within the harmonised area.

Table 3.2: Estimated mean peat depths of PD-ML, PEATGRIDS (Widyastuti et al., 2024), and Hugelius et al. (2020). Note that PD-ML and Hugelius et al. (2020) are masked by reduced resolution versions of PEATGRIDS to allow for more direct comparisons.

Region	**Source**	**Mean Peat Depth (cm)**
Global	PD-ML	182.4 ± 7.3
	PEATGRIDS	213.5
Northern Hemisphere (> 23°N)	PD-ML	183.1 ± 7.3
	PEATGRIDS	213.6
	Hugelius et al. (2020)	238.5
Tropics (23.5°S-23.5°N)	PD-ML	133.7 ± 23.5
	PEATGRIDS	226.3

Higher training data availability may contribute to PEATGRIDS (Widyastuti et al., 2024) reaching a significantly deeper mean depth in the tropics than PD-ML. PEATGRIDS has more non-zero cm peat depth training data available for tropical regions than PD-ML, particularly for Indonesia. This difference in training data supply is partially due to PEATGRIDS' inclusion of depths that are randomly sampled from existing peat depth maps, such as the Indonesian national map of peat depth categories (Anda et al., 2021). The midpoint of the associated peat depth category was used as the depth value for the sampled points. This approach has the added uncertainty that spatially extrapolated peat depths, i.e. the national peat depth map, were treated as discrete soil cores. The bootstrapping of PD-ML has demonstrated that the depth value assigned to training data points in this process could potentially have a strong impact on a model's behaviour.

While the various approaches are not directly comparable (e.g. PEATGRIDS

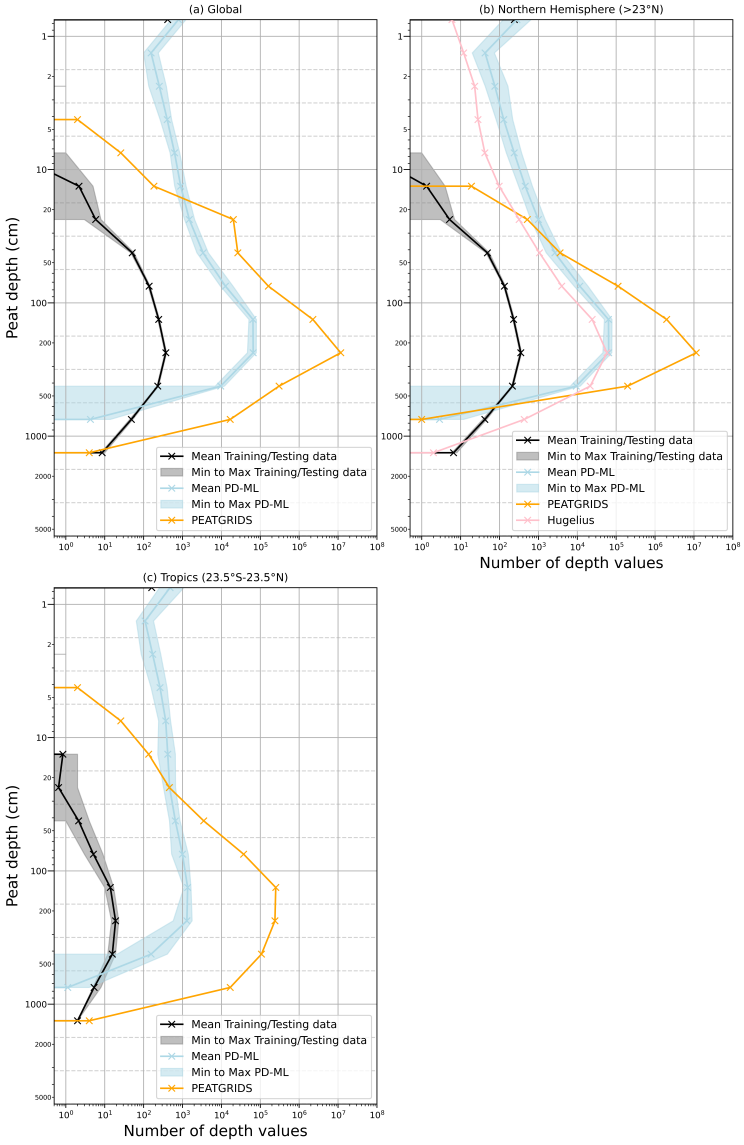

Figure 3.11: Histograms showing peat depth distributions of the PD-ML training and testing data, PD-ML model run outputs, Hugelius et al. (2020), and PEATGRIDS (Widyastuti et al., 2024) across different regions: (a) global, (b) northern latitudes (> 23°N), and (c) tropics (23.5°S-23.5°N). Where there are multiple datasets present for PD-ML due to bootstrapping, the mean distribution is represented as a solid dark line and the minimum to maximum range of the distribution in a lighter band. PD-ML, Hugelius et al. (2020), and PEATGRIDS were harmonized as described in the main text. Note that the axes of all panels are presented in log scales, which can inflate the prominence of smaller values.

trained regional models and Hugelius et al. (2020) only exists for high latitudes), the relative shallowness of PD-ML overall in Table 3.2 could be the result of our zero-inflated training data and a common tendency of ML algorithms to predict values toward the mean of their training data (Zhang and Lu, 2012; Song, 2015). Hugelius et al. (2020) found this to be the case for their RF model and Xu et al. (2016) experienced the same behaviour with both an RF model and a Maximum Entropy model when attempting to predict mean vegetation canopy height. The empirical cumulative distribution of Figure 3.12 indicates that PD-ML is also following this trend when compared to the training data. Figure 3.12 shows that just over 90% of all the training and testing data have a value of zero cm depth, while about 75% of the PD-ML bootstrap predicted depths are at or near zero cm over the same grid cells. The distribution of the observations and PD-ML bootstrap predictions over the same grid cells intersect at about 150 cm with the PD-ML values reaching nearly 100% of their distribution by a depth of about 400 cm, whereas the training and testing data has deeper values and does not approach 100% as shallowly. Thus PD-ML does not predict end members as well. When considering the mean of non-zero and zero cm depth data, PD-ML closely matches its training data. However, when only considering non-zero cm depth data, PD-ML is shallower overall (Figure 3.12).

3.2.3 Model Performance Estimation

In addition to the previous qualitative assessments, we conducted a more quantitative estimation of PD-ML performance using a variety of metrics. We started with a grid cell comparison between the depth values predicted by the model and the observations at these locations. For these comparisons we calculate the RMSE, MBE, and a version of the normalised mean error (NME) developed by Kelley et al. (2013). RMSE and MBE are popular metrics for assessing ML models (Plevris et al., 2022), with PEATGRIDS and Hugelius et al. (2020) both using RMSE, and Peat-ML calculating RMSE as well as MBE. The Coefficient of Determination (R^2) is used by Peat-ML and PEATGRIDS, however we have decided against its inclusion due to its ambiguous representation of non-linear model performance (Plevris et al., 2022). We chose to calculate Kelley et al. (2013)'s NME as we follow their process of comparison to two null models: the observed mean null model and the observed random resampling null model (explained in Table 3.3 and analysed in Figure 3.14d and 3.15). Table 3.3 shows the equations of our chosen metrics and explains how to interpret them.

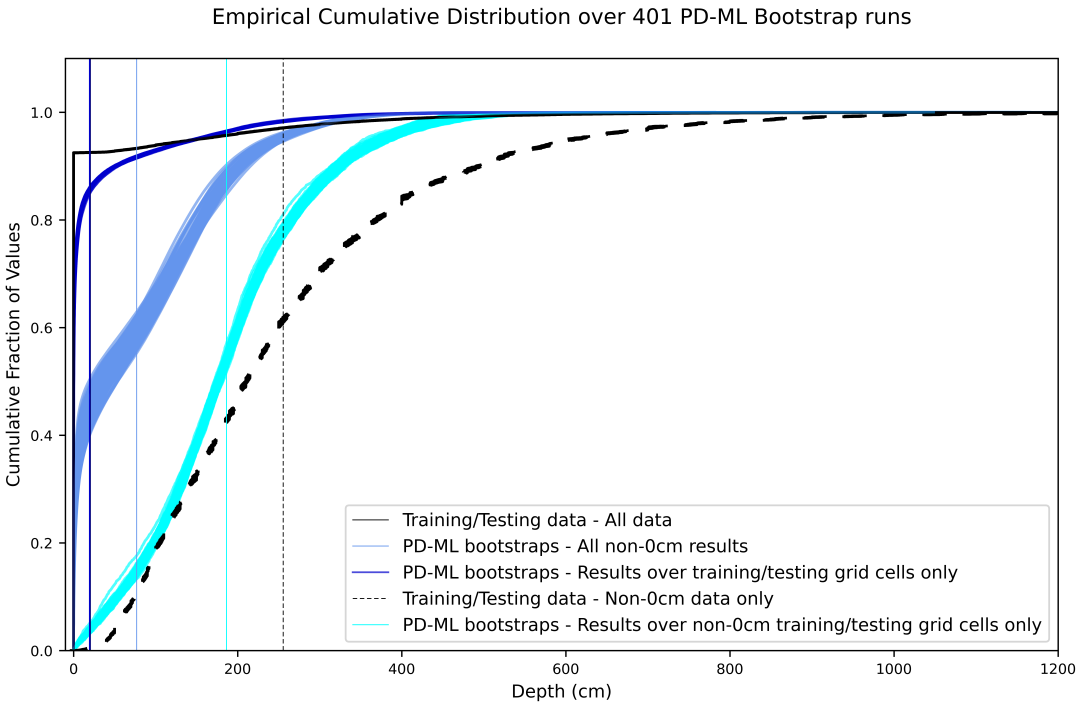

Figure 3.12: Empirical cumulative distribution (ECD) of PD-ML results and its training and testing data. The black line indicates the ECD of all of the training and testing peat depth data. The dashed black line indicates the ECD of the non-zero cm depth grid cells in the training and testing peat depth data. The dark blue lines indicate the ECD of the outputs of the PD-ML bootstrap runs but only over the grid cells which have training and testing data. The light blue lines indicate the ECD of all non-zero cm depth data in the outputs of the PD-ML bootstrap runs. The cyan lines indicate the ECD of the outputs of the PD-ML bootstrap runs but only over the grid cells which have non-zero cm depth training and testing data. The vertical lines indicate the mean of the corresponding datasets.

Table 3.3: Equations used in assessing PD-ML model performance and their meanings in this context. Here, p refers to the values predicted by PD-ML, r is the 'real' observed values (where all observations within a grid cell are meaned as described in Chapter 3.1.2.1), $\bar{r}$ is the mean of the observed values over all grid cells, and N is the number grid cells for which there are both predicted and observed values.

Equations	**Values**	**Interpretations in this study**
$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(p_i - r_i\right)^2}$	Values are in centimetres and can range from 0 to $+\infty$, with 0 meaning the predicted and observed values are the same.	Represents the difference between the predicted values and the observed values. RMSE can be more sensitive to outliers due to the squaring of the error (Plevris et al., 2022).
$MBE = \frac{1}{N}\sum_{i=1}^{N}\left(p_i - r_i\right)$	Values are in centimetres and can range from $-\infty$ to $+\infty$. An MBE of 0 occurs when positive and negative errors cancel out or when the predicted and observed values are the same.	Represents the average error of the predicted values. MBE can be negative if the predicted values are generally smaller than observed or positive if the predicted values are bigger. However the presence of both positive and negative errors can result in an MBE that is near 0 even when errors are large (Plevris et al., 2022).

Continued on next page

Equations	**Values**	**Interpretations in this study**
$NME = \frac{\sum_{i=1}^{N} \lvert p_i - r_i \rvert}{\sum_{i=1}^{N} \lvert r_i - \bar{r} \rvert}$ Version of the formula proposed by Kelley et al. (2013) where the mean absolute error is normalised by the variance in the observations.	Values are unitless and can range from 0 to $+\infty$, with 0 meaning the predicted and observed values are the same. A value of 1 indicates agreement with the observed mean null model. A value greater than 1 indicates the model performs worse than this null model. The observed mean null model is a dataset equal in size to that produced by the original model, where every entry is set to the mean of the observations. Kelley et al. (2013) also produce random null models for analysis. The random null models are a series of datasets equal in size to that produced by the original model, filled by randomly sampling the observations with replacement (bootstrapping).	Represents the mean of the absolute model error normalised by the variance in the observations. This NME allows us to assess whether the model approaches the true observed values, is simply following the mean of the observed values, or performing worse than the mean of the observed values (Plevris et al., 2022; Kelley et al., 2013).

We calculated these metrics over selected areas for each bootstrap model run, using

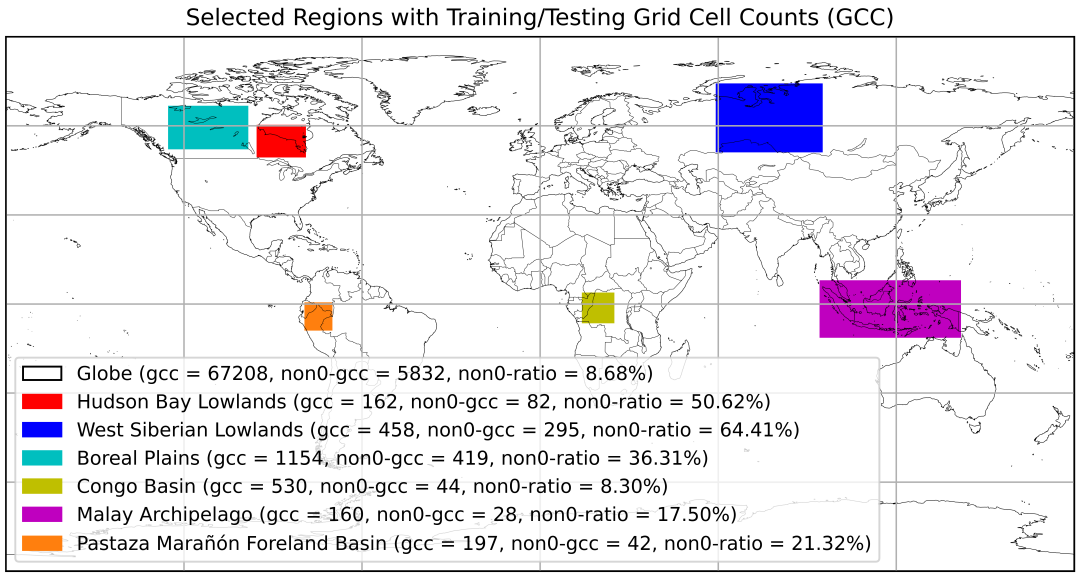

Figure 3.13: Geographical distribution of selected regions for the PD-ML performance assessment. While the entire predicted area is global, the model predictions occur only over grid cells where the Light-GBM algorithm has coverage from all selected predictors. The number of observed grid cells contained within each area is recorded in the legend, with gcc referring to the total grid cell count and non0-gcc referring to the count of non-zero cm depth grid cells. The ratio of non-zero cm depth grid cells to total observed grid cells is also recorded (non0-ratio).

the cross-validation results. Figure 3.13 shows the regions we selected for focussed assessment. We sought to have an equal amount of high latitude and tropical areas for analysis, and chose regions which have high peatland fractional coverage according to Peat-ML (Figure 1.1) while also including a fair portion of observed non-zero cm peat depth grid cells (Figure 3.4). The output of each bootstrap model run is compared to the bootstrapped version of the observed data used in training and testing that model instance. By calculating these metrics for each bootstrap run, we further demonstrate the variation in model behaviour due to the sampling uncertainty in the observed peat depth data. However, this same uncertainty affects the accuracy we are capable of reaching within our performance assessment. For a more truthful representation of model ability, the model results produced through blocked-leave-one-out cross-validation (BLOOCV) are used in the metric calculations (see Chapter 1.1.2 and Figure 3.1). The BLOOCV results are PD-ML's predictions without the advantage of having learned from the observed peat data in the current area.

Figure 3.14 shows the mean depth and performance metrics for the PD-ML boot-

strap runs. Throughout Figure 3.14a, 3.14b, and 3.14c the variance of the bootstraps in the selected regions is broadly consistent with fluctuations of roughly 10 to 25 cm. Figure 3.14b indicates that the Congo Basin has the best performance of the selected peatland regions. We hypothesise that this pattern may be due to the ratio of observed non-zero cm depth data in the Congo Basin mirroring that of the globe (non0-ratio in Figure 3.13), therefore the broad patterns the model has learned globally may be more applicable in this region. Meanwhile, the Malay Archipelago tends to have the worst performance in RMSE (Figure 3.14b). The Malay Archipelago also has the most severe underprediction of peat depth overall (Figure 3.14c). We believe the poor performance in this region is the result of insufficient training data in the Malay Archipelago (Figure 3.13 shows the Malay Archipelago has the lowest gcc and non0-gcc of all selected regions, despite its size). Additionally, we are also likely unable to adequately assess the performance in parts of this region given that PD-ML is generally not applicable there according to the AOA (Figure 3.9).

Figure 3.14 further illustrates PD-ML's tendency to predict towards the mean of the training data, a mean that is shallow due to the strong presence of zero cm peat depth data. For example, the variance in mean peat depth across the bootstraps is stable globally and matches that of the training and testing data (Figure 3.14a). Similarly, the variance of the other metrics are all generally stable globally, suggesting PD-ML is able to repeatedly capture broad trends to consistent levels of success despite the variations in training data across the bootstraps. The MBE indicates that within peatland regions, the magnitude of PD-ML's overly shallow predictions is greater than the magnitude of the model's overly deep predictions (Figure 3.14c). We suggest that this is indicative of PD-ML's bias towards the zero-inflated shallow mean.

The NME results are noticeably different compared to the other metrics (Figure 3.14d). In general, the NME is between 0.6 and 1.0 for most peatland regions and the globe. However, the amount of variation across regions is less consistent, with the Congo Basin showing a particularly wide range of NME scores across the bootstrap runs. Since the denominator of the NME is the observational variance (Table 3.3), our bootstrapping method has the potential to create more wide ranging values here. Additionally, as the ratio of observed non-zero cm depth data (non0-ratio in Figure 3.13) approaches zero so does the NME denominator, such that in the most extreme case where all observations are zero cm, the NME will be infinity. The Congo Basin contains several grid cells with some of the most peat depth measurements within

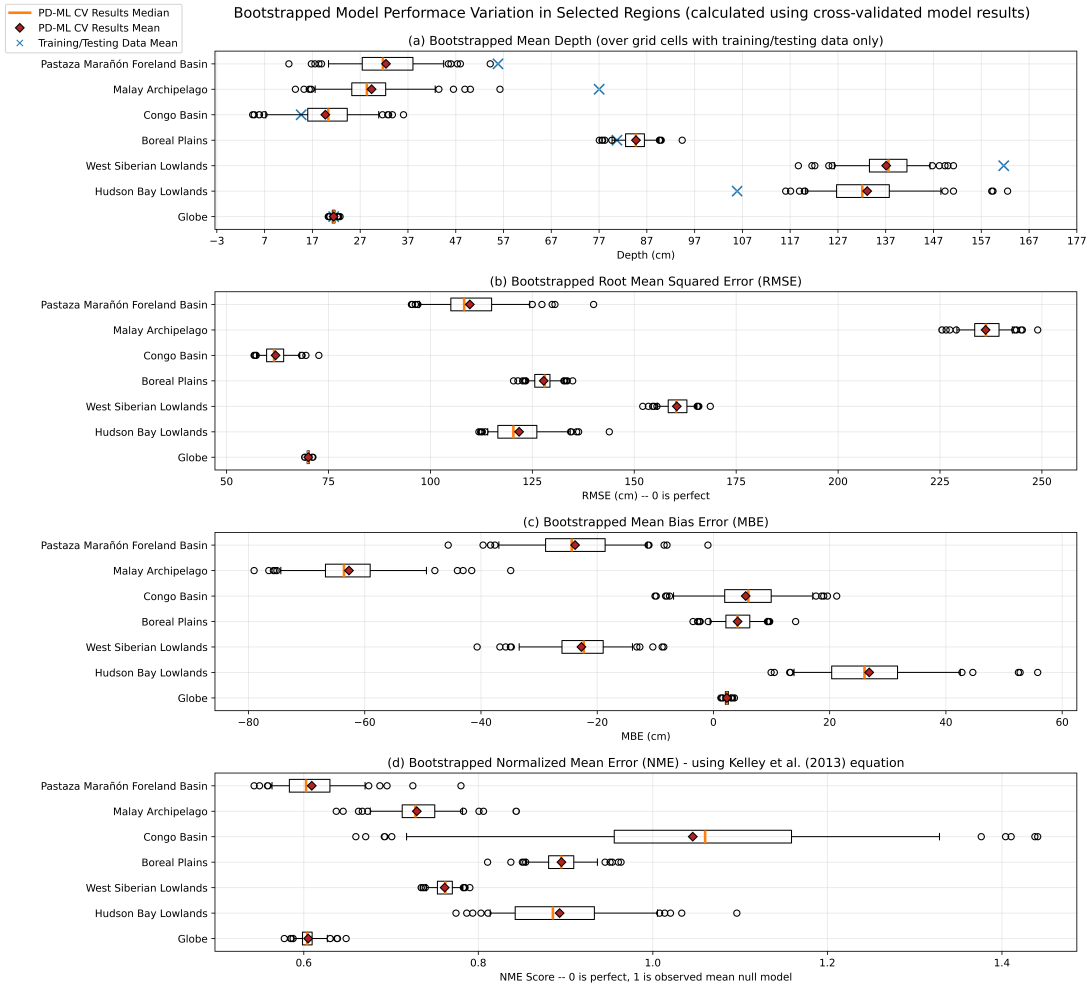

Figure 3.14: Box and whisker plots of PD-ML performance metrics over all bootstrap runs for the selected regions (see Figure 3.13), where the whiskers extend to the 5th and 95th percentiles and outliers are indicated with empty circles. (a) The variation in mean depth over all bootstrap runs with blue crosses indicating the mean of the bootstrapped observed datasets. (b) The variation in RMSE over all bootstrap runs. (c) The variation in MBE over all bootstrap runs. (d) The variation in Kelley et al. (2013)'s NME over all bootstrap runs (see Table 3.3 for NME explanation).

them compared to the majority of grid cells, thus the mean observed depth formed by the bootstrapping within these grid cells has the opportunity to vary more strongly. Also, the Congo Basin has the lowest ratio of observed non-zero cm depth data of all the selected peatland regions (non0-ratio in Figure 3.13). Taken together, the Congo Basin can experience a high observational variance, which can serve to exacerbate any poor model performance statistics in this location.

Following Kelley et al. (2013)'s implementation of the NME, we can compare PD-ML's NME to that of observed random resampling null models (see Table 3.3 for an explanation of the null models). These random null models are produced by applying a bootstrapping process to our 67 208 observed grid cells, such that we create various datasets made up of a random selection of these while still having the same total number of grid cells. We made 1000 random null models from the version of the observed data that did not undergo any bootstrapping within its grid cells (Figure 3.4). The NME is calculated for each of the random null models using the formula in Table 3.3 and plotted in Figure 3.15a. The random null model results are compared to the global PD-ML bootstrap results in Figure 3.15b to assess whether PD-ML can perform better than random resampling. As a whole, PD-ML performs better than the observed mean null model and the random null model. However, Figure 3.14d shows that for some peatland regions, the uncertainty in the training data may result in extreme cases that perform worse than both null models (e.g. outlier bootstraps in the Congo).

The RMSE of PD-ML can be compared to that of Hugelius et al. (2020) and PEATGRIDS (Widyastuti et al., 2024), with the acknowledgement of key differences beyond our inclusion of non-peat training data. Hugelius et al. (2020) appear to use 10-fold cross-validation to validate their model results, with their chosen R package using a form of random sampling that attempts to balance the range of values seen in each fold (Kuhn, 2008). PEATGRIDS conducts random sampling to split apart 30% of their data for testing. As discussed in Chapter 1.1.2, random k-Fold cross-validation can result in highly optimistic representations of model performance, with randomly sampled training-testing splits facing the same challenge (Meyer and Pebesma, 2022). Thus, our accounting of spatial autocorrelation in our cross-validation block selection (see Chapter 1.1.2 and 3.1.3) allows for a more accurate or possibly pessimistic representation of model performance (Milà et al., 2022), thus our performance metrics could be comparatively lower as a result.

Hugelius et al. (2020) and PEATGRIDS (Widyastuti et al., 2024) are regional

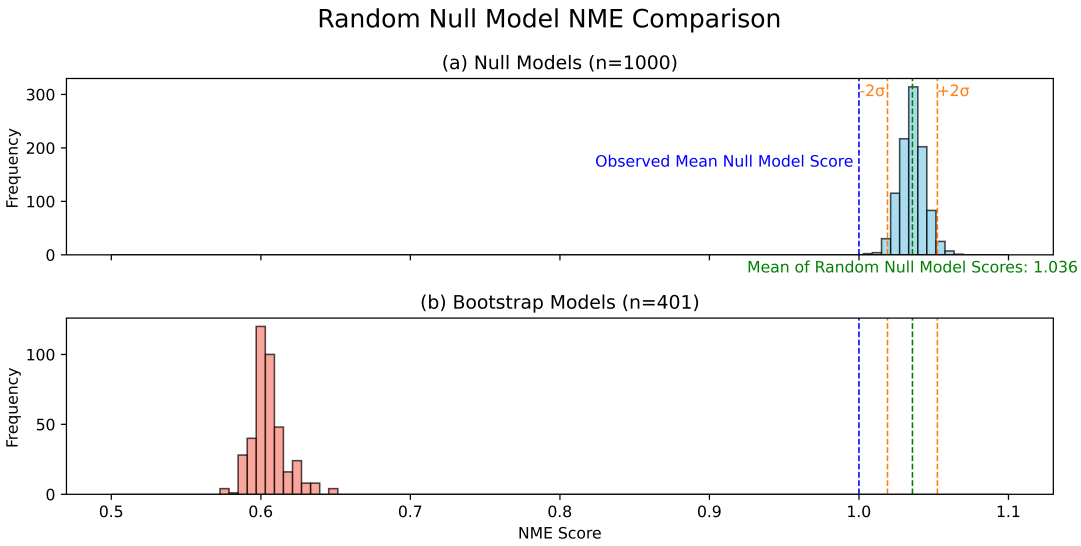

Figure 3.15: Histograms of model NME score distributions (see Table 3.3 for an explanation of the NME and null models). (a) The distribution of the NME scores of the observed random resampling null models, with the number of models indicated in brackets. (b) The distribution of the NME scores of the PD-ML bootstrap model runs, with the number of models indicated in brackets. The blue line indicates the NME of the observed mean null model, the green line indicates the mean NME score of the random null models, and the orange lines indicate the position of two standard deviations away from this mean.

models, or a combination thereof. Hugelius et al. (2020) model northern latitude peatlands only, therefore an RMSE comparison to PD-ML is only relevant over that area. While PEATRGRIDS present a global map, it is produced by combining the output of six regional models (North America, Europe and Russia, Latin America, Africa, South-Southeast Asia, Australia and New Zealand) which each have their own RMSE scores. The exact bounds of these six models are not provided (see Figure 1 in Widyastuti et al. (2024)), therefore a more precise RMSE comparison to PD-ML in these areas is not possible. The difference in resolution across Hugelius et al. (2020), PEATGRIDS, and PD-ML also complicates the ability to do an exact comparison.

Prior to performing bias correction on their model output, Hugelius et al. (2020) reported an RMSE of 142.2 cm. When we limit the performance assessment of PD-ML to $\geq 23°$N to match Hugelius et al. (2020), PD-ML has a mean RMSE of 88.3 ± 1.1 cm. PEATGRIDS report the following RMSE scores for their models : 168 cm for North America, 96 cm for Europe and Russia, 58 cm for Latin America, 103 cm for Africa, 191 cm for South and Southeast Asia, and 85 cm for Australia and New Zealand (Table 2 in Widyastuti et al. (2024)). The average RMSE across the six models is 116.8 cm. Comparatively, PD-ML has a global mean RMSE of 70.1 ± 0.9 cm over all the bootstrap runs. Our selection of the Malay Archipelago region for assessment of PD-ML is fairly similar to the area selected for PEATGRIDS' South and Southeast Asia model and overlaps with much of their training data there (Figure 1 and 2 in Widyastuti et al. (2024)), although our AOA suggests this region is challenging for PD-ML. PD-ML has a mean RMSE of 236.2 ± 6.8 cm over the Malay Archipelago. Again, we hypothesise that PD-ML's strong decrease in performance in this region is partially due to a paucity of training data. PEATGRIDS has more data available for this area through their use of peat depth maps, yet their highest RMSE score still occurs there, which suggests this region may present unique challenges for modelling as a whole.

PD-ML was developed at the global scale for several reasons. PEATGRIDS (Widyastuti et al., 2024) predicted over several model domains with the final global map produced by stitching together these subdomains. If PD-ML were to adopt this same methodology, the stitching together of multiple model outputs would require the harmonisation of these outputs at their boundaries based on some assumption or heuristic. The effects of this harmonisation on the AOA would also need to be considered. Depending on how regions are divided, there could be a significant disparity in the amount of peat observations available for training in each region. This

disparity could impact the quality of each respective model (Somarathna et al., 2017; Meyer and Pebesma, 2021), while also potentially requiring different distances of spatial autocorrelation to be determined for cross-validation processes for each model as well as specific tuning of the amount of non-peat data in each region. A few preliminary multi-model tests were run for PD-ML throughout development. We noted some extreme behaviour occurred, such as the selection of only one predictor, depending on the regions chosen. We suspect these results were partially due to an even greater zero-inflation of some of the regional training datasets. Furthermore, our bootstrapping method has revealed the significant influence of uncertainty in the observations on the model. Our results in Figure 3.14 suggest the model sensitivity to this uncertainty could potentially increase at smaller regional scales as well as be region dependent.

3.2.4 Preliminary Estimation of Carbon Stocks Using Model Results

We can use our peat depth results from PD-ML in some basic C stock estimates. To calculate these estimates, we generally follow a similar approach to PEATGRIDS (Widyastuti et al., 2024) with the equations set out below:

$$C_{dens} = OC \times BD \tag{3.1}$$

$$C_{stock} = C_{dens} \times D_{peat} \times A_{cell} \tag{3.2}$$

where organic C content (OC) is in kg C kg^{-1}, bulk density (BD) is in kg m^{-3}, C_{dens} is C density in kg C m^{-3}, D_{peat} is peat depth in m, A_{cell} is the area of a grid cell in m^2, and C_{stock} is C stocks in kg which we then convert to Pg C. We use the peatland fractional coverage of Peat-ML to find A_{cell} by calculating the trapezoidal area of each grid cell in m, then scaling the solution by the fraction of peat coverage within it. The application of Peat-ML and the resulting A_{cell} within the C stock estimate calculations is necessary, as the peat depths predicted by PD-ML represent the mean peat depth of only the peatlands within a grid cell (see Chapter 3.1.2.1).

We determine different C stock estimates using C_{dens}, OC, and BD values established in other studies such that we can compare to their results. Specifically, we test using mean C_{dens} values from PEATGRIDS' modelled OC and BD results (see Table 3 in Widyastuti et al. (2024)). Additionally, we test using OC and BD values from

two different approaches in Page et al. (2011). In both methods, Page et al. (2011) use the results from their literature review for the tropical OC and BD estimates, while employing a high latitude C_{dens} value based on the estimates of Immirzi et al. (1992), which used an average peat depth of 1.5 m, for one method, and recalculating a high latitude C_{dens} value by combining the data of Immirzi et al. (1992) with an average peat depth of 2.3 m provided by Gorham (1991) for the second method. Both PEATGRIDS and Page et al. (2011) have different respective values for high latitude regions and the tropics, which they delineate at 23.5°N and 23.5°S. Thus we also calculate C_{stock} separately for these regions before merging the results together for a global value. However, for our calculations, we chose to use 30°N and 30°S as our boundaries based on expert opinion (Personal Communication - A. Gallego-Sala, 2024) and to include known peatlands that are more similar to tropical ones (e.g. Florida). We get a range of results for all PD-ML bootstrap runs. We also perform the same calculations using the mean depth of 254.8 cm from the non-zero cm measurements in Peat-DBase (Chapter 2) for another element of comparison.

PD-ML estimates a range of 327-378 Pg C in peatlands globally, with 6.9-14.6% located in tropical peats (Figure 3.16). This percentage is comparable to PEATGRIDS and Page et al. (2011) which have 14.5% and 14.5-18.5% in the tropics, respectively. Figure 3.16 shows that PD-ML estimates smaller peatland C stocks compared to the other products. PEATGRIDS reports particularly high C stock estimates, with a global value of 1029 Pg C. They attribute these elevated values to the significant peat coverage within the Global Peat Map (6.57 million km^2 whereas Peat-ML reports 4.04 million km^2). Nonetheless, compared to the more moderate estimates from Page et al. (2011) and Hugelius et al. (2020), PD-ML is still lower. PD-ML's potential shallow bias is a possible contributor to its lower C stock estimates, particularly in the tropics. The C stocks calculated from the mean depth of the non-zero cm measurements in Peat-DBase are among the higher values in most cases (this mean depth of over 2.5 m exceeds the depths of 1.5 m and 2.3 m used by Page et al. (2011) and those of PEATGRIDS and Hugelius et al. (2020) (Table 3.2)). Despite PD-ML producing lower peatland C stocks of roughly 300-400 Pg C, its estimates still fall within the range (113 Pg to 612 Pg C) previously reported in the literature (Minasny et al., 2019), excluding the recent high estimate of PEATGRIDS.

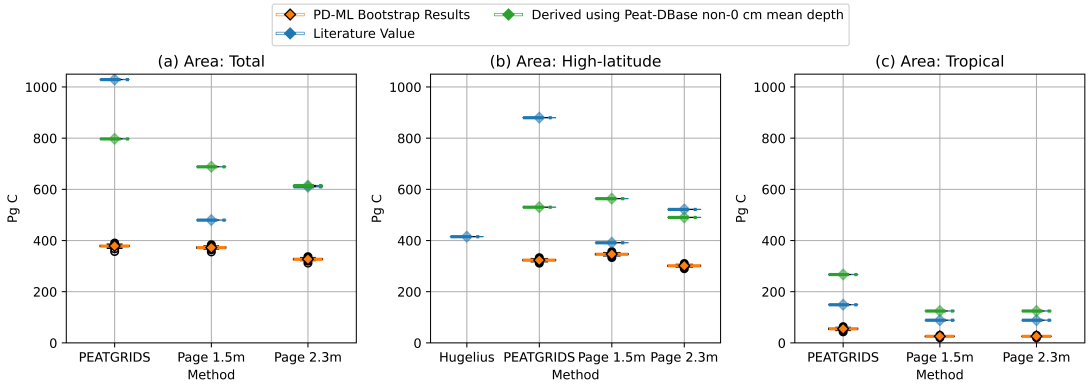

Figure 3.16: C stock estimates calculated based on PEATGRIDS (Widyastuti et al., 2024) and different approaches in Page et al. (2011). C stock estimates from Hugelius et al. (2020) are shown for comparison where applicable. (a) Global C stock estimates. (b) C stock estimates for high latitude regions. (c) C stock estimates for low latitude regions. C stocks for PD-ML are calculated with a low latitude boundary of 30°N to 30°S, all other products used 23.5°N to 23.5°S.

3.2.5 Model Limitations and Future Work

We have modelled peat depth globally and provided a detailed analysis of the uncertainty present in our method, however further limitations exist. As was recognized by Melton et al. (2022), we cannot easily draw conclusions about possible peat forming conditions from the predictors selected by an ML model, as it is challenging to distinguish between cause and effect in these choices. Our examination of PD-ML's sensitivity to uncertainty in the training data revealed that predictor selection can be highly variable, making assumptions about their relationship to peat development potentially even more ambiguous.

PD-ML is clearly limited by the availability of peat depth observations on which to train. Both Widyastuti et al. (2024) and Melton et al. (2022) explain that the accuracy of PEATGRIDS and Peat-ML respectively are impacted by training data availability, with their training datasets each being biassed in favour of the high latitudes. This bias is also present for PD-ML and its influence may be seen in our seemingly poor performance in the Malay Archipelago region and more significant underprediction of peat depth in low latitude regions in general (Figure 3.14). We have demonstrated that sampling uncertainty in the observed data used for model training is also highly impactful on model performance. However, the PD-ML framework is publicly available and as more peat depth measurements become available it can be redeployed to create updated results.

We focused our efforts on predicting peat depth with the LightGBM algorithm based on the Peat-ML Framework (Figure 3.1), however other algorithms may offer better performance. Hugelius et al. (2020) and PEATGRIDS (Widyastuti et al., 2024) show that RF models can also be implemented to predict peat depth. Both RF models and LightGBM are ensemble models built from several decision trees. Decision trees are generally effective at dealing with missing data on their target variable, but not necessarily as suited to modelling data with a small sample size (Haixiang et al., 2017). Therefore, other algorithms, such as those based on neural networks, may achieve higher accuracy when working from a highly imbalanced dataset like ours (Haixiang et al., 2017; Chen et al., 2024). Alternatively, starting with a classification step to establish peatland versus non-peatland regions and then proceeding with the prediction of depth could also yield improved results (Rožanec et al., 2023).

Options exist for dealing with PD-ML's tendency towards the training data mean in the future. Alternative custom scoring methods could be tested, such as trying a

scheme where weights are assigned based on the imbalance ratio between zero cm and non-zero cm depths in the training data (Haixiang et al., 2017). Another possible approach to improving PD-ML's prediction distribution could be some form of bias correction. Hugelius et al. (2020) bias correct their peat depth results using a residual rotation method to account for their model's trend towards the training data mean, while Goodling et al. (2024) bias correct their results using an empirical distribution matching method.

There are other aspects of our approach which could benefit from further investigation. Conducting more tests on the influence of the random sampling of desert data may reveal additional model sensitivities. Including more paleoclimate predictors could benefit the model by providing further representation of longer time scales. For example, the modern climate predictors we have used are unlikely to be representative of the climate over the extensive periods many peatlands have developed. Goodling et al. (2024) propose alternative model assessment methods, such as a distribution scale assessment approach using empirical cumulative distribution results. Implementation of this distribution scale assessment may yield more details on PD-ML's performance. More thorough tests of a regional modelling approach like that of PEATGRIDS could be conducted to examine how the model sensitivity and AOA change. However, steps would likely need to be taken to account for the more extreme zero-inflation in the training data for some regions first. More detailed C stock estimations could also be performed. For example, Hugelius et al. (2020) employed a linear relationship developed from peat core data which estimated C stocks from peat depth.

3.3 Conclusion

We developed PD-ML, a global peat depth model, and conducted a detailed quality and uncertainty assessment to address model sensitivity to potential sampling bias in observed peat depth data. PD-ML is built on the Peat-ML Framework, using many of the same predictor datasets for climate, soil, vegetation, and terrain. New predictors were added to account for hydrology and paleo-environmental conditions as additional indicators of peat depth. To train PD-ML, a large database of peat depth measurements (Peat-DBase) was converted to a gridded format by taking the mean of the peat measurements within each grid cell. The resulting dataset was supplemented with additional non-peat data in desert regions. Multiple training datasets

were created by bootstrapping data within grid cells to test PD-ML's sensitivity to uncertainty in the peat depth database. To reduce the impact of strong zero-inflation within the training data, a custom scoring method prioritising performance on non-zero cm depth grid cells was implemented within the model.

PD-ML was then run on the bootstrapped training datasets. Predictor selection by the model was found to be highly sensitive to the individual bootstraps due to changes in the training data provided by each bootstrap. Performance metrics were calculated using the cross-validated results of the bootstrap runs and it was found that model accuracy also varied, particularly at the regional scale. Overall, PD-ML achieved a root mean square error of 70.1 $\pm$ 0.9 cm, a mean bias error of 2.1 $\pm$ 0.7 cm, and a normalised mean error of 0.6 $\pm$ 0.0 (non-standard equation). Where possible, PD-ML was compared to other peat depth maps and achieved similar or better results. PD-ML showed a tendency to predict towards the mean of its training data, which was relatively shallow due to the large amount of zero cm depths present. This shallowing within the model resulted in a lower overall mean depth and C stock estimate compared to other similar studies. PD-ML would likely improve with additional peat depth measurements for training, particularly in the tropics, and bias correction methods may assist in resolving the model's trend towards the training data mean.

3.4 Data Availability

The PD-ML model Python script and other associated code is stored here: `https://doi.org/10.5281/zenodo.15530817`. NetCDF files containing the mean of the PD-ML bootstrap runs, and the mean of the equivalent cross-validated model results are also stored in that location, along with Figure 3.9 and predictor importance results from all 401 model runs.

Chapter 4

Conclusion

The objective of this thesis was to answer the following questions:

1. What is the global distribution of peatland depths?
2. How much C is stored within peatlands?

To that end, I first combined a large amount of observed peat depth data to create Peat-DBase (Chapter 2). Peat-DBase was then used as the training and testing data of PD-ML, a newly developed ML based digital soil mapping approach, to produce a spatially continuous global map of peat depth. Bootstrapping and AOA calculations were applied within the PD-ML workflow to capture the sensitivity of this ML modelling method to uncertainties within the training data and to produce quantified estimates of the resulting prediction uncertainty. The peat depth results of PD-ML were then used to calculate some initial C stock estimates based on organic C content and bulk density values provided by previous studies (Chapter 3). In this chapter, I provide a qualitative exploration of my global peatland depth and C stock solutions in the context of these previous studies.

4.1 Global Peatland Depth Distribution Conclusion

Peat-DBase provides foundational information for global peat depth mapping, which is an advancement over previous products. It is the largest global database available, has at least some data for most known peatland complexes, and contains data on which areas are not peat. There are two primary sources of uncertainty that must

be kept in mind when using Peat-DBase. First is the poor data coverage for some regions, such as the Tropics and Eastern Russia, which is not readily solvable. Second is the potential bias that can arise from peat measurement approaches, such as taking a single core to represent an entire peatland.

Peat-DBase only contains information at points on the land surface where a measurement was made. Peat-DBase can thus be used to inform spatial interpolation approaches. PD-ML then builds upon Peat-DBase by linking geophysical variables to peat depth to allow peat depth predictions to be spatially continuous rather than only at points. PD-ML thus provides a framework to produce a global peat depth distribution that is not possible with Peat-DBase alone. Additionally, the PD-ML process took care to address the uncertainties of Peat-DBase through the use of the AOA and bootstrapping.

PD-ML provides an improved global peat depth distribution compared to PEATGRIDS (Widyastuti et al., 2024) and Hugelius et al. (2020) in multiple ways. First, by training on peat and non-peat data, the map produced by PD-ML is spatially continuous and does not need to be combined with a peatland extent map to be considered reasonable. In contrast, PEATGRIDS and Hugelius et al. (2020) rely on a peatland extent map to mask out non-peat regions before and after their depth mapping processes, respectively. Secondly, PEATGRIDS and Hugelius et al. (2020) use forms of random sampling within their performance assessment processes, which can produce overly optimistic views of model skill. The block selection steps employed within PD-ML's cross validation approach accounts for spatial autocorrelation and therefore provides a more conservative (even potentially pessimistic) illustration of model performance. Even with PD-ML's more conservative approach to cross-validation, PD-ML still achieved comparable performance to these other products, although it should be noted that such comparisons are approximate because each product handles non-peat areas in a different way and operates at different resolutions. Finally, the presentation of the AOA and bootstrap model run results of PD-ML clearly demonstrate the model sensitivity to uncertainty in the training data and allows end users to make informed decisions about their use of this data. Comparatively, Hugelius et al. (2020) provide uncertainty ranges for their overall values, but do not explore the uncertainty in as much detail. Widyastuti et al. (2024) acknowledge that an uncertainty assessment was not conducted for PEATGRIDS. A spatially continuous global peat depth distribution, as provided by PD-ML, can then be used to produce more detailed global peat C stock estimates.

4.2 Global Peatland C Stock Estimate Conclusion

Using a simple approach of combining PD-ML output with a fractional peatland coverage map and different literature estimates of bulk density and organic C content, we created viable global peat C stocks of comparable quality to other estimates (i.e, Widyastuti et al., 2024; Hugelius et al., 2020; Page et al., 2011). As discussed in Chapter 1 and 3.2.4, the literature contains peatland C stock estimates ranging from 113 to 1029 Pg (Minasny et al., 2019; Widyastuti et al., 2024), of which all of the PD-ML estimates are within. Looking more closely, each C stock estimate to which PD-ML was specifically compared comes with pros and cons, which results in the quality of these products being similar overall (Figure 3.16). For example, PEAT-GRIDS uses a more complex approach to model organic C content and bulk density globally, giving them a final C stock estimate (1029 Pg C) that is well outside previous estimates (Minasny et al., 2019). Hugelius et al. (2020) applied a more simple linear relationship to estimate peat C stocks based on the depth, however their results are limited to the northern latitudes. Page et al. (2011) take the simplest approach based on single best-estimates for peat depth, bulk density, and organic C content for general regions or countries when making their C stock inventory, thereby losing spatial fidelity. The use of the mean peat depth of Peat-DBase in place of values from PD-ML also results in a loss of spatial fidelity. While C stock estimates calculated from PD-ML are based on single spatially uniform estimates of C density, organic C content, and bulk density, global coverage was maintained and the peat depth values were spatially varying (and informed by an extensive peat depth database in Peat-DBase). Thus, the C inventories emerging from PD-ML appear reasonable in the context of previous research.

At this time, PD-ML's trend towards the mean of its training data likely contributes to its lower peatland C stock values. Thus the quality of these C stock estimates could potentially increase with the future implementation of bias correction of PD-ML outputs. Because the first focus of this thesis was on understanding peat depth, the organic C content and bulk density elements of the PD-ML C stock estimates were kept more simplistic. The application of a more detailed C stock calculation like that of Hugelius et al. (2020) could potentially produce higher quality C stocks based on PD-ML in the future.

Peat-DBase and PD-ML represent comprehensive new datasets of peat depth information. Peat-DBase provides a harmonised global observational database of peat

depth measurements at a size that was not previously available. PD-ML produces a spatially continuous global peat depth dataset that does not rely on a map of peatland extent to indicate non-peat areas and includes thorough uncertainty estimates, features that were absent in previous peat depth maps of similar spatial scales. Taken together, Peat-DBase and PD-ML can serve as the foundations of further peatland research or policy initiatives, such as field campaigns and conservation efforts.

Bibliography

Abatzoglou, J.T., Dobrowski, S.Z., Parks, S.A., Hegewisch, K.C., 2018. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958-2015. Scientific Data 5, 170191. doi:10.1038/sdata.2017.191.

Akumu, C.E., McLaughlin, J.W., 2014. Modeling peatland carbon stock in a delineated portion of the Nayshkootayaow river watershed in Far North, Ontario using an integrated GIS and remote sensing approach. Catena 121, 297–306. doi:10.1016/j.catena.2014.05.025.

Altdorff, D., Bechtold, M., van der Kruk, J., Vereecken, H., Huisman, J.A., 2016. Mapping peat layer properties with multi-coil offset electromagnetic induction and laser scanning elevation data. Geoderma 261, 178–189. doi:10.1016/j.geoderma.2015.07.015.

Amatulli, G., McInerney, D., Sethi, T., Strobl, P., Domisch, S., 2020. Geomorpho90m, empirical evaluation and accuracy assessment of global high-resolution geomorphometric layers. Scientific Data 7, 162. doi:10.1038/s41597-020-0479-6.

Anda, M., Ritung, S., Suryani, E., Sukarman, Hikmat, M., Yatno, E., Mulyani, A., Subandiono, R.E., Suratman, Husnain, 2021. Revisiting tropical peatlands in indonesia: Semi-detailed mapping, extent and depth distribution assessment. Geoderma 402, 115235. doi:10.1016/j.geoderma.2021.115235.

Bansal, S., Creed, I.F., Tangen, B.A., Bridgham, S.D., Desai, A.R., Krauss, K.W., Neubauer, S.C., Noe, G.B., Rosenberry, D.O., Trettin, C., Wickland, K.P., Allen, S.T., Arias-Ortiz, A., Armitage, A.R., Baldocchi, D., Banerjee, K., Bastviken, D., Berg, P., Bogard, M.J., Chow, A.T., Conner, W.H., Craft, C., Creamer, C., DelSontro, T., Duberstein, J.A., Eagle, M., Fennessy, M.S., Finkelstein, S.A., Göckede, M., Grunwald, S., Halabisky, M., Herbert, E., Jahangir, M.M.R., Johnson, O.F.,

Jones, M.C., Kelleway, J.J., Knox, S., Kroeger, K.D., Kuehn, K.A., Lobb, D., Loder, A.L., Ma, S., Maher, D.T., McNicol, G., Meier, J., Middleton, B.A., Mills, C., Mistry, P., Mitra, A., Mobilian, C., Nahlik, A.M., Newman, S., O'Connell, J.L., Oikawa, P., van der Burg, M.P., Schutte, C.A., Song, C., Stagg, C.L., Turner, J., Vargas, R., Waldrop, M.P., Wallin, M.B., Wang, Z.A., Ward, E.J., Willard, D.A., Yarwood, S., Zhu, X., 2023. Practical guide to measuring wetland carbon pools and fluxes. Wetlands (Wilmington) 43, 105. doi:10.1007/s13157-023-01722-2.

Batjes, N.H., Ribeiro, E., van Oostrum, A., 2020. Standardised soil profile data to support global mapping and modelling (WoSIS snapshot 2019). Earth System Science Data 12, 299–320. doi:10.5194/essd-12-299-2020.

Batjes, N.H., Ribeiro, E., Van Oostrum, A., 2019. Standardised soil profile data for the world (WoSIS Snapshot–September 2019). doi:10.17027/isric-wdcsoils.20190901.

Batjes, N.H., Van Oostrum, A.J.M., 2023. World Soil Information Service (WoSIS) Procedures for standardizing soil analytical method descriptions. Technical Report. doi:10.17027/isric-1dq0-1m83.

Bauer, I.E., Davies, M.A., Bona, K.A., Hararuk, O., Shaw, C.H., Thompson, D.K., Kurz, W.A., Webster, K.L., Garneau, M., McLaughlin, J.W., Packalen, M.S., Prystupa, E., Sanderson, N.K., Tarnocai, C., 2024. Peat profile database from peatlands in Canada. Ecology , e4398doi:10.1002/ecy.4398.

Bechtold, M., De Lannoy, G.J.M., Koster, R.D., Reichle, R.H., Mahanama, S.P., Bleuten, W., Bourgault, M.A., Brümmer, C., Burdun, I., Desai, A.R., Devito, K., Grünwald, T., Grygoruk, M., Humphreys, E.R., Klatt, J., Kurbatova, J., Lohila, A., Munir, T.M., Nilsson, M.B., Price, J.S., Röhl, M., Schneider, A., Tiemeyer, B., 2019. PEAT-CLSM: A specific treatment of peatland hydrology in the NASA catchment land surface model. Journal of Advances in Modeling Earth Systems 11, 2130–2162. doi:10.1029/2018MS001574.

Beilman, D.W., MacDonald, G.M., Smith, L.C., Reimer, P.J., 2009. Carbon accumulation in peatlands of West Siberia over the last 2000 years. Global Biogeochemical Cycles 23. doi:10.1029/2007gb003112.

Beilman, D.W., Vitt, D.H., Bhatti, J.S., Forest, S., 2008. Peat carbon stocks in the southern Mackenzie River Basin: uncertainties revealed in a high-resolution case

study: Southern Mackenzie Basin Peat Carbon Stocks. Global Change Biology 14, 1221–1232. doi:10.1111/j.1365-2486.2008.01565.x.

van Bellen, S., Dallaire, P.L., Garneau, M., Bergeron, Y., 2011. Quantifying spatial and temporal Holocene carbon accumulation in ombrotrophic peatlands of the Eastmain region, Quebec, Canada. Global Biogeochemical Cycles doi:10.1029/2010GB003877.

Benfield, A.J., Yu, Z., Benavides, J.C., 2021. Environmental controls over Holocene carbon accumulation in Distichia muscoides-dominated peatlands in the eastern Andes of Colombia. Quaternary Science Reviews 251, 106687. doi:10.1016/j.quascirev.2020.106687.

Blaauw, M., Christen, J.A., 2005. Radiocarbon peat chronologies and environmental change. The Journal of the Royal Statistical Society, Series C (Applied Statistics) 54, 805–816. doi:10.1111/j.1467-9876.2005.00516.x.

Brockett, B.E., Lawson, D.E., 1985. Prototype drill for core sampling fine-grained perennially frozen ground. Technical Report. URL: http://hdl.handle.net/11681/9352.

de Bruin, S., Brus, D.J., Heuvelink, G.B.M., van Ebbenhorst Tengbergen, T., Wadoux, A.M.J.C., 2022. Dealing with clustered samples for assessing map accuracy by cross-validation. Ecological Informatics 69, 101665. doi:10.1016/j.ecoinf.2022.101665.

Brun, P., Zimmermann, N.E., Hari, C., Pellissier, L., Karger, D.N., 2022. CHELSA-BIOCLIM+ a novel set of global climate-related predictors at kilometre-resolution. doi:https://www.doi.org/10.16904/envidat.332.

Canadell, J.G., Monteiro, P.M.S., Costa, M.H., Cotrim da Cunha, L., Cox, P.M., Eliseev, A.V., Henson, S., Ishii, M., Jaccard, S., Koven, C., Lohila, A., Patra, P.K., Piao, S., Rogelj, J., Syampungani, S., Zaehle, S., Zickfeld, K., 2021. 2021: Global Carbon and other Biogeochemical Cycles and Feedbacks. In Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Technical Report. Cambridge, United Kingdom and New York, NY, USA. doi:10.1017/9781009157896.007.

Chadburn, S.E., Burke, E.J., Gallego-Sala, A.V., Smith, N.D., Bret-Harte, M.S., Charman, D.J., Drewer, J., Edgar, C.W., Euskirchen, E.S., Fortuniak, K., Gao, Y., Nakhavali, M., Pawlak, W., Schuur, E.A.G., Westermann, S., 2022. A new approach to simulate peat accumulation, degradation and stability in a global land surface scheme (JULES vn5.8_accumulate_soil) for northern and temperate peatlands. Geoscientific Model Development 15, 1633–1657. doi:10.5194/gmd-15-1633-2022.

Chen, W., Yang, K., Yu, Z., Shi, Y., Chen, C.L.P., 2024. A survey on imbalanced learning: latest research, applications and future directions. Artificial Intelligence Review 57, 1–51. doi:10.1007/s10462-024-10759-6.

Cole, L.E.S., Bhagwat, S.A., Willis, K.J., 2015. Long-term disturbance dynamics and resilience of tropical peat swamp forests. Journal of Ecology 103, 16–30. doi:10.1111/1365-2745.12329.

Comas, X., Terry, N., Slater, L., Warren, M., Kolka, R., Kristiyono, A., Sudiana, N., Nurjaman, D., Darusman, T., 2015. Imaging tropical peatlands in indonesia using ground-penetrating radar (GPR) and electrical resistivity imaging (ERI): implications for carbon stock estimates and peat soil characterization. Biogeosciences 12, 2995–3007. doi:10.5194/bg-12-2995-2015.

Crezee, B., Dargie, G.C., Ewango, C.E.N., Mitchard, E.T.A., Emba B., O., Kanyama T., J., Bola, P., Ndjango, J.B.N., Girkin, N.T., Bocko, Y.E., Ifo, S.A., Hubau, W., Seidensticker, D., Batumike, R., Imani, G., Cuní-Sanchez, A., Kiahtipes, C.A., Lebamba, J., Wotzka, H.P., Bean, H., Baker, T.R., Baird, A.J., Boom, A., Morris, P.J., Page, S.E., Lawson, I.T., Lewis, S.L., 2022. Mapping peat thickness and carbon stocks of the central Congo Basin using field data. Nature Geoscience 15, 639–644. doi:10.1038/s41561-022-00966-7.

Davies, M.A., Blewett, J., Naafs, B.D.A., Finkelstein, S.A., 2021. Ecohydrological controls on apparent rates of peat carbon accumulation in a boreal bog record from the Hudson Bay Lowlands, northern Ontario, Canada. Quaternary Research 104, 14–27. doi:10.1017/qua.2021.22.

Davies, M.A., Mclaughlin, J.W., Packalen, M.S., Finkelstein, S.A., 2023a. Holocene carbon storage and testate amoeba community structure in treed peatlands of the

western Hudson Bay Lowlands margin, Canada. Journal of Quaternary Science 38, 92–106. doi:10.1002/jqs.3465.

Davies, M.A., McLaughlin, J.W., Packalen, M.S., Finkelstein, S.A., 2023b. Using holocene paleo-fire records to estimate carbon stock vulnerabilities in Hudson Bay Lowlands peatlands. Facets (Ott) 8, 1–26. doi:10.1139/facets-2022-0162.

Didan, K., Barreto, A., 2018. VIIRS/NPP vegetation indices 16-day L3 global 500m SIN grid V001. doi:10.5067/VIIRS/VNP13A1.001.

Efron, B., Tibshirani, R.J., 1994. An introduction to the bootstrap. Biometrics 50, 890. doi:10.2307/2532810.

Environment, U.N., 2022. Global peatlands assessment: The state of the world's peatlands. https://www.unep.org/resources/global-peatlands-assessment-2022. Accessed: 2023-4-3.

Fan, Y., Li, H., Miguez-Macho, G., 2013. Global patterns of groundwater table depth. Science 339, 940–943. doi:10.1126/science.1229881.

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., Tu, X.M., 2014. Log-transformation and its implications for data analysis. Shanghai Archives of Psychiatry 26, 105–109. doi:10.3969/j.issn.1002-0829.2014.02.009.

Fenton, J.H.C., 1980. The rate of peat accumulation in antarctic moss banks. Journal of Ecology 68, 211. doi:10.2307/2259252.

Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology 37, 4302–4315. doi:10.1002/joc.5086.

Fluet-Chouinard, E., Stocker, B.D., Zhang, Z., Malhotra, A., Melton, J.R., Poulter, B., Kaplan, J.O., Goldewijk, K.K., Siebert, S., Minayeva, T., Hugelius, G., Joosten, H., Barthelmes, A., Prigent, C., Aires, F., Hoyt, A.M., Davidson, N., Finlayson, C.M., Lehner, B., Jackson, R.B., McIntyre, P.B., 2023. Extensive global wetland loss over the past three centuries. Nature 614, 281–286. doi:10.1038/s41586-022-05572-6.

Friedl, M., Gray, J., Sulla-Menashe, D., 2019. MCD12Q2 MODIS/Terra+Aqua land cover dynamics yearly L3 global 500m SIN grid V006. doi:10.5067/MODIS/MCD12Q2.006.

Galdi, P., Tagliaferri, R., 2019. Data mining: Accuracy and error measures for classification and prediction, in: Ranganathan, S., Gribskov, M., Nakai, K., Schönbach, C. (Eds.), Encyclopedia of Bioinformatics and Computational Biology. Academic Press, Oxford, pp. 431–436. doi:10.1016/B978-0-12-809633-8.20474-3.

Gatis, N., Luscombe, D.J., Carless, D., Parry, L.E., Fyfe, R.M., Harrod, T.R., Brazier, R.E., Anderson, K., 2019. Mapping upland peat depth using airborne radiometric and lidar survey data. Geoderma 335, 78–87. doi:10.1016/j.geoderma.2018.07.041.

Getis, A., 2010. Spatial autocorrelation, in: Fischer, M.M., Getis, A. (Eds.), Handbook of Applied Spatial Analysis. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 255–278. doi:10.1007/978-3-642-03647-7_14.

Getis, A., Ord, J.K., 1992. The analysis of spatial association by use of distance statistics. Geographical Analysis 24, 189–206. doi:10.1111/j.1538-4632.1992.tb00261.x.

Goodling, P., Belitz, K., Stackelberg, P., Fleming, B., 2024. A spatial machine learning model developed from noisy data requires multiscale performance evaluation: Predicting depth to bedrock in the Delaware river basin, USA. Environmental Modelling and Software 179, 106124. doi:10.1016/j.envsoft.2024.106124.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google earth engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment 202, 18–27. doi:10.1016/j.rse.2017.06.031.

Gorham, E., 1957. The development of peat lands. Q. Rev. Biol. 32, 145–166. doi:10.1086/401755.

Gorham, E., 1991. Northern peatlands: Role in the carbon cycle and probable responses to climatic warming. Ecological Applications 1, 182–195. doi:10.2307/1941811.

Gorham, E., Lehman, C., Dyke, A., Clymo, D., Janssens, J., 2012. Long-term carbon sequestration in North American peatlands. Quaternary Science Reviews 58, 77–82. doi:10.1016/j.quascirev.2012.09.018.

Gowan, E.J., Zhang, X., Khosravi, S., Rovere, A., Stocchi, P., Hughes, A.L.C., Gyllencreutz, R., Mangerud, J., Svendsen, J.I., Lohmann, G., 2021. A new global ice sheet reconstruction for the past 80 000 years. Nature Communications 12, 1199. doi:10.1038/s41467-021-21469-w.

Grand-Clement, E., Anderson, K., Smith, D., Angus, M., Luscombe, D.J., Gatis, N., Bray, L.S., Brazier, R.E., 2015. New approaches to the restoration of shallow marginal peatlands. Journal of Environmental Management 161, 417–430. doi:10.1016/j.jenvman.2015.06.023.

Gumbricht, T., Roman-Cuesta, R.M., Verchot, L., Herold, M., Wittmann, F., Householder, E., Herold, N., Murdiyarso, D., 2017. An expert system model for mapping tropical wetlands and peatlands reveals South America as the largest contributor. Global Change Biology 23, 3581–3599. doi:10.1111/gcb.13689.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G., 2017. Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications 73, 220–239. doi:10.1016/j.eswa.2016.12.035.

Hawkins, D.M., 2004. The problem of overfitting. Journal of Chemical Information and Computer Sciences 44, 1–12. doi:10.1021/ci0342472.

Helbig, M., Waddington, J.M., Alekseychik, P., Amiro, B.D., Aurela, M., Barr, A.G., Black, T.A., Blanken, P.D., Carey, S.K., Chen, J., Chi, J., Desai, A.R., Dunn, A., Euskirchen, E.S., Flanagan, L.B., Forbrich, I., Friborg, T., Grelle, A., Harder, S., Heliasz, M., Humphreys, E.R., Ikawa, H., Isabelle, P.E., Iwata, H., Jassal, R., Korkiakoski, M., Kurbatova, J., Kutzbach, L., Lindroth, A., Löfvenius, M.O., Lohila, A., Mammarella, I., Marsh, P., Maximov, T., Melton, J.R., Moore, P.A., Nadeau, D.F., Nicholls, E.M., Nilsson, M.B., Ohta, T., Peichl, M., Petrone, R.M., Petrov, R., Prokushkin, A., Quinton, W.L., Reed, D.E., Roulet, N.T., Runkle, B.R.K., Sonnentag, O., Strachan, I.B., Taillardat, P., Tuittila, E.S., Tuovinen, J.P., Turner, J., Ueyama, M., Varlagin, A., Wilmking, M., Wofsy, S.C., Zyrianov, V., 2020. Increasing contribution of peatlands to boreal evapotranspiration in a warming climate. Nature Climate Change 10, 555–560. doi:10.1038/s41558-020-0763-7.

Hengl, T., 2018. Clay content in % (kg / kg) at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution. doi:10.5281/zenodo.2525663.

Hengl, T., Mendes de Jesus, J., Heuvelink, G.B.M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS One 12, e0169748. doi:10.1371/journal.pone.0169748.

Hengl, T., MacMillan, R.A., 2019. Predictive soil mapping with R. https://soilmapper.org/. Accessed: 2024-10-15.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., Jean-Noël Thépaut, 2020. The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society 146, 1999–2049. doi:10.1002/qj.3803.

Hesterberg, T., 2011. Bootstrap. Wiley Interdisciplinary Reviews: Computational Statistics 3, 497–526. doi:10.1002/wics.182.

Heyvaert, Z., Scherrer, S., Dorigo, W., Bechtold, M., De Lannoy, G., 2024. Joint assimilation of satellite-based surface soil moisture and vegetation conditions into the Noah-MP land surface model. Science of Remote Sensing 9, 100129. doi:10.1016/j.srs.2024.100129.

Householder, J.E., Janovec, J.P., Tobler, M.W., Page, S., Lähteenoja, O., 2012. Peatlands of the Madre de Dios River of Peru: distribution, geomorphology, and habitat diversity. Wetlands 32, 359–368.

Hribljan, J.A., Hough, M., Lilleskov, E.A., Suarez, E., Heckman, K., Planas-Clarke, A.M., Chimner, R.A., 2023. Elevation and temperature are strong predictors of long-term carbon accumulation across tropical Andean mountain peatlands.

Mitigation and Adaptation Strategies for Global Change 29, 1. doi:10.1007/s11027-023-10089-y.

Hribljan, J.A., Suárez, E., Heckman, K.A., Lilleskov, E.A., Chimner, R.A., 2016. Peatland carbon stocks and accumulation rates in the Ecuadorian páramo. Wetlands Ecology and Management 24, 113–127. doi:10.1007/s11273-016-9482-2.

Huang, Z., Zhao, T., Lai, R., Tian, Y., Yang, F., 2023. A comprehensive implementation of the log, Box-Cox and log-sinh transformations for skewed and censored precipitation data. Journal of Hydrology 620, 129347. doi:10.1016/j.jhydrol.2023.129347.

Hugelius, G., Bockheim, J.G., Camill, P., Elberling, B., Grosse, G., Harden, J.W., Johnson, K., Jorgenson, T., Koven, C.D., Kuhry, P., Michaelson, G., Mishra, U., Palmtag, J., Ping, C.L., O'Donnell, J., Schirrmeister, L., Schuur, E.A.G., Sheng, Y., Smith, L.C., Strauss, J., Yu, Z., 2013. A new data set for estimating organic carbon storage to 3 m depth in soils of the northern circumpolar permafrost region. Earth System Science Data 5, 393–402. doi:10.5194/essd-5-393-2013.

Hugelius, G., Loisel, J., Chadburn, S., Jackson, R.B., Jones, M., MacDonald, G., Marushchak, M., Olefeldt, D., Packalen, M., Siewert, M.B., Treat, C., Turetsky, M., Voigt, C., Yu, Z., 2020. Large stocks of peatland carbon and nitrogen are vulnerable to permafrost thaw. Proceedings of the National Academy of Sciences 117, 20438–20446. doi:10.1073/pnas.1916387117.

Immirzi, C.P., Maltby, E., Clymo, R.S., 1992. The global status of peatlands and their role in carbon cycling. Friends of The Earth, London, England. ISBN: 978-18-575-0105-6.

IUSS Working Group WRB, 2022. World Reference Base for Soil Resources. International soil classification system for naming soils and creating legends for soil maps. 4th edition. International Union of Soil Sciences (IUSS), Vienna, Austria.

Jackson, R.B., Lajtha, K., Crow, S.E., Hugelius, G., Kramer, M.G., Piñeiro, G., 2017. The ecology of soil carbon: Pools, vulnerabilities, and biotic and abiotic controls. Annual Review of Ecology, Evolution, and Systematics 48, 419–445. doi:10.1146/annurev-ecolsys-112414-054234.

Jeglum, J.K., Rothwell, R.L., Berry, G.J., Smith, G.K.M., 1991. New volumetric sampler increases speed and accuracy of peat surveys. Technical Report Frontline Technical Note 9. Forestry Canada, Ontario Region, Sault Ste. Marie, Ontario. URL: https://ostrnrcan-dostrncan.canada.ca/handle/1845/242503.

Johnson, R.W., 2001. An introduction to the bootstrap. Teaching Statistics 23, 49–54. doi:10.1111/1467-9639.00050.

Joosten, H., Clarke, D., 2002. Wise Use of Mires and Peatlands. International Mire Conservation Group and International Peat Society. ISBN: 978-951-97744-8-0.

Jowsey, P.C., 1966. An improved peat sampler. New Phytologist 65, 245–248. doi:10.1111/j.1469-8137.1966.tb06356.x.

Karger, D.N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W., Zimmermann, N.E., Linder, H.P., Kessler, M., 2017. Climatologies at high resolution for the earth's land surface areas. Scientific Data 4, 170122. doi:10.1038/sdata.2017.122.

Karger, D.N., Wilson, A.M., Mahony, C., Zimmermann, N.E., Jetz, W., 2021. Global daily 1 km land surface precipitation based on cloud cover-informed downscaling. Scientific Data 8, 307. doi:10.1038/s41597-021-01084-6.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y., 2017. LightGBM: A highly efficient gradient boosting decision tree. 31st Conference on Neural Information Processing Systems .

Kelley, D.I., Prentice, I.C., Harrison, S.P., Wang, H., Simard, M., Fisher, J.B., Willis, K.O., 2013. A comprehensive benchmarking system for evaluating global vegetation models. Biogeosciences 10, 3313–3340. doi:10.5194/bg-10-3313-2013.

Kelly, T.J., Lawson, I.T., Roucoux, K.H., Baker, T.R., Honorio Coronado, E.N., 2020. Patterns and drivers of development in a west Amazonian peatland during the late Holocene. Quaternary Science Reviews 230, 106168. doi:10.1016/j.quascirev.2020.106168.

Keys, D., Henderson, R.E., 1987. An investigation of the peat resources of New Brunswick. Technical Report Minerals and Petroleum Publications and Assessment Reports Information System (PARIS) OF 83-10. New Brunswick Depart-

ment of Natural Resources and Energy. ISSN: 0712-4562 URL: https://dnr-mrn.gnb.ca/ParisWeb/PublicationDetails.aspx.

Koster, E., Favier, T., 2005. Peatlands, past and present, in: The Physical Geography of Western Europe. Oxford University Press. doi:10.1093/oso/9780199277759.003.0018.

Krankina, O.N., Pflugmacher, D., Friedl, M., Cohen, W.B., Nelson, P., Baccini, A., 2008. Meeting the challenge of mapping peatlands with remotely sensed data. Biogeosciences 5, 2075–2101. doi:10.5194/bgd-5-2075-2008.

Krawczyk, B., 2016. Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence 5, 221–232. doi:10.1007/s13748-016-0094-0.

Kuhn, M., 2008. Building predictive models in R Using the caret Package. Journal of Statistical Software 28, 1–26. doi:10.18637/jss.v028.i05.

Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer, New York.

Lawson, I.T., Åkesson, C., Baker, T.R., Cordova Oroche, C.J., Dargie, G.C., del Aguila-Pasquel, J., Grandez Ríos, J., Hastie, A., Honorio Coronado, E.N., Mitchard, E.T., Roucoux, K.H., Reyna Huaymacari, J., Williams, M., 2023. Peat depths from the Pastaza-Marañón Basin, Amazonian Peru, 2019-2020. doi:10.5285/ab13a06f-392f-4bc6-b1bf-06dd8b020307.

Loidi, J., Navarro-Sánchez, G., Vynokurov, D., 2023. A vector map of the world's terrestrial biotic units: subbiomes, biomes, ecozones and domains. Vegetation Classification and Survey 4, 59–61. doi:10.3897/vcs.99167.

Loisel, J., van Bellen, S., Pelletier, L., Talbot, J., Hugelius, G., Karran, D., Yu, Z., Nichols, J., Holmquist, J., 2017. Insights and issues with estimating northern peatland carbon stocks and fluxes since the Last Glacial Maximum. Earth-Science Reviews 165, 59–80. doi:10.1016/j.earscirev.2016.12.001.

Lourenco, M., Fitchett, J.M., Woodborne, S., 2022. Peat definitions: A critical review. Progress in Physical Geography: Earth and Environment , 03091333221118353doi:10.1177/03091333221118353.

Lähteenoja, O., Reátegui, Y.R., Räsänen, M., Torres, D.D.C., Oinonen, M., Page, S., 2012. The large Amazonian peatland carbon sink in the subsiding Pastaza-Marañón foreland basin, Peru. Global Change Biology 18, 164–178. doi:10.1111/j.1365-2486.2011.02504.x.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52. doi:10.1016/S0016-7061(03)00223-4.

Melton, J.R., Arora, V.K., Wisernig-Cojoc, E., Seiler, C., Fortier, M., Chan, E., Teckentrup, L., 2020. CLASSIC v1.0: the open-source community successor to the Canadian Land Surface Scheme (CLASS) and the Canadian Terrestrial Ecosystem Model (CTEM) – Part 1: Model framework and site-level performance. Geoscientific Model Development 13, 2825–2850. doi:10.5194/gmd-13-2825-2020.

Melton, J.R., Chan, E., Millard, K., Fortier, M., Winton, R.S., Martín-López, J.M., Cadillo-Quiroz, H., Kidd, D., Verchot, L.V., 2022. A map of global peatland extent created using machine learning (Peat-ML). Geoscientific Model Development 15, 4709–4738. doi:10.5194/gmd-15-4709-2022.

Meyer, H., Pebesma, E., 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods in Ecology and Evolution 12, 1620–1633. doi:10.1111/2041-210x.13650.

Meyer, H., Pebesma, E., 2022. Machine learning-based global maps of ecological variables and the challenge of assessing them. Nature Communications 13, 2208. doi:10.1038/s41467-022-29838-9.

Meyer, H., Reudenbach, C., Wöllauer, S., Nauss, T., 2019. Importance of spatial predictor variable selection in machine learning applications – moving from data reproduction to spatial prediction. Ecological Modelling 411, 108815. doi:10.1016/j.ecolmodel.2019.108815.

Milà, C., Mateu, J., Pebesma, E., Meyer, H., 2022. Nearest neighbour distance matching Leave-One-Out Cross-Validation for map validation. Methods in Ecology and Evolution 13, 1304–1316. doi:10.1111/2041-210x.13851.

Minasny, B., Berglund, O., Connolly, J., Hedley, C., de Vries, F., Gimona, A., Kempen, B., Kidd, D., Lilja, H., Malone, B., McBratney, A., Roudier, P., O'Rourke, S.,

Rudiyanto, Padarian, J., Poggio, L., ten Caten, A., Thompson, D., Tuve, C., Widyatmanti, W., 2019. Digital mapping of peatlands – a critical review. Earth-Science Reviews 196, 102870. doi:10.1016/j.earscirev.2019.05.014.

Moore, P.D., 1989. The ecology of peat-forming processes: a review. Int. J. Coal Geol. 12, 89–103. doi:10.1016/0166-5162(89)90048-7.

Müller, J., Joos, F., 2020. Global peatland area and carbon dynamics from the Last Glacial Maximum to the present – a process-based model investigation. Biogeosciences 17, 5285–5308. doi:10.5194/bg-17-5285-2020.

Ofiti, N.O.E., Schmidt, M.W.I., Abiven, S., Hanson, P.J., Iversen, C.M., Wilson, R.M., Kostka, J.E., Wiesenberg, G.L.B., Malhotra, A., 2023. Climate warming and elevated CO2 alter peatland soil carbon sources and stability. Nature Communications 14, 7533. doi:10.1038/s41467-023-43410-z.

Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D., Powell, G.V.N., Underwood, E.C., Itoua, I., Strand, H.E., Morrison, J.C., Loucks, C.J., Allnutt, T.F., Ricketts, T.H., Kura, Y., Lamoreux, J.F., Wettengel, W.W., Hedao, P., Kassem, K.R., 2001. Terrestrial ecoregions of the world: A new map of life on earth: A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. Bioscience 51, 933–938. doi:10.1641/0006-3568(2001)051[0933:TEOTWA]2.0.CO;2.

ONeill, P.E., Chan, S., Njoku, E.G., Jackson, T., Bindlish, R., Chaubell, M.J., Colliander, A., 2021. SMAP enhanced L3 radiometer global and polar grid daily 9 km EASE-grid soil moisture, version 5. doi:10.5067/4DQ54OUIJ9DL.

Ord, J.K., Getis, A., 1995. Local spatial autocorrelation statistics: Distributional issues and an application. Geographical Analysis 27, 286–306. doi:10.1111/j.1538-4632.1995.tb00912.x.

Osman, K.T., 2013. Organic matter of forest soils, in: Forest Soils. Springer International Publishing, Cham, pp. 63–76. doi:10.1007/978-3-319-02541-4_4.

Page, S.E., Baird, A.J., 2016. Peatlands and global change: Response and resilience. Annual Review of Environment and Resources 41, 35–57. doi:10.1146/annurev-environ-110615-085520.

Page, S.E., Banks, C.J., Rieley, J.O., 2007. Tropical peatlands: Distribution, extent and carbon storage-uncertainties and knowledge gaps. Peatlands International Volume 2. ISSN:1455-8491.

Page, S.E., Rieley, J.O., Banks, C.J., 2011. Global and regional importance of the tropical peatland carbon pool. Global Change Biology 17, 798–818. doi:10.1111/j.1365-2486.2010.02279.x.

Parry, L.E., West, L.J., Holden, J., Chapman, P.J., 2014. Evaluating approaches for estimating peat depth. Journal of Geophysical Research: Biogeosciences 119, 567–576. doi:10.1002/2013JG002411.

Pekel, J.F., Cottam, A., Gorelick, N., Belward, A.S., 2016. High-resolution mapping of global surface water and its long-term changes. Nature 540, 418–422. doi:10.1038/nature20584.

Peng, Y., Nagata, M.H., 2020. An empirical overview of nonlinearity and overfitting in machine learning using COVID-19 data. Chaos Solitons Fractals 139, 110055. doi:10.1016/j.chaos.2020.110055.

Plevris, V., Solorzano, G., Bakas, N., Ben Seghier, M., 2022. Investigation of performance metrics in regression analysis and machine learning-based prediction models, in: 8th European Congress on Computational Methods in Applied Sciences and Engineering, CIMNE. doi:10.23967/eccomas.2022.155.

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., Pélissier, R., 2020. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nature Communications 11, 4540. doi:10.1038/s41467-020-18321-y.

Quik, C., Palstra, S.W.L., van Beek, R., van der Velde, Y., Candel, J.H.J., van der Linden, M., Kubiak-Martens, L., Swindles, G.T., Makaske, B., Wallinga, J., 2022. Dating basal peat: The geochronology of peat initiation revisited. Quaternary Geochronology 72, 101278. doi:10.1016/j.quageo.2022.101278.

Ratnayake, A.S., 2020. Characteristics of lowland tropical peatlands: Formation, classification, and decomposition. Journal of Tropical Forestry and Environment 10. doi:10.31357/jtfe.v10i1.4685.

Ribeiro, K., Pacheco, F.S., Ferreira, J.W., de Sousa-Neto, E.R., Hastie, A., Krieger Filho, G.C., Alvalá, P.C., Forti, M.C., Ometto, J.P., 2021. Tropical peatlands and their contribution to the global carbon cycle and climate change. Global Change Biology 27, 489–505. doi:10.1111/gcb.15408.

Rosa, E., Larocque, M., Pellerin, S., Gagné, S., Fournier, B., 2009. Determining the number of manual measurements required to improve peat thickness estimations by ground penetrating radar. Earth Surface Processes and Landforms 34, 377–383. doi:10.1002/esp.1741.

Rouault, E., Warmerdam, F., Schwehr, K., Kiselev, A., Butler, H., Łoskot, M., Szekeres, T., Tourigny, E., Landa, M., Miara, I., Elliston, B., Chaitanya, K., Plesea, L., Morissette, D., Jolma, A., Dawson, N., 2023. GDAL. doi:10.5281/ZENODO.7764163.

Rožanec, J.M., Petelin, G., Costa, J., Bertalanič, B., Cerar, G., Guček, M., Papa, G., Mladenić, D., 2023. Dealing with zero-inflated data: achieving SOTA with a two-fold machine learning approach. arXiv [cs.LG] arXiv:2310.08088.

Rudiyanto, Minasny, B., Setiawan, B.I., Arif, C., Saptomo, S.K., Chadirin, Y., 2016. Digital mapping for cost-effective and accurate prediction of the depth and carbon stocks in Indonesian peatlands. Geoderma 272, 20–31. doi:10.1016/j.geoderma.2016.02.026.

Running, S., Mu, Q., Zhao, M., 2011. MOD17A3 MODIS/terra net primary production yearly L4 global 1km SIN grid V055 [data set]. NASA EOSDIS Land Processes DAAC .

Ruppel, M., Väliranta, M., Virtanen, T., Korhola, A., 2013. Postglacial spatiotemporal peatland initiation and lateral expansion dynamics in North America and northern Europe. Holocene 23, 1596–1606. doi:10.1177/0959683613499053.

Russell, S., Norvig, P., 2020. Chapter 19. Learning from Examples, in: Artificial Intelligence: A modern approach. 4 ed.. Pearson, Upper Saddle River, NJ. ISBN:9780134610993.

Ruwaimana, M., Anshari, G.Z., Silva, L.C.R., Gavin, D.G., 2020. The oldest extant tropical peatland in the world: a major carbon reservoir for at least 47 000 years. Environmental Research Letters 15, 114027. doi:10.1088/1748-9326/abb853.

Rydin, H., Jeglum, J.K., 2013a. Peatland habitats, in: The Biology of Peatlands. Oxford University Press, pp. 1–20. doi:10.1093/acprof:osobl/9780199602995.003.0001.

Rydin, H., Jeglum, J.K., 2013b. Peatland succession and development, in: Rydin, H., Jeglum, J.K., Bennett, K.D. (Eds.), The Biology of Peatlands, 2nd ed. Oxford : Oxford University Press, p. 127–147. doi:10.1093/acprof:osobl/9780199602995.003.0007.

Rydin, H., Jeglum, J.K., 2013c. The Biology of Peatlands. Biology of Habitats Series. 2 ed., Oxford University Press, London, England. doi:10.1093/acprof:osobl/9780199602995.001.0001.

Sarracino, F., Mikucka, M., 2017. Bias and efficiency loss in regression estimates due to duplicated observations: a monte carlo simulation. SRM doi:10.18148/srm/2017.v11i1.7149.

Schulzweida, U., 2022. CDO user guide. doi:10.5281/ZENODO.7112925.

Seiler, C., Melton, J.R., Arora, V.K., Wang, L., 2020. CLASSIC v1.0: the open-source community successor to the Canadian Land Surface Scheme (CLASS) and the Canadian Terrestrial Ecosystem Model (CTEM) – Part 2: Global Benchmarking doi:10.5194/gmd-2020-294.

Shimada, M., Itoh, T., Motooka, T., Watanabe, M., Shiraishi, T., Thapa, R., Lucas, R., 2014. New global forest/non-forest maps from ALOS PALSAR data (2007–2010). Remote Sensing of Environment 155, 13–31. doi:10.1016/j.rse.2014.04.014.

Shotyk, W., Noernberg, T., 2020. Sampling, handling, and preparation of peat cores from bogs: review of recent progress and perspectives for trace element research. Canadian Journal of Soil Science 100, 363–380. doi:10.1139/cjss-2019-0160.

Silvestri, S., Knight, R., Viezzoli, A., Richardson, C.J., Anshari, G.Z., Dewar, N., Flanagan, N., Comas, X., 2019a. Field-lab data and analyses results West Kalimantan, Indonesia. doi:10.5281/zenodo.3572061.

Silvestri, S., Knight, R., Viezzoli, A., Richardson, C.J., Anshari, G.Z., Dewar, N., Flanagan, N., Comas, X., 2019b. Quantification of peat thickness and stored carbon

at the landscape scale in tropical peatlands: A comparison of airborne geophysics and an empirical topographic method. Journal of Geophysical Research: Earth Surface 124, 3107–3123. doi:10.1029/2019jf005273.

Smith, K.B., Smith, C.E., Forest, S.F., Richard, A.J., 2007. A Field Guide to the Wetlands of the Boreal Plains Ecozone of Canada. Technical Report. Ducks Unlimited Canada, Western Boreal Office: Edmonton, Alberta.

Soil Classification Working Group, 1998. The Canadian System of Soil Classification, 3rd ed, in: Agriculture and Agri-Food Canada Publication 1646, p. 187pp.

Soil Survey Staff, 1999. Soil taxonomy: basic system soil classification making interpreting soil surveys, in: *Natural Resources Conservation Service. U.S. Department of Agriculture Handbook*, p. 436.

Somarathna, P.D.S.N., Minasny, B., Malone, B.P., 2017. More data or a better model? Figuring out what matters most for the spatial prediction of soil carbon. Soil Science Society of America Journal 81, 1413–1426. doi:10.2136/sssaj2016.11.0376.

Song, J., 2015. Bias corrections for random forest in regression using residual rotation. Journal of the Korean Statistical Society 44, 321–326. doi:10.1016/j.jkss.2015.01.003.

Sun, J., Gallego-Sala, A., Yu, Z., 2023. Topographic and climatic controls of peatland distribution on the Tibetan Plateau. Scientific Reports 13, 14811. doi:10.1038/s41598-023-39699-x.

Tarnocai, C., Kettles, I.M., Lacelle, B., 2011. Peatlands of Canada. Technical Report. doi:10.4095/288786.

Thibault, J., 1992. The New Brunswick Peatland Database. Technical Report Minerals and Petroleum Publications and Assessment Reports Information System (PARIS) MRR 6. New Brunswick Department of Natural Resources and Energy Development. URL: https://dnr-mrn.gnb.ca/ParisWeb/PublicationDetails.aspx.

Treat, C.C., Broothaerts, N., Dalton, A.S., Dommain, R., Douglas, T., Drexler, J., Finkelstein, S.A., Grosse, G., Hope, G., Hutchings, J.A., Jones, M.C., Kleinen, T., Kuhry, P., Lacourse, T., Lähteenoja, O., Loisel, J., Notebaert, B., Payne, R.J., Peteet, D.M., Sannel, A.B.K., Stelling, J., Strauss, J., Swindles, G.T., Talbot, J.,

Tarnocai, C., Verstraeten, G., Williams, C.J., Xia, Z., Yu, Z., Brovkin, V., 2017. (table S2) global dataset of peatland basal ages. doi:10.1594/PANGAEA.873065.

Treat, C.C., Kleinen, T., Broothaerts, N., Dalton, A.S., Dommain, R., Douglas, T.A., Drexler, J.Z., Finkelstein, S.A., Grosse, G., Hope, G., Hutchings, J., Jones, M.C., Kuhry, P., Lacourse, T., Lähteenoja, O., Loisel, J., Notebaert, B., Payne, R.J., Peteet, D.M., Sannel, A.B.K., Stelling, J.M., Strauss, J., Swindles, G.T., Talbot, J., Tarnocai, C., Verstraeten, G., Williams, C.J., Xia, Z., Yu, Z., Väliranta, M., Hättestrand, M., Alexanderson, H., Brovkin, V., 2019. Widespread global peatland establishment and persistence over the last 130,000 y. Proceedings of the National Academy of Sciences 116, 4822–4827. doi:10.1073/pnas.1813305116.

Turetsky, M.R., St. Louis, V.L., 2006. Disturbance in boreal peatlands, in: Wieder, R.K., Vitt, D.H. (Eds.), Boreal Peatland Ecosystems. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 359–379. doi:10.1007/978-3-540-31913-9_16.

Turunen, J., Tomppo, E., Tolonen, K., Reinikainen, A., 2002. Estimating carbon accumulation rates of undrained mires in Finland–application to boreal and subarctic regions. Holocene 12, 69–80. doi:10.1191/0959683602hl522rp.

United Nations Environment Programme, 2021. The global peatland map 2.0 URL: https://wedocs.unep.org/20.500.11822/37571.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors, 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods 17, 261–272. doi:10.1038/s41592-019-0686-2.

Voroney, R.P., Heck, R.J., Kuzyakov, Y., 2024. The habitat of the soil biota, in: Soil Microbiology, Ecology and Biochemistry. Elsevier, pp. 13–40. doi:10.1016/b978-0-12-822941-5.00002-8.

Wade, C., 2020. *Chapter 4*: From gradient boosting to XGBoost, in: Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform accessible machine learning

and extreme gradient boosting with Python. Packt Publishing, Birmingham, England.

Wang, B., Ding, Q., 2008. Global monsoon: Dominant mode of annual variation in the tropics. Dynamics of Atmospheres and Oceans 44, 165–183. doi:10.1016/j.dynatmoce.2007.05.002.

Wang, D., 2021. MODIS/Terra+Aqua photosynthetically active radiation daily/3-hour L3 global 0.05Deg CMG V061. https://lpdaac.usgs.gov/products/mcd18c2v061/. doi:10.5067/MODIS/MCD18C2.061. accessed: 2023-7-31.

Warren, M., Hergoualc'h, K., Kauffman, J.B., Murdiyarso, D., Kolka, R., 2017. An appraisal of Indonesia's immense peat carbon stock using national peatland maps: uncertainties and potential losses from conversion. Carbon Balance and Management 12, 12. doi:10.1186/s13021-017-0080-2.

Warren, M.W., Kauffman, J.B., Murdiyarso, D., Anshari, G., Hergoualc'h, K., Kurnianto, S., Purbopuspito, J., Gusmayanti, E., Afifudin, M., Rahajoe, J., Alhamd, L., Limin, S., Iswandi, A., 2012. A cost-efficient method to assess carbon stocks in tropical peat soil. Biogeosciences 9, 4477–4485. doi:10.5194/bg-9-4477-2012.

West, R.M., 2022. Best practice in statistics: The use of log transformation. Annals of Clinical Biochemistry 59, 162–165. doi:10.1177/00045632211050531.

Widyastuti, M.T., Minasny, B., Padarian, J., Maggi, F., Aitkenhead, M., Beucher, A., Connolly, J., Fiantis, D., Kidd, D., Ma, Y., Macfarlane, F., Robb, C., Rudiyanto, Setiawan, B.I., Taufik, M., 2024. PEATGRIDS: Mapping thickness and carbon stock of global peatlands via digital soil mapping. Earth System Science Data Discussions , 1–29doi:10.5194/essd-2024-333.

Wilson, M.J., 2019. The importance of parent material in soil classification: A review in a historical context. Catena 182, 104131. doi:10.1016/j.catena.2019.104131.

Wu, Y., Verseghy, D.L., Melton, J.R., 2016. Integrating peatlands into the coupled Canadian Land Surface Scheme (CLASS) v3.6 and the Canadian Terrestrial Ecosystem Model (CTEM) v2.0. Geoscientific Model Development 9, 2639–2663. doi:10.5194/gmd-9-2639-2016.

Xu, J., Morris, P.J., Liu, J., Holden, J., 2018. PEATMAP: Refining estimates of global peatland distribution based on a meta-analysis. Catena 160, 134–140. doi:10.1016/j.catena.2017.09.010.

Xu, L., Saatchi, S.S., Yang, Y., Yu, Y., White, L., 2016. Performance of non-parametric algorithms for spatial mapping of tropical forest structure. Carbon Balance and Management 11, 18. doi:10.1186/s13021-016-0062-9.

Yu, Z., Loisel, J., Brosseau, D.P., Beilman, D.W., Hunt, S.J., 2010. Global peatland dynamics since the Last Glacial Maximum. Geophysical Research Letters 37. doi:10.1029/2010gl043584.

Zender, C.S., 2008. Analysis of self-describing gridded geoscience data with netCDF operators (NCO). Environmental Modelling & Software 23, 1338–1342. doi:10.1016/j.envsoft.2008.03.004.

Zeng, J., Zhang, Q., 2020. The trends in land surface heat fluxes over global monsoon domains and their responses to monsoon and precipitation. Scientific Reports 10, 5762. doi:10.1038/s41598-020-62467-0.

Zhang, G., Lu, Y., 2012. Bias-corrected random forests in regression. J. Appl. Stat. 39, 151–160. doi:10.1080/02664763.2011.578621.

Zhao, Y., Tang, Y., Yu, Z., Li, H., Yang, B., Zhao, W., Li, F., Li, Q., 2014. Holocene peatland initiation, lateral expansion, and carbon dynamics in the Zoige Basin of the eastern Tibetan Plateau. Holocene 24, 1137–1145. doi:10.1177/0959683614538077.

Zinck, J.A., 2011. Tropical and subtropical peats: An overview, in: Zinck, J.A., Huber, O. (Eds.), Peatlands of the Western Guayana Highlands, Venezuela: Properties and Paleogeographic Significance of Peats. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 5–28. doi:10.1007/978-3-642-20138-7_2.