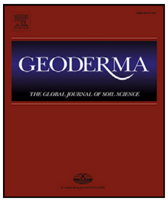
Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma



Benefits of hierarchical predictions for digital soil mapping—An approach to map bimodal soil pH

Madlene Nussbaum [a,*], Stephan Zimmermann [b], Lorenz Walthert [b], Andri Baltensweiler [b]

[a] Bern University of Applied Sciences (BFH), School for Agricultural, Forest and Food Sciences (HAFL), Länggasse 85, 3052 Zollikofen, Switzerland
[b] Swiss Federal Institute for Forest, Snow and Landscape Research (WSL), Zürcherstrasse 111, 8903 Birmensdorf, Switzerland

ARTICLE INFO

Handling Editor: Budiman Minasny

Keywords:
Digital soil mapping
Hierarchical predictions
Bimodal response
Soil pH
Acidity buffer ranges
Prediction uncertainty

ABSTRACT

Maps of soil pH are an important tool for making decisions in sustainable forest management. Accurate pH mapping, therefore, is crucial to support decisions by authorities or forest companies. Soil pH values typically exhibit a distinct distribution characterized by two frequently encountered pH ranges, wherein aluminium oxides (Al_2O_3) and carbonates ($CaCO_3$) act as the primary buffer agents. Soil samples with moderately acid pH values (pH $CaCl_2$ of 4.5-6) are less commonly observed due to their weaker buffering capacity. The different strength of buffer agents results in a distinct bimodal distribution of soil pH values with peaks at pH of around 4 and 7.5. Commonly used approaches for spatial mapping neglect this often observed characteristic of soil pH and predict unimodal distributions with too many moderately acid pH values. For ecological map applications this might result in misleading interpretations.

This article presents a novel approach to produce pH maps that are able to reproduce pedogenic processes. The procedure is suitable for bimodal responses where the response distribution is naturally inherent and needs to be reproduced for the predictions. It is model-agnostic, namely independent from the used statistical prediction method. Calibration data is optimally split into two parts corresponding each to a data culmination, i.e. for soil pH values belonging to the ranges of the two principal buffer agents (Al_2O_3 and $CaCO_3$). For each subset a separate model is then built. In addition, a binary model is fitted to assign every new prediction location a probability to belong either to Al_2O_3 or $CaCO_3$ buffer range. Predictions are combined by weighted mean. Weights are derived from probabilities predicted by the binary model. Degree of smoothness is chosen by sigmoid transform which allows for optimal continuous transition of the pH values between Al_2O_3 and $CaCO_3$ buffer ranges. For each location uncertainty distributions may be combined by using the same weights.

We illustrated application of the new approach to a medium and strong bimodal distributed response (1) pH in 0–5 cm and (2) pH in 60–100 cm of forest soils in Switzerland (2 530 calibration sites). While model performance measured at 354 validation sites slightly dropped compared to a common modelling approach (drop of R^2 of 0.02–0.03) distributional properties of the predictions are much more meaningful from a pedogenic point of view. We were able to demonstrate the benefits of considering specific distributional properties of responses within the prediction process and expanded model assessment by comparing observed and predicted distributions.

1. Introduction

Soil pH is critical for a wide range of applications and stakeholders. Besides providing information on relative acidity and alkalinity, pH is an indication of nutrient availability and has an impact on physical structure, metal dissolution, decomposition processes or (micro-)biological activity (Blume et al., 2016, Sec. 5.6). Spatial information on soil pH supports soil function assessments like acidity of arable land to plan liming (Bolan et al., 2003), acidity status of forests (Zimmermann et al., 2011) or to assess filtering or binding and decomposition capacity of contaminants or pollutants (Bechler and Toth, 2010; Greiner et al., 2017) .

Soil pH datasets commonly exhibit a characteristic bimodal distribution observed at local (Baltensweiler et al., 2020), regional (Vaysse et al., 2017), national (Baltensweiler et al., 2021; Helfenstein et al.,

* Corresponding author.
E-mail address: madlene.nussbaum@bfh.ch (M. Nussbaum).

https://doi.org/10.1016/j.geoderma.2023.116579
Received 24 March 2023; Received in revised form 19 June 2023; Accepted 20 June 2023
Available online 14 July 2023

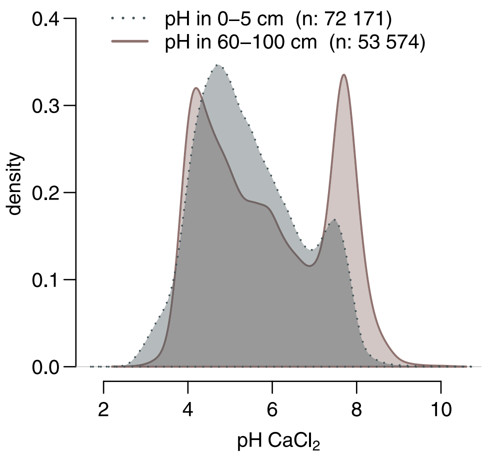


Fig. 1. Density plots of soil pH ($CaCl_2$ extraction) in 0–5 and 60–100 cm depth for complete 2019 snapshot of the World Soil Data Base (WoSIS, Batjes et al., 2019, n: number of observations). Topsoil pH (dotted grey) exhibits weak and pH in deeper soil layers (solid red) strong bimodal distribution.

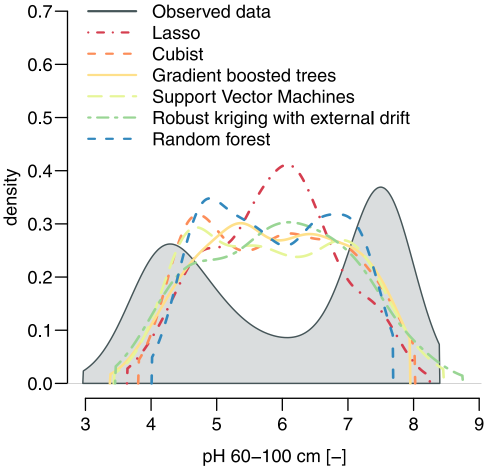


Fig. 2. Density of predictions computed for the validation data not used for modelling (n = 325, Section 3.4) by a variety of methods fitted to response pH in 60–100 cm in Swiss forest soils (calibration data n = 2357) compared to the density of original observed validation data (solid line with grey fill, densities cropped at minimum and maximum value; for details on methods and their application see Baltensweiler et al., 2021).

2022) or global level (Fig. 1) with often more pronounced bimodality with increasing depth (Mulder et al., 2016). During soil development natural acidification takes place in the course of which pH value cascades down the soil profile according to the effective buffering processes. In carbonate parent materials or carbonate-bearing mixed rocks, carbonates ($CaCO_3$) buffer incoming acids very efficiently (Blume et al., 2016, Sec. 5.6.4). After carbonates get depleted mainly silicates and the cation exchangers become buffer agents (pH $CaCl_2$ range between 4.5 and 6.5). Silicates are often present in large quantities. However, their buffer efficiency is limited because buffer reactions are kinetically slow. Exchange kinetics of cation exchangers are very fast, but capacity of exchangers is limited. Overall, buffer efficiency is limited in the pH range between 4.5 and 6.5 due to kinetic inhibition or limited buffer capacity. The pH value of fine earth of a horizon, therefore, passes through this range relatively quickly during acidification and only few soil samples are found with a moderately acid pH value. As soon as buffering by aluminium oxides (Al_2O_3) becomes dominant, buffering efficiency is significantly larger, since there are abundant oxides in the soil and the buffering reactions are not kinetically inhibited. This leads to a bimodal pH value distribution with a maximum of values in the pH ranges efficiently buffered by $CaCO_3$ and Al_2O_3 (Sparks et al., 2023, chap. 9). Although bimodality is for above reasons to be expected in many pH datasets it might not always be present, i.e. in datasets with a large share of arable land that is artificially kept at pH levels between 5.5–6.5 by application of lime (Bolan et al., 2003) or if alkaline conditions are rare as found by Roudier et al. (2020) for New Zealand.

Producing maps of basic soil properties by digital soil mapping is an established practice (Malone et al., 2017). Mapping of soil pH has been of interest for a long time (Laslett et al., 1987; Makungwe et al., 2021). According to GlobalSoilMap specification soil pH is among the relevant soil properties to be mapped globally (Arrouays et al., 2014b). In recent national to continental digital soil mapping efforts pH was among the most frequently predicted properties just after soil organic carbon and texture (Chen et al., 2022). SoilGrids, a global pH soil map, was recently updated (version 2.0, Poggio et al., 2021).

The largest error in quantitative mapping of pH for a given site results among other error sources (e.g. spatial accuracy of location, measurement method, pedotransfer functions) from the spatial prediction (Libohova et al., 2019). To our best knowledge no mapping study specifically addressed bimodality of soil pH datasets in their prediction approach. Some authors discussed bimodality of their datasets (Mulder et al., 2016; Baltensweiler et al., 2021; Helfenstein et al., 2022), but for most studies it remains unclear if their pH data set was bimodal or not. Even if a study focused mainly or solely on soil pH (e.g. Adhikari et al., 2014; Wang et al., 2019; Lu et al., 2023) reported descriptive statistics of moments were not able to reveal such a distributional feature nor was it mentioned or shown in corresponding graphs.

If no covariate is able to distinguish the underlying processes of the two data peaks, unimodal approaches have difficulty reproducing the original distribution. Even models with relaxed assumptions on error distribution such as random forest (RF) or gradient boosted trees fail to predict the original distribution (Fig. 2). Neglecting bimodality may result in maps with unimodal distributions (Fig. 2) with a large number of predictions in a soil pH range where only few horizons were observed. As a result strongly acid soil horizons may be represented as moderately acid (i.e. not yet in the Al_2O_3 buffer range) and alkaline horizons as slightly acidic (i.e. already beyond $CaCO_3$ buffer capacity and prone to faster acidification). Subsequent map interpretations for example for ecological applications may then result in misleading conclusions. Moreover, ignoring bimodal response distribution may result in over-pessimistic prediction intervals (Helfenstein et al., 2022).

Bimodal or multimodal distributions of response variables are not limited to soil pH. Styc and Lagacherie (2021) reported bimodal distributed plant available soil water capacity in deeper soil depth, likely due to uneven particle size distribution resulting in two size levels of soil pores (Zhang and Chen, 2005). Rawlins et al. (2009) and de Brogniez et al. (2015) found bimodal soil organic carbon (SOC) data because of large areas of peat within their area of focus (Northern Ireland, Europe). de Brogniez et al. (2015) used a default unimodal modelling approach, but concluded that this might have been the cause for their underestimation of large SOC contents. Rawlins et al. (2009) dealt with bimodality by splitting the study area into mineral, organo-mineral and peat according to an existing soil map and fitted separate geostatistical models to each sub-dataset. To deal with a bimodal response in econometrics (Gostkowski and Gajowniczek, 2020) adapted quantile regression forest (QRF) with performance based weights for each tree with minor benefit in reproducing bimodal prediction distribution compared to default QRF.

Using soil pH we illustrate how to introduce knowledge on soil pedogenic processes into digital soil mapping (Wadoux et al., 2021) and how to retain the specific distributional feature of bimodality of a response within predictions.

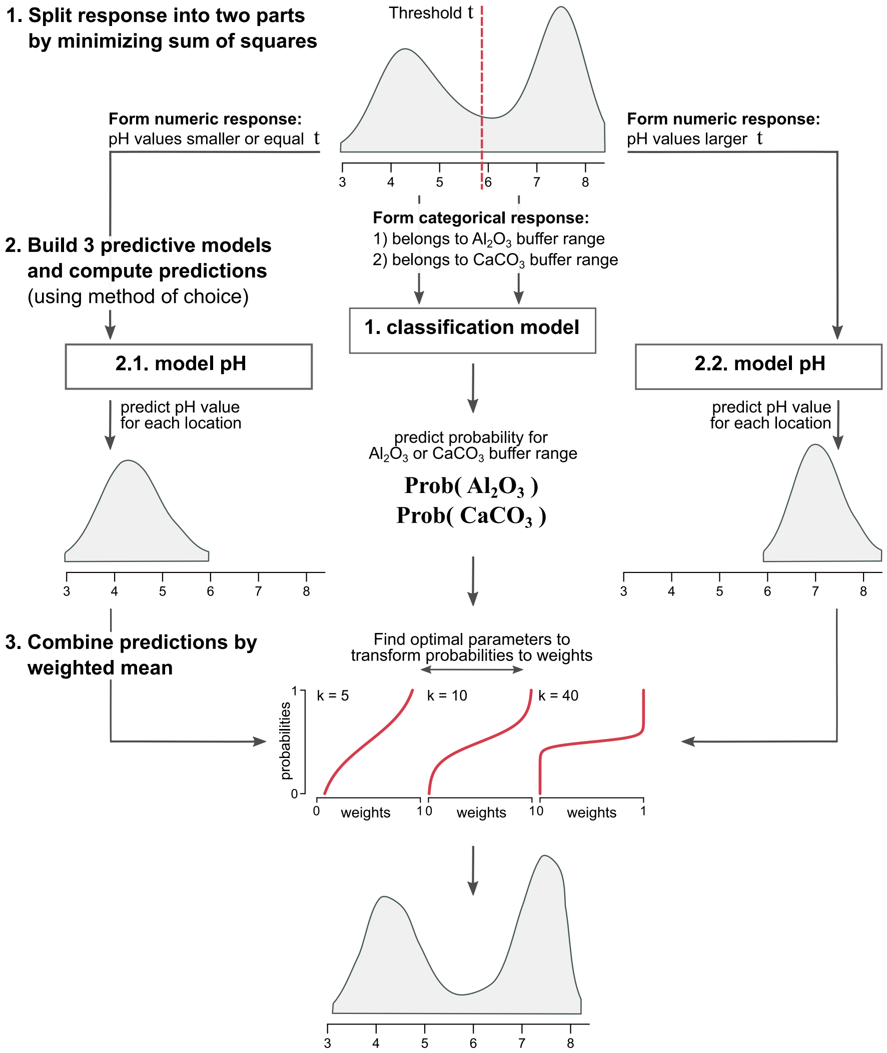


Fig. 3. Flowchart of hierarchical three step approach to model bimodal distributed responses by splitting response data, building three predictive models and then combining predictions by weighted means (k: tuning parameter of sigmoid function controlling smoothness to combine predictions of model 2.1 and 2.2. Parameter j_0 allows to shift function inflection in vertical direction which is not illustrated in the figure).

We demonstrate

1. how to compute hierarchical predictions considering underlying soil pH buffering processes and at the same time deriving relevant model building decisions from the present dataset;
2. how to quantify and validate uncertainty of such hierarchical predictions;
3. how to include degree of reproduction of observed response distribution into model building and validation by using a *goodness-of-fit* statistic to complement commonly used accuracy measures such as R^2.

First, we formalize the hierarchical model building approach (Section 2). Thereafter, we apply the presented framework to a medium and strong bimodal distributed dataset from pedogenic highly variable forest soils of Switzerland (Section 3) and present and discuss the results based on these example datasets (Sections 4 and 5).

2. Hierarchical prediction of bimodal distributed responses

Predictions for bimodal distributed responses may be modelled in a hierarchical two step approach (Fig. 3). Calibration data is split at the local minimum between peaks in its density function ideally resulting in nearly Gaussian distributed subsets. Models are built for each subset separately. To assign predictions from each model within the study region a binary classification is used. Response $Y(\mathbf{s})$ at location $\mathbf{s}$ needs to be split and predictions reunited in an optimal way as follows:

1. Divide the calibration data $Y(\mathbf{s})$ into two groups by a threshold t that minimizes variance within and maximizes variance inbetween the groups. t is estimated by minimizing the sum of squares

$$\tilde{t} = \arg\min_{t} \left\{ \sum_{i\in\{Y(\mathbf{s})|Y(\mathbf{s}_i)\le t\}} (Y(\mathbf{s}_i) - \bar{Y}_1)^2 + \sum_{i\in\{Y(\mathbf{s})|Y(\mathbf{s}_i)>t\}} (Y(\mathbf{s}_i) - \bar{Y}_2)^2 \right\} \qquad (1)$$

with $\bar{Y}_1$ and $\bar{Y}_2$ being the mean of each group $\{Y(\mathbf{s})|Y(\mathbf{s}_i) \le t\}$ and $\{Y(\mathbf{s})|Y(\mathbf{s}_i) > t\}$, respectively. Splitting by optimal threshold $\tilde{t}$ allows to form three responses:

$$G(\mathbf{s}) = I(y \in \{Y(\mathbf{s})|Y(\mathbf{s}_i) \le \tilde{t}\}) \qquad (2)$$

$$Y_a(\mathbf{s}) = y \in \{Y(\mathbf{s})|Y(\mathbf{s}_i) \le \tilde{t}\} \qquad (3)$$

$$Y_c(\mathbf{s}) = y \in \{Y(\mathbf{s})|Y(\mathbf{s}_i) > \tilde{t}\} \qquad (4)$$

$G(\mathbf{s})$ represents a binary outcome for each location to belong either to *aluminium oxide* (1) or *carbonate* buffer range (0) while

$Y_a(\mathbf{s})$ and $Y_c(\mathbf{s})$ contain the observed values below and above threshold $\tilde{t}$, respectively.

2. Build predictive models with new responses $G(\mathbf{s})$, $Y_a(\mathbf{s})$ and $Y_c(\mathbf{s})$ including tuning of model parameters or selection of relevant covariates as e.g.

$$Y_a(\mathbf{s}) = \hat{f}(X(\mathbf{s})) + \epsilon \tag{5}$$

with environmental covariates $X(\mathbf{s})$ at locations $\mathbf{s}$. From the classification model fitted to the binary response $G(\mathbf{s})$ probabilities $\tilde{J}(\mathbf{s}_+) = \widehat{Prob}(Y(\mathbf{s}_+) = 1|X(\mathbf{s}_+))$ are predicted for all $\mathbf{s}_+$ location in the study region. For the two numeric responses standard predictions $\tilde{Y}_a(\mathbf{s}_+)$ and $\tilde{Y}_c(\mathbf{s}_+)$ are computed.

3. Predictions $\tilde{Y}_a(\mathbf{s}_+)$ and $\tilde{Y}_c(\mathbf{s}_+)$ are combined by weights

$$\tilde{Y}(\mathbf{s}_+) = \tilde{W}(\mathbf{s}_+)\,\tilde{Y}_a(\mathbf{s}_+) + (1 - \tilde{W}(\mathbf{s}_+))\,\tilde{Y}_c(\mathbf{s}_+). \tag{6}$$

Weights vector $\tilde{W}(\mathbf{s}_+)$ is determined by transforming predicted probabilities $\tilde{J}(\mathbf{s}_+)$ for each prediction location $\mathbf{s}_+$ by sigmoid function

$$\tilde{W}(\mathbf{s}_+) = \frac{1}{1 + e^{-k(\tilde{J}(\mathbf{s}_+) - j_0)}} \tag{7}$$

where k determines the width of the function and hence the strengths of transformation of $\tilde{J}(\mathbf{s}_+)$. If $k = 0$ equal weights of 0.5 are used for all locations $\mathbf{s}_+$. For very large k probabilities $\tilde{J}(\mathbf{s}_+)$ become close to 0 for $\tilde{J}(\mathbf{s}_+) \leq j_0$ and close to 1 for $\tilde{J}(\mathbf{s}_+) > j_0$ resulting in either $\tilde{Y}_a(\mathbf{s})$ or $\tilde{Y}_c(\mathbf{s})$ predictions with potential crisp transitions in-between. j_0 refers to the inflection point of the sigmoid function and allows to optimize the probability threshold which is used to assign a location to either Al_2O_3 or $CaCO_3$ buffer range.

Optimal values for k and j_0 are estimated by a grid search from the calibration data by minimizing prediction variance measured as mean squared error skill score SS_{mse} and distributional deviation of observed $Y(\mathbf{s})$ and predicted $\tilde{Y}(\mathbf{s})$ measured by Kolmogorov–Smirnov test statistic D (Kolmogorov, 1933; Smirnov, 1939)

$$\{\tilde{k}, \tilde{j}_0\} = \underset{\{k, j_0\}}{\arg\min}\left\{SS_{\mathrm{mse}} + \mathrm{D}\right\}. \tag{8}$$

SS_{mse} (Wilks, 2011; Nussbaum et al., 2017) is defined as

$$SS_{\mathrm{mse}} = R^2 = MEC = 1 - \frac{\sum_{i=1}^{n}\left(Y(\mathbf{s}_i) - \tilde{Y}(\mathbf{s}_i)\right)^2}{\sum_{i=1}^{n}\left(Y(\mathbf{s}_i) - \frac{1}{n}\sum_{i=1}^{n} Y(\mathbf{s}_i)\right)^2}, \tag{9}$$

and also called R^2 or model efficiency coefficient (MEC, e.g. Helfenstein et al., 2022), with $SS_{\mathrm{mse}} = 1$ for perfect predictions, $SS_{\mathrm{mse}} = 0$ if predictions have the same variance as the observed data and $SS_{\mathrm{mse}} < 0$ for predictions with larger variance.

D is defined as distance between empirical cumulative distribution functions (ECDF) of $Y(\mathbf{s})$ and $\tilde{Y}(\mathbf{s})$ and takes the largest absolute difference between the two distribution functions across all X of the ECDF. D ranges from 0 to 1.

Optimal response splitting and merging of predictions is estimated from the data. The approach is model agnostic, i.e. it can be applied to any predictive method given the method allows to model continuous responses and binary responses with probability predictions.

Predictive distribution $\tilde{F}_+(Y(\mathbf{s}))$ for each new location $\mathbf{s}_+$ may be formed by combining the predictive distributions $\tilde{F}_+(Y_a(\mathbf{s}))$ and $\tilde{F}_+(Y_c(\mathbf{s}))$ of the two numerical models. Random sampling is done from pooled distribution with inclusion probabilities

$$Prob\left(\tilde{F}_+(Y_a(\mathbf{s})) \in \tilde{F}_+(Y(\mathbf{s}))\right) = \tilde{W}(\mathbf{s}) \tag{10}$$

$$Prob\left(\tilde{F}_+(Y_c(\mathbf{s})) \in \tilde{F}_+(Y(\mathbf{s}))\right) = 1 - \tilde{W}(\mathbf{s}) \tag{11}$$

according to weights $\tilde{W}(\mathbf{s})$ of Eq. (7). Full distribution of prediction errors from the two numerical models are hence merged based on the same weights used to combine mean predictions. Two-sided prediction intervals for α, for example, are then derived from resulting combined distributions by selecting $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles as

$$\left[\tilde{F}_+(Y(\mathbf{s}))_{\frac{\alpha}{2}};\ \tilde{F}_+(Y(\mathbf{s}))_{1-\frac{\alpha}{2}}\right]. \tag{12}$$

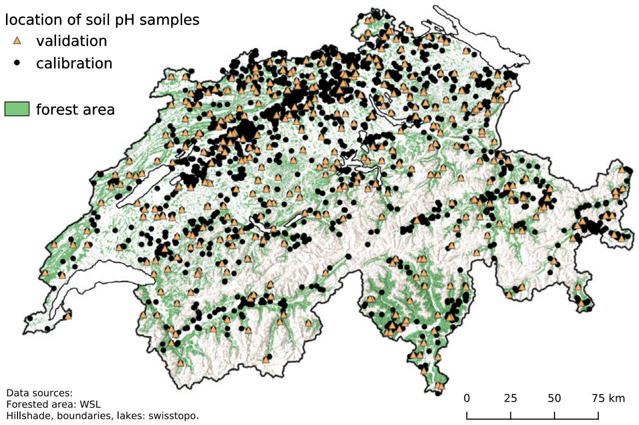


Fig. 4. Forested area of Switzerland with locations used for model calibration (black dots, n = 2530) and validation (orange triangles, n = 354) for pH in 0–5 cm soil depth.

3. Case study – materials and methods

3.1. Study region

Our study focused on forest soils of Switzerland (Nussbaum et al., 2014; Baltensweiler et al., 2021). Delineation of forests was done as by the definition of the national forest inventory (NFI, version 11.2017). With an area of 13 200 km^2 forests cover roughly 32% of Switzerland (Brändli et al., 2020).

Due to the high variability of topography, climate, and geology, Switzerland has strong environmental gradients compared to its small area. Soils formed on very diverse lithology, altogether yielding high spatial variability of soil properties such as soil pH. Low soil pH values prevail mostly in topsoils and on silicate parent materials whereas high pH values are found predominantly in deeper soil layers on parent material consisting of calcareous rocks or containing calcareous fractions (Walthert et al., 2010).

3.2. Data

3.2.1. Observed soil pH data

We assembled data from several sources:

- WSL soil data base (1833 sites where 67 profiles were included from data sources of Cantons),
- Swiss soil dataset (1051 locations, Service center NABODAT, 2018).

There was a total of 2884 soil profile locations with pH measurements for soil depth of 0–5 cm available (Fig. 4) of which 2682 also had pH measurements in 60–100 cm depth. Soil samples were taken from genetic horizons, dried and sieved to 2 mm. Soil pH was measured potentiometrically in 0.01 M $CaCl_2$ with a solid-extractant ratio of 1:2 (Thomas, 1996). Soil pH values were transferred to two fixed depth intervals by weighted mean where weights were derived according to thickness of each horizon fully or partly included within the depth interval (Baltensweiler et al., 2021). pH in 0–5 cm and pH in 60–100 cm were selected because the former had weak and the latter strong bimodal distribution (see red shaded density plots in bottom row of Fig. 5). Observed pH ranged from 2.6 to 8.0 in 0–5 cm and from 3.0 to 8.6 in 60–100 cm.

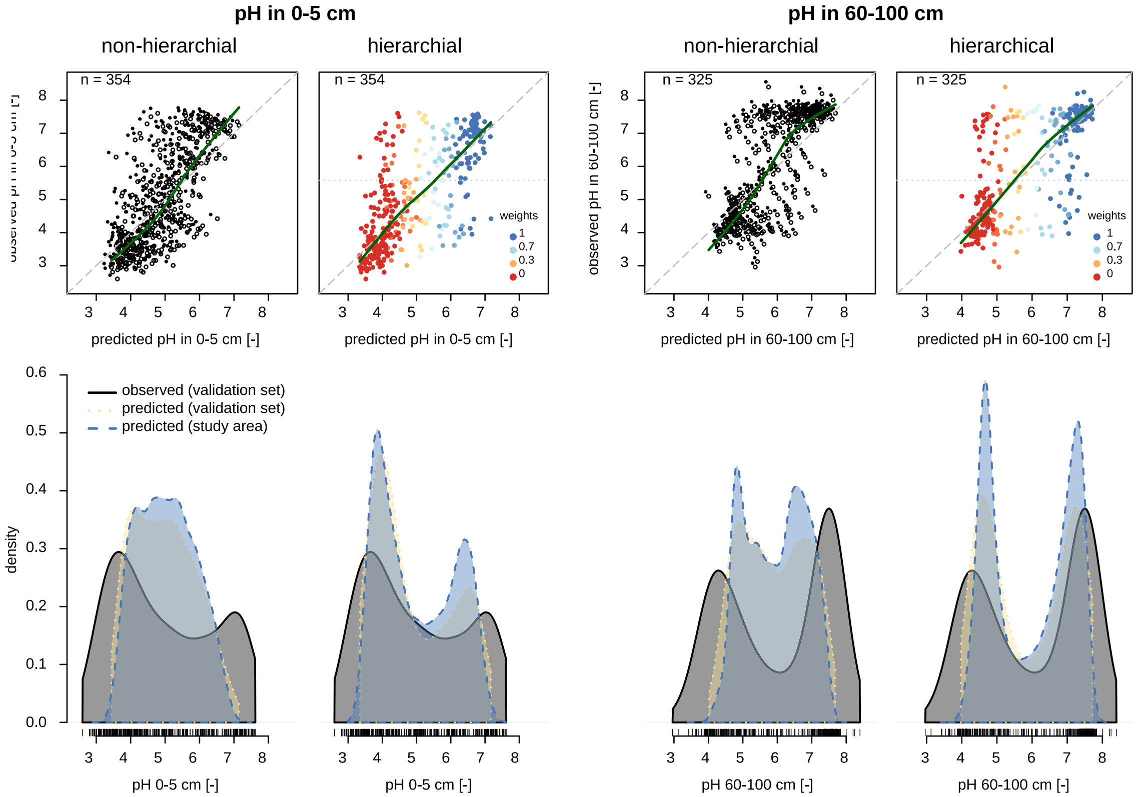


Fig. 5. Scatterplots of predicted vs. observed pH values computed for the validation data in 0–5 cm and 60–100 cm soil depth each for the standard non-hierarchical and the hierarchical modelling approach (top row). Colours in scatterplots for hierarchical predictions indicate weights used to combine $\tilde{Y}_a(\mathbf{s})$ and $\tilde{Y}_c(\mathbf{s})$ predictions according to Eq. (6). Bottom row displays density plots of observed validation data (solid black line), predictions computed for the validation locations (dotted yellow line) and prediction computed for all map pixels for forest soils of Switzerland (dashed blue line). Density plots have been cropped at their respective minimum and maximum value and rugs (short black vertical lines) were plotted for observed values.

3.2.2. *Explanatory covariates*

To represent soil forming conditions, spatial datasets from 18 different data sources were available. Used covariates and data sources were detailed in Baltensweiler et al. (2021). We used climate maps with monthly resolution from different climatic periods (1961–1990, 1981–1990, 1981–2010, 1975–2010) modelled at different spatial resolution (25 m, 100 m, 250 m, 2 km, total 53 climate covariates).

Terrain attributes (TA) were derived from 5 and 25 m resolution elevation models and smoothed by 2D convolutional filters with a Gaussian weighing scheme to represent different spatial scales (radii 15–200 m). Calculation of TA was done with following main computational functions: convexity, curvature, flow accumulation, flow length, flow path, slope, specific catchment area, ruggedness, stream power, topographic position index, topographic wetness index and valley depth (total of 86 terrain covariates).

Parent material and partitioning of the landscape was broadly represented by several overview maps: last Glacial Maximum (map scale 1:500 000), Hydrogeological Map (1:500 000), Geotechnical Map (1:200 000), a landscape classification (1:200 000) and the large-scaled Swiss Soil Suitability Map (1:200 000, with attributes like soil depth or nutrient storage capacity). To latter, median pH per map unit was assigned from a poor-quality pH dataset of 10 865 topsoil observations. In addition to Baltensweiler et al. (2021) a map representing the distance to the nearest rock outcrop (Swisstopo, 2022) was used (total of 93 covariates).

Vegetation was represented by satellite image derivatives from Sentinel-2 (10 m resolution) and Landsat time series by averaging over years 1985–2015 (30 m resolution). Different indices like green and default normalized difference vegetation indices or pigment specific simple ratio were computed. Besides, a canopy height model (25 m resolution), biogeographic regions (1:200 000) and the proportion of coniferous trees was used (total of 37 vegetation covariates).

Categorical polygon based maps with crisp boundaries (e.g. soil and geology maps, biogeographic regions) were transferred to their indicator representation (1: inside unit, 0: outside unit, for n categories resulting in $n-1$ new raster layers). To account for the uncertainty of unit boundaries we computed the mean within a moving rectangular window whose size was determined by the scale of the original map product to correspond to 2–4 mm if the map were printed (i.e. 800 m for map scale of 1:200 000).

To represent relative position, we additionally added oblique coordinate axis with rotations of 30° and 60° (Møller et al., 2020) resulting in 274 covariates in total. Covariates were all resampled to a pixel width of 5 m by bilinear interpolation regardless the original resolution.

3.3. *Statistical analysis*

Steps 1 to 3 detailed in Section 2 were applied to pH values in 0–5 and 60–100 cm soil depth separately (2.5D approach, Ma et al., 2021, 3.2.1). To model the three constructed responses (step 1) for each variable to be mapped we used random forest (RF, Breiman, 2001). RF performed best on a smaller data set for the same study area (Baltensweiler et al., 2021).

RF algorithm establishes a large number of fully grown classification or regression trees (CART). Individual trees are decorrelated by two resampling procedures: (1) for each tree only a random subset of observations is used and (2) for each node only $\mathrm{m}_{\mathrm{try}} > p$ randomly selected covariates are tested as candidates for binary splitting. Final predictions are means of predictions produced by all n_{tree} trees.

We first reduced the large number of covariates by sequential recursive backward elimination (Brungard et al., 2015) based on node-impurity covariate importance (Hastie et al., 2009, Sect. 15.3.2). We removed 5 to 10 covariates at each step fitting models with $5, 10, 15, \ldots,$ $100,\ 110, \ldots, 170,\ p$ covariates. Covariates were further decorrelated by limiting degree of correlation ϵ (Hertzog, 2017; Baltensweiler et al., 2021). For covariate removal we used default $\mathrm{m_{try}} = \frac{p}{3}$. To find optimal $\mathrm{m_{try}}$ for the final covariate set we minimized out-of-bag RMSE by iterating through $\mathrm{m_{try}} = 1, 2, \ldots, p$. Number of trees $\mathrm{n_{tree}}$ and minimal number of observations remaining in each tree end node n_{min} were left on default values ($\mathrm{n_{tree}}$: 500, $\mathrm{n_{min}}$: 5 for regression and 10 for classification to predict probabilities). Model selection was done minimizing out-of-bag Gini index for classification (binary response according Eq. (2) in Section 2) and mean squared error for regression (responses according Eqs. (3) and (4) in Section 2).

Prediction intervals were computed by quantile regression forest (QRF, Meinshausen, 2006).

3.4. Model validation

Soil data was split into a calibration (88%) and a validation set (12%) used only to compute the model performance statistics. We used locations with measured pH of the same validation dataset as in previous analysis (Hertzog, 2017; Baltensweiler et al., 2021, sec 2.3). Validation locations were chosen by a stratified weighted random selection. Inclusion probability was proportional to the area of physiographic units (strata) of the Swiss soil suitability map (SSSM, FSO, 2000). To avoid overrepresentation of validation points in spatial clusters with a large profile density, validation points were sampled (without replacement) with probability weights corresponding to the forest area in the Voronoi polygons. The selected validation data therefore represents a broad approximation for a true design-based validation sample (Brus et al., 2011).

To evaluate overall predictive model performance we computed bias, RMSE, $\mathrm{SS_{mse}}$ (Eq. (9)) and Lin's concordance correlation coefficient (Lin, 1989) for all continuous predictions and bias ratio, percent correct and pierce skill score (PSS, Wilks, 2011) for class predictions (Eq. (2)).

All analysis was done in R (R Core Team, 2022) using packages *ranger* (Wright and Ziegler, 2017) to fit RF, *sf* (Pebesma, 2018) and *raster* (Hijmans, 2022) for spatial data analysis and management, *twosamples* (Dowd, 2022) to compute D statistic, *DescTools* for CCC (Signorell, 2023) and *verification* (NCAR, 2015) for PSS.

4. Results

4.1. Models for pH in 0–5 cm and 60–100 cm soil depth

Table 1 reports model parameters selected to compute final predictions. The optimal threshold for data splitting into Al_2O_3 and $CaCO_3$ buffer ranges (resulting in $Y_a(\mathbf{s})$ and $Y_c(\mathbf{s})$) was slightly larger for pH response in 60–100 cm (pH 5.25 vs. pH 5.86). For both responses, width of sigmoid function was narrow (large $\tilde{k}$) leading to large weights for $\tilde{Y}_a(\mathbf{s})$ and small weights for $\tilde{Y}_c(\mathbf{s})$ predictions and vice-versa. Hence, buffer ranges are clearly assigned in the final result with only little smoothing in between the $\tilde{Y}_a(\mathbf{s})$ and $\tilde{Y}_c(\mathbf{s})$ prediction maps.

Calibration data was imbalanced especially for pH in 0–5 cm (35.2% of observations in $CaCO_3$ vs. 64.8% in Al_2O_3 buffer range) and less for pH in 60–100 cm (41.5% in $CaCO_3$ and 58.5% in Al_2O_3 buffer range). A strategy to consider imbalanced data is to estimate optimal thresholds to convert probability predictions into classes. Thresholds t are chosen with calibration data by iterating e.g. through $0, 0.01, 0.02, .., 0.99, 1$ and minimizing a categorical measure like PSS. The chosen threshold t is then applied to transfer probabilities to classes for the validation set and predictions for new locations. For imbalanced data thresholds often clearly deviate from midpoint of 0.5 (e.g. Nussbaum et al., 2017, Table 4). In the presented approach thresholds are constituted by the inflection point of the sigmoid function ($\tilde{j}_0$ in Table 1). Although calibration data was imbalanced optimal selection of inflection points were close to 0.5 and did not exhibit the need to correct for the imbalanced setting.

If measured by RMSE and $\mathrm{SS_{mse}}$ out-of-bag validation resulted in slightly inferior model performance for hierarchical models (Table 1) while CCC showed opposite conclusion. Comparing the resulting distributions by D, however, indicated a closer fit for both hierarchical models compared to their non-hierarchical counterparts.

4.2. Validation of predicted pH value with independent data

Class predictions to assign each location of the independent validation data set to either Al_2O_3 or $CaCO_3$ buffer range resulted in no bias (bias ratio close to one). With about 80% percent correct and PSS of 0.55 and 0.61 for pH in 0–5 and 60–100 cm, respectively, classification resulted in satisfactory predictions (Table 2).

None of the final pH models showed a bias (Table 3). The final predictions combined from three hierarchical models according to Eq. (6) resulted in somewhat larger RMSE and lower $\mathrm{SS_{mse}}$ compared to non-hierarchical predictions with CCC again showing an opposite trend. $\mathrm{SS_{mse}}$ dropped by 2.1 percent points for pH in 0–5 cm and by 2.6 for pH in 60–100 cm.

Distribution of errors was quite different for non-hierarchical and hierarchical predictions. Overall absolute errors were clearly smaller for hierarchical predictions with their median being 0.56 for pH in 0–5 cm and 0.42 for pH in 60–100 cm. Non-hierarchical models had medians of absolute errors of 0.68 and 0.64, respectively. Evident from validation scatterplots (top row in Fig. 5) some locations were wrongly assigned by the first step of the hierarchical modelling. Therefore, there was a slight increase in very large differences between predicted and observed: non-hierarchical models resulted in 17 and 21 locations with errors larger than 2 pH units for 0–5 and 60–100 cm while hierarchical models had 30 and 37 such large deviations, respectively.

Distributions of hierarchical predictions themselves were closer to the observed validation data resulting in smaller deviation as measured by D (7.4 percent points for topsoil and 9.3 for pH in 60–100 cm). Bimodal distributions were reproduced much closer if density plots are compared (Fig. 5, bottom row). Weak bimodal distribution as in pH 0–5 cm resulted in nearly unimodal predictions from a conventional non-hierarchical approach. For stronger bimodal distribution as observed in 60–100 cm, non-hierarchical predictions resulted in weak bimodal predictions, but a large number of locations were still predicted within a pH range that is rarely observed. Overall, hierarchical models displayed pH prediction frequencies in pedologically meaningful ranges, although peaks of bimodal distributions are slightly shifted towards the median. In addition, upper and especially lower tails were poorly predicted for any of the approaches.

4.3. Validation of prediction intervals

Prediction intervals were evaluated by comparing their nominal probabilities to their coverage of observed values in the validation data set (Fig. 6a–d). Non-hierarchical models predicted intervals that were too wide, i.e. too pessimistic, compared to the observed data. Coverage was even more pessimistic for the more pronounced bimodal distributed pH in 60–100 cm soil depth. For a 90%-prediction interval this results in only 4.3% of the observations being outside the intervals instead of the expected 10% (Fig. 6, panel e).

Interval coverage for hierarchical models showed a nearly exact fit for medium bimodal distributed pH in 0–5 cm and became slightly too pessimistic for pH in 60–100 cm with a coverage of 8.6% for the 90%-prediction intervals. Fig. 6 (panels e and f) show clear reduction of uncertainty for pH predicted smaller than 4.5 and a slight reduction for

Table 1
Optimal model parameters and validation statistics computed with out-of-bag predictions of the data used for model calibration (n: number of calibration locations with observed pH for given soil depth; $\tilde{t}$: threshold to split data set before model fit; $\tilde{k}$: smoothness to merge predictions/width of sigmoid function; $\tilde{j}_0$: binarisation threshold/inflection of sigmoid function; p: number of covariates in the final model; m_{try}: number of randomly selected covariates to test at each split, given for 3 hierarchical models in order: binary classification, Al_2O_3 buffer and $CaCO_3$ buffer range; bias: negative mean error; RMSE: root mean squared error; SS_{mse}: mean squared error skill score aka. R^2; CCC: concordance correlation coefficient; D: Kolmogorov–Smirnov test statistic).

		$\tilde{t}$	$\tilde{k}$	$\tilde{j}_0$	p	m_{try}	bias	RMSE	SS_{mse}	CCC	D
pH 0–5 cm (n = 2530)	Non-hierar.				50	14	0.132	0.957	0.559	0.692	0.228
	Hierarchical	5.25	14.2	0.52	90/15/12	9/4/7	−0.082	0.970	0.547	0.726	0.157
pH 60–100 cm (n = 2357)	Non-hierar.				39	11	0.001	1.029	0.550	0.690	0.244
	Hierarchical	5.86	12.6	0.48	20/29/20	5/5/2	−0.085	1.057	0.525	0.719	0.191

Table 2
Confusion matrix and validation statistics for binary classification model to divide between Al_2O_3 and $CaCO_3$ buffer ranges computed for the independent validation dataset. To convert probability predictions optimal thresholds $\tilde{j}_0$ were used (pH 0–5 cm: 0.52, pH 60–100 cm: 0.48, see Table 1, n: number of validation locations with observed pH for given soil depth; bias: bias ratio, with 1 = no bias, <1 = presence of Al_2O_3 buffer range is underpredicted, >1 = Al_2O_3 range is overpredicted; PC: percentage correct; PSS: Pierce Skill Score with −1 = opposite prediction, 0 = random prediction, 1 = perfect prediction).

	Predicted buffer range	Observed buffer range		Bias	PC [%]	PSS
		Al_2O_3	$CaCO_3$			
pH 0–5 cm (n = 354)	Al_2O_3	183	43	1.06	79.4	0.554
	$CaCO_3$	30	98			
pH 60–100 cm (n = 325)	Al_2O_3	126	37	1.07	80.6	0.615
	$CaCO_3$	26	136			

Table 3
Validation statistics computed on predictions for the validation dataset not used in any step of model building (n: number of validation locations with observed pH for given soil depth; bias: negative mean error; RMSE: root mean squared error; SS_{mse}: mean squared error skill score aka. R^2; CCC: concordance correlation coefficient; D: Kolmogorov–Smirnov test statistic).

		Bias	RMSE	SS_{mse}	CCC	D
pH 0–5 cm (n = 354)	Non-hierarchical	0.003	1.017	0.514	0.647	0.232
	Hierarchical	0.090	1.038	0.493	0.688	0.158
pH 60–100 cm (n = 325)	Non-hierarchical	0.097	1.046	0.529	0.663	0.311
	Hierarchical	−0.094	1.075	0.503	0.701	0.218

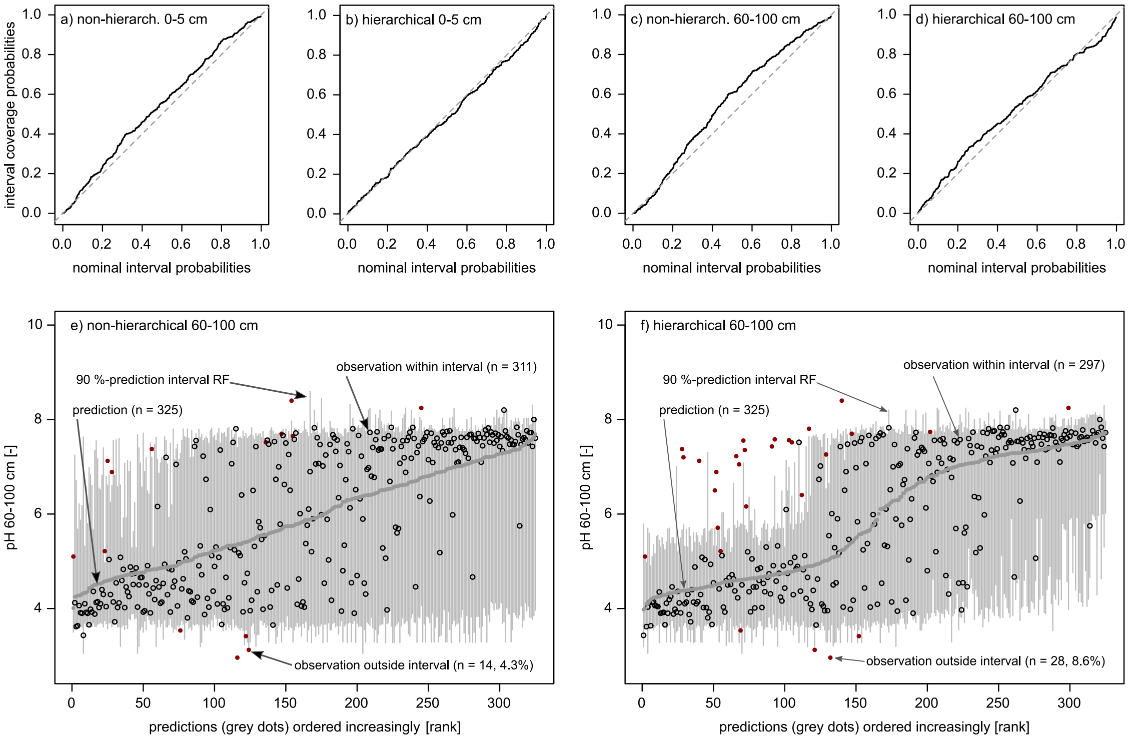
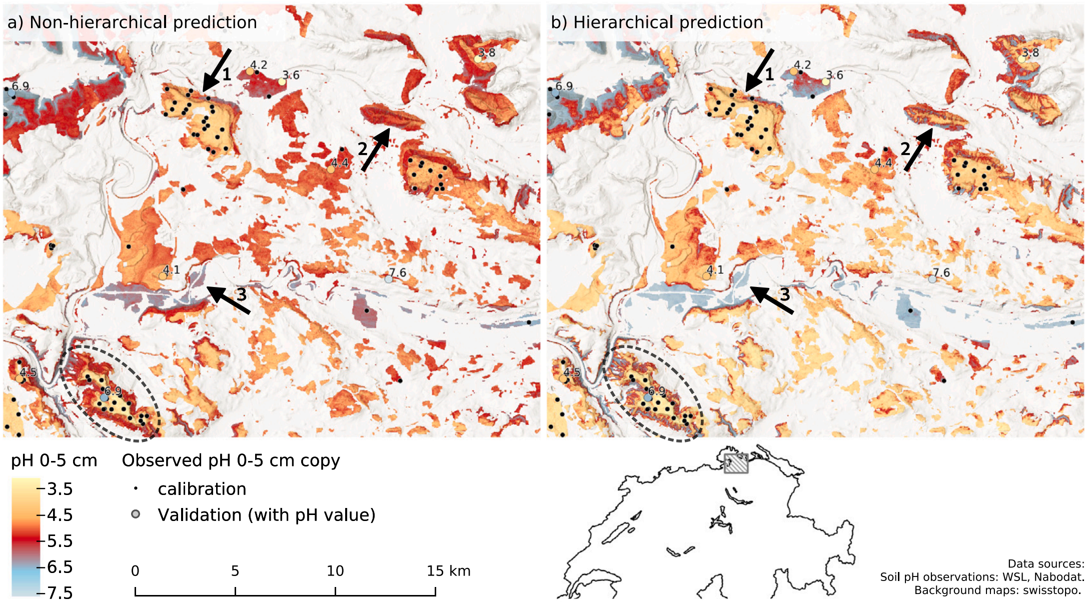


Fig. 6. Top row: Nominal probabilities plotted against the actual coverage of two-sided QRF prediction intervals compared to the independent validation samples (0–5 cm: n = 354, 60–100 cm: n = 325; dashed 1:1-line: ideal interval width where nominal equals actual coverage, solid line above 1:1-line: coverage too large/pessimistic, solid line below 1:1-line: coverage too small/optimistic). Bottom row: Ordered predictions (grey circles) of pH in 60–100 cm along with 90%-prediction intervals (vertical grey lines). Observed pH inside the intervals are plotted by open circles, those outside by red filled symbols.

Table 4
Rating of uncertainties for pH of the validation data set according to GlobalSoilMap Tiers specification (Arrouays et al., 2014a, #: number of observations, %: percentage of observations of validation data set, rating based on width of 90%-prediction interval smaller than or equal the value give in brackets). .

		AAA (≤1)		AA (≤2)		A (≤3)		None (>3)	
		#	%	#	%	#	%	#	%
pH 0–5 cm (n = 354)	Non-hierarchical	0	0	6	1.7	35	9.9	313	88.4
	Hierarchical	0	0	56	15.8	119	33.6	179	50.6
pH 60–100 (n = 325)	Non-hierarchical	0	0	9	2.8	19	5.8	297	91.4
	Hierarchical	0	0	54	17.0	82	25.0	189	58.0



Data sources:
Soil pH observations: WSL, Nabodat.
Background maps: swisstopo.

Fig. 7. Map section of standard non-hierarchical (a) and hierarchical predictions (b) for soil pH in 0–5 cm with colour scale emphasizing on pH values around 5.5, i.e. between Al_2O_3 and $CaCO_3$ buffer ranges. The section was chosen to illustrate an area with spatially clustered calibration points and larger surfaces away from calibration sites predicted in intermediate pH range in (a). Validation locations are plotted with grey circles and their observed pH values is given. Arrows 1 to 3 and dashed ellipse are areas referred to in the text.

pH predicted above 7.5. Within the intermediate pH range uncertainty remains large also with the two-step approach proposed in the present article.

The more accurate uncertainty prediction results in a substantial improvement regarding the so called *Tiers* according to Arrouays et al. (2014a, thresholds published in Helfenstein et al., 2022, Table A1). With standard non-hierarchical predictions the majority (88 and 91%) of validation observations falls outside of the rating (Table 4). With the new approach this ratio can be reduced to 51% for pH in 0–5 cm and 58% for pH in 60–100 cm. AA rating is achieved for 16-17% percent as compared to 2-3% for the standard predictions.

4.4. Mapping of soil pH

As displayed in the density plots (Fig. 5, bottom row) pH in intermediate range (pH 4.5–6) was predicted for much fewer pixels using the two-step hierarchical approach compared to a standard non-hierarchical approach. Fig. 7 shows a detail map section emphasizing on the differences in the intermediate pH range. Non-hierarchical predictions (panel a) performed well for areas with locations with calibration data (arrow 1). In areas with few or no calibration locations (arrow 2) there is a tendency towards an intermediate pH (pH 4.5–6). The hierarchical predictions (panel b, arrow 2) were able to predict small and large pH also further away from calibration data locations.

5. Discussion

5.1. Accuracy of predictions

Average prediction accuracy measured by SS_{mse} was below the median of 0.6 reported by Chen et al. (2022, Fig. 9) for pH mapping of large areas comparable to this study, but within the full range of below 0.1 to above 0.8. The difference between hierarchical and non-hierarchical approach was only small. RMSE became dominated by large errors due to the misclassification of some observations in the binary model. A similar outcome was observed by Rawlins et al. (2009). Analogously to the binary model in the present study they used units of an existing soil map to subsequently fit two models. A substantial number of observations were wrongly assigned by the soil map and resulted in large absolute errors and thus large RMSE (Rawlins et al., 2009).

We are aware that conclusions from comparing density plots as done in Fig. 5 are possibly biased. Validation data was not represented by a design-based sample (Brus et al., 2011), but data splitting was done to best approximate unbiased conclusions. We do not know of other studies that benchmarked *goodness-of-fit* of observed and predicted distribution in digital soil mapping. For Mulder et al. (2016) and Helfenstein et al. (2022) it can be implied from their figures that observed bimodal distributions were not reproduced by the predictions. A statistic is however not provided. The Kolmogorov–Smirnov test statistic D used here was well able to detect deviations located at the centre of the distribution. It might not be a suitable statistic for all responses and foci of end-users. Especially to evaluate performance of predictions towards the tails of a distribution, other *goodness-of-fit* statistics like the Anderson–Darling two sample test (Scholz and Stephens, 1987) are more reliable. While Anderson–Darling test statistic is more sensitive to the tails, it lacks the convenient property of ranging between 0 and 1.

Figs. 2 and 5 (bottom row) reveal that tails indeed were under- or over-predicted by all approaches. This seems to be a general problem of currently used predicting methods (Mulder et al., 2016; Wang et al., 2019; Roudier et al., 2020). Smoothing behaviour around the lower and upper margins of response distribution is typically pronounced with RF when relationships between covariates and responses are rather weak (Baltensweiler et al., 2021). RF predictions are computed by averaging twice (mean of observations in terminal tree nodes, then mean over all trees, Hastie et al., 2009) and are thus applying stronger smoothing at the tails than other machine learning approaches (Fig. 2).

5.2. *Performance of uncertainty quantification*

A bimodal or multimodal distribution likely has a larger standard deviation which then is propagated into predictive distributions obtained by non-parametric bootstrap (e.g. QRF). Helfenstein et al. (2022) and our study indeed observed too large prediction intervals by not considering bimodality. By using the presented hierarchical approach the problem could be removed or at least reduced.

Other studies often reported average coverage of 90% prediction intervals which were accurate (Poggio et al., 2021), too pessimistic/large (coverage of 0.94–0.98, Padarian et al., 2017) or too optimistic/small (0.76–0.86, Viscarra Rossel et al., 2015; 0.88–0.9, Mulder et al., 2016). Not all publications used the same approach to quantify uncertainty, hence it remains difficult to disentangle the effect of bimodal response distribution on the reported interval coverage.

5.3. *Evaluation of mapped pH patterns*

Please note that only few validation samples were available and the true soil pH remains unknown throughout the largest part of the map. Moreover, small scale variability is considerably large and given the observation density presented mapping remains on an overview scale.

The spatial pattern of the hierarchical predictions, however, followed closer of what would be expected from a pedogenic point of view. Large pH values in the lower plains as for example along river Thur (Fig. 7, arrow 3) are feasible due to weakly weathered recent alluvial deposits containing carbonates. Hilltops and plateaus in the displayed map section are often strongly weathered with hardly any erosion. Hence, small pH values are expected and spatial patterns at arrow 2 in Fig. 7 appear to be more reliable predicted with the hierarchical approach (panel b). On steep slopes falling from these plateaus topsoils have often been eroded and show intermediate pH or even alkaline conditions on very steep sites. Again, a more differentiated pattern was predicted by the hierarchical approach while the non-hierarchical predictions were likely not able to identify strongly eroded areas with large pH above 6. The above mentioned pH patterns predicted more accurately by the hierarchical approach were confirmed for Irchel hill marked by an ellipse in Fig. 7 (written communication, 01.2023, construction directorate of Canton of Zurich, detailed survey from 2018, data not public).

5.4. *Applicability of approach*

The present study used RF as prediction method, but the developed hierarchical approach is not limited to RF. Any other suitable prediction method could be implemented if numeric responses can be modelled and probabilities can be predicted for binary responses. Moreover, the full predictive distribution needs to be established to combine uncertainties. If a method does not provide an inherent bootstrap procedure like QRF, a model-based or non-parametric bootstrap may be used (Davison and Hinkley, 1997). In addition, all tuning parameters are estimated from the data, limiting arbitrary choices.

A hierarchical two step approach is a more complex procedure with larger implementation effort. Three models have to be fitted and three maps have do be computed for the whole area. Further, three additional tuning parameters need to be estimated (optimal data split t, k and j_0 of sigmoid function). Moreover, it makes a bimodal distributed response a special case. As soil function assessments are based on numerous soil properties (Greiner et al., 2018) pedometricians often map more than one soil property for multiple depth (Chen et al., 2022). If each response needs its own consideration and specific modelling approach, a substantial increase in workload is imposed. Bayes approaches allowing to insert pedologically informed priors into the model would likely resolve having multiple models to map one response.

6. Summary and conclusion

We presented a multi-step procedure to reproduce bimodal data distributions in spatial predictions. Such data is often found for soil pH and sometimes for soil organic carbon or properties relating to pore size distribution. For each data culmination of the bimodal response distribution a separate model is fitted. Another model is needed to assign each new prediction location to either of the two modes. The statistical approach for the three models required to be fitted may be chosen by the data scientist. Some prediction errors might increase, but overall the additional effort is rewarded by predictions being more similar to the distribution of the originally observed values. In addition, narrower and more accurate prediction intervals are to be expected.

The present article lets us conclude:

- Mean summary statistics (e.g bias, SS_{mse}, RMSE) are not be the sole measures to evaluate the quality of a predicted map. Map validation should be extended by comparing *goodness-of-fit* of observed and predicted distributions.
- Refinement of digital soil mapping should consider specific pedogenic characteristics of soil properties resulting in characteristic data distributions (such as bimodal distributions).
- Additional effort in data analysis and collaboration among statisticians and pedologists are needed to avoid misleading conclusions by end-users.

CRediT authorship contribution statement

Madlene Nussbaum: Designed the study, Developed the approach and analysed the forest soil data, Prepared the manuscript and handled the review process, Commented and improved on the draft of the article. **Stephan Zimmermann:** Assembled the soil data from WSL and NABODAT databases, Specifically contributed pedological background information to introduction, results and discussion, Commented and improved on the draft of the article. **Lorenz Walthert:** Specifically contributed pedological background information to introduction, results and discussion, Commented and improved on the draft of the article. **Andri Baltensweiler:** Preprocessed all geodata and derived the spatial covariates, Commented and improved on the draft of the article.

Declaration of competing interest

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

Data availability

The authors do not have permission to share data.

Acknowledgements

We are thankful to Swiss Federal Office for the Environment (FOEN) for funding a previous study that eased data preparation for the current work. We are grateful to Sabine Augustin (FOEN) for providing impulses to further improve mapping of forest soils in Switzerland. We would like to thank members of NABODAT association for making soil pH data accessible for this research. Moreover, we like to thank the numerous collaborators of the Swiss Federal Institute for Forest, Snow and Landscape Research (WSL) who contributed to the ever growing soil data base over the past 40 years.

References

Adhikari, K., Bou Kheir, R., Greve, M.B., Greve, M.H., Malone, B.P., Minasny, B., McBratney, A.B., 2014. Mapping soil pH and bulk density at multiple soil depths in Denmark. In: GlobalSoilMap: Basis of the Global Spatial Soil Information System - Proceedings of the 1st GlobalSoilMap Conference. pp. 155–160.

Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G., Hong, S.Y., Lagacherie, P., Lelyk, G., McBratney, A.B., McKenzie, N.J., Mendonca-Santos, M.D., Minasny, B., Montanarella, L., Odeh, I., Sanchez, P.A., Thompson, J.A., Zhang, G.-L., 2014a. GlobalSoilMap. Toward a fine-resolution global grid of soil properties. Adv. Agron. 125, 93–134. http://dx.doi.org/10.1016/B978-0-12-800137-0.00003-0.

Arrouays, D., McBratney, A.B., Minasny, B., Hempel, J.W., Heuvelink, G.B.M., MacMillan, R.A., Hartemink, A.E., Lagacherie, P., McKenzie, N.J., 2014b. The GlobalSoilMap project specifications. In: GlobalSoilMap Basis of the Global Spatial Soil Information System. CRC Press, pp. 9–12. http://dx.doi.org/10.1201/b16500-4.

Baltensweiler, A., Heuvelink, G.B., Hanewinkel, M., Walthert, L., 2020. Microtopography shapes soil pH in flysch regions across Switzerland. Geoderma 380, 114663. http://dx.doi.org/10.1016/j.geoderma.2020.114663.

Baltensweiler, A., Walthert, L., Hanewinkel, M., Zimmermann, S., Nussbaum, M., 2021. Machine learning based soil maps for a wide range of soil properties for the forested area of Switzerland. Geoderma Reg. 27, e00437. http://dx.doi.org/10.1016/j.geodrs.2021.e00437.

Batjes, N.H., Ribeiro, E., van Oostrum, A., Van Oostrum, A., Mendes, J., Standardised soil profile data for the world: WoSIS Snapshot – September 2019, http://dx.doi.org/10.17027/isric-wdcsoils.20190901.

Bechler, K.H., Toth, O., 2010. Bewertung von Böden nach ihrer Leistungsfähigkeit, URL http://www.fachdokumente.lubw.baden-wuerttemberg.de/servlet/is/99474/Bodenschutz_23_Lesefassung_aktuell.pdf?command=downloadContent&filename=Bodenschutz_23_Lesefassung_aktuell.pdf&FIS=199.

Blume, H.-P., Brümmer, G.W., Fleige, H., Horn, R., Kandeler, E., Kögel-Knabner, I., Kretzschmar, R., Stahr, K., Wilke, B.-M., 2016. Scheffer/Schachtschabel Soil Science. Springer, Berlin and Heidelberg and New York and Dordrecht and London, http://dx.doi.org/10.1007/978-3-642-30942-7.

Bolan, N.S., Adriano, D.C., Curtin, D., 2003. Soil acidification and liming interactions with nutrient and heavy metal transformation and bioavailability. Adv. Agron. 78 (21), 5–272.

Brändli, U.-B., Abegg, M., Allgaier Leuch, B., 2020. Schweizerisches Landesforstinventar. Ergebnisse der vierten Erhebung 2009–2017.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

de Brogniez, D., Ballabio, C., Stevens, A., Jones, R.J.A., Montanarella, L., van Wesemael, B., 2015. A map of the topsoil organic carbon content of Europe generated by a generalized additive model. J. Soil Sci. 66 (1), 121–134. http://dx.doi.org/10.1111/ejss.12193.

Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., Edwards Jr., T.C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239–240, 68–83. http://dx.doi.org/10.1016/j.geoderma.2014.09.019.

Brus, D.J., Kempen, B., Heuvelink, G.B.M., 2011. Sampling for validation of digital soil maps. Eur. J. Soil Sci. 62 (3), 394–407. http://dx.doi.org/10.1111/j.1365-2389.2011.01364.x.

Chen, S., Arrouays, D., Leatitia Mulder, V., Poggio, L., Minasny, B., Roudier, P., Libohova, Z., Lagacherie, P., Shi, Z., Hannam, J., Meersmans, J., Richer-de Forges, A.C., Walter, C., 2022. Digital mapping of GlobalSoilMap soil properties at a broad scale: A review. Geoderma 409, 115567. http://dx.doi.org/10.1016/j.geoderma.2021.115567.

Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and their Applications. Cambridge University Press, Cambridge, http://dx.doi.org/10.1017/cbo9780511802843.

Dowd, C., 2022. Twosamples: Fast permutation based two sample tests: R package version 2.0.0. URL https://CRAN.R-project.org/package=twosamples.

FSO, 2000. Swiss soil suitability map. In: BFS GEOSTAT. Swiss Federal Statistical Office.

Gostkowski, M., Gajowniczek, K., 2020. Weighted quantile regression forests for bimodal distribution modeling: A loss given default case. Entropy (Basel, Switzerland) 22 (5), http://dx.doi.org/10.3390/e22050545.

Greiner, L., Keller, A., Grêt-Regamey, A., Papritz, A., 2017. Soil function assessment: review of methods for quantifying the contributions of soils to ecosystem services. Land Use Policy 69, 224–237. http://dx.doi.org/10.1016/j.landusepol.2017.06.025, URL http://www.sciencedirect.com/science/article/pii/S0264837717305719.

Greiner, L., Nussbaum, M., Papritz, A., Fraefel, M., Zimmermann, S., Schwab, P., Grêt-Regamey, A., Keller, A., 2018. Assessment of soil multi-functionality to support the sustainable use of soil resources on the swiss plateau. Geoderma Reg. 14, e00181. http://dx.doi.org/10.1016/j.geodrs.2018.e00181.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, second ed. Springer, http://dx.doi.org/10.1007/978-0-387-84858-7.

Helfenstein, A., Mulder, V.L., Heuvelink, G.B., Okx, J.P., 2022. Tier 4 maps of soil pH at 25 m resolution for the Netherlands. Geoderma 410, 115659. http://dx.doi.org/10.1016/j.geoderma.2021.115659.

Hertzog, M., 2017. Modelling Soil Attributes with the Random Forest Method for the Swiss Forest Area, (Master Thesis), ETH Zürich, Switzerland.

Hijmans, R.J., 2022. Raster: Geographic data analysis and modeling: R package version 3.5-15. URL https://CRAN.R-project.org/package=raster.

Kolmogorov, A., 1933. Sulla determinazione empirica di una lgge di distribuzione. Inst. Ital. Attuari, Giorn. 4, 83–91, URL https://ci.nii.ac.jp/naid/10010480527/.

Laslett, G.M., McBratney, A.B., Pahl, P.J., Hutchinson, M.F., 1987. Comparison of several spatial prediction methods for soil pH. J. Soil Sci. 38, 325–341.

Libohova, Z., Seybold, C., Adhikari, K., Wills, S., Beaudette, D., Peaslee, S., Lindbo, D., Owens, P.R., 2019. The anatomy of uncertainty for soil pH measurements and predictions: Implications for modellers and practitioners. Eur. J. Soil Sci. 70 (1), 185–199. http://dx.doi.org/10.1111/ejss.12770.

Lin, L.I.-K., 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45, 255–268.

Lu, Q., Tian, S., Wei, L., 2023. Digital mapping of soil pH and carbonates at the European scale using environmental variables and machine learning. Sci. Total Environ. 856 (Pt 2), 159171. http://dx.doi.org/10.1016/j.scitotenv.2022.159171.

Ma, Y., Minasny, B., McBratney, A., Poggio, L., Fajardo, M., 2021. Predicting soil properties in 3D: Should depth be a covariate? Geoderma 383, 114794. http://dx.doi.org/10.1016/j.geoderma.2020.114794.

Makungwe, M., Chabala, L.M., Chishala, B.H., Lark, R.M., 2021. Performance of linear mixed models and random forests for spatial prediction of soil pH. Geoderma 397, 115079. http://dx.doi.org/10.1016/j.geoderma.2021.115079.

Malone, B.P., Minasny, B., McBratney, A.B., 2017. using R for Digital Soil Mapping. Springer International Publishing, Cham, http://dx.doi.org/10.1007/978-3-319-44327-0_1.

Meinshausen, N., 2006. Quantile regression forests. J. Mach. Learn. Res. 7, 983–999.

Møller, A.B., Beucher, A.M., Pouladi, N., Greve, M.H., 2020. Oblique geographic coordinates as covariates for digital soil mapping. Soil 6 (2), 269–289. http://dx.doi.org/10.5194/soil-6-269-2020.

Mulder, V.L., Lacoste, M., Richer-de Forges, A.C., Arrouays, D., 2016. GlobalSoilMap France: High-resolution spatial modelling the soils of France up to two meter depth. Sci. Total Environ. 573, 1352–1369. http://dx.doi.org/10.1016/j.scitotenv.2016.07.066.

NCAR, 2015. Verification: Weather forecast verification utilities: R package version 1.42, research applications laboratory (NCAR). URL https://CRAN.R-project.org/package=verification.

Nussbaum, M., Papritz, A., Baltensweiler, A., Walthert, L., 2014. Estimating soil organic carbon stocks of swiss forest soils by robust external-drift kriging. Geosci. Model Dev. 7 (3), 1197–1210. http://dx.doi.org/10.5194/gmd-7-1197-2014.

Nussbaum, M., Walthert, L., Fraefel, M., Greiner, L., Papritz, A., 2017. Mapping of soil properties at high resolution in Switzerland using boosted geoadditive models. Soil 3, 191–210. http://dx.doi.org/10.5194/soil-2017-13.

Padarian, J., Minasny, B., McBratney, A.B., 2017. Chile and the Chilean soil grid: A contribution to GlobalSoilMap. Geoderma Reg. 9, 17–28. http://dx.doi.org/10.1016/j.geodrs.2016.12.001.

Pebesma, E., 2018. Simple features for R: Standardized support for spatial vector data. R J. 10 (1), 439–446. http://dx.doi.org/10.32614/RJ-2018-009.

Poggio, L., de Sousa, L.M., Batjes, N.H., Heuvelink, G.B.M., Kempen, B., Ribeiro, E., Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil 7 (1), 217–240. http://dx.doi.org/10.5194/soil-7-217-2021, URL https://soil.copernicus.org/articles/7/217/2021/.

R Core Team, 2022. R: A language and environment for statistical computing: version 4.2.1. URL https://www.R-project.org.

Rawlins, B.G., Marchant, B.P., Smyth, D., Scheib, C., Lark, R.M., Jordan, C., 2009. Airborne radiometric survey data and a DTM as covariates for regional scale mapping of soil organic carbon across Northern Ireland. Eur. J. Soil Sci. 60 (1), 44–54. http://dx.doi.org/10.1111/j.1365-2389.2008.01092.x.

Roudier, P., Burge, O.R., Richardson, S.J., McCarthy, J.K., Grealish, G.J., Ausseil, A.-G., 2020. National scale 3D mapping of soil pH using a data augmentation approach. Remote Sens. 12 (18), 2872. http://dx.doi.org/10.3390/rs12182872.

Scholz, F.W., Stephens, M.A., 1987. K -sample Anderson–darling tests. J. Am. Stat. Assoc. 82 (399), 918–924. http://dx.doi.org/10.1080/01621459.1987.10478517.

Service center NABODAT, 2018. Swiss soil dataset – documentation version 3. URL https://www.nabodat.ch/index.php/de/service/bodendatensatz.

Signorell, A., 2023. Desctools: Tools for descriptive statistics. URL https://CRAN.R-project.org/package=DescTools, R package version 0.99.49.

Smirnov, A., 1939. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. Moscow Univ. Math. Bull. 2, 3–16.

Sparks, D.L., Singh, B., Siebecker, M.G., 2023. Environmental Soil Chemistry. Elsevier.

Styc, Q., Lagacherie, P., 2021. Uncertainty assessment of soil available water capacity using error propagation: A test in Languedoc-Roussillon. Geoderma 391, 114968. http://dx.doi.org/10.1016/j.geoderma.2021.114968.

Swisstopo, 2022. swissTLM3D: The large-scale topographic landscape model of Switzerland, version 2.0. URL https://www.swisstopo.admin.ch/en/geodata/landscape/tlm3d.html.

Thomas, G.W., 1996. Soil pH and soil acidity: Chapter 16. In: Sparks, D.L., Swift, R.S., et al. (Eds.), Methods of Soil Analysis: Part 3. Chemical Methods. In: SSSA Book Series, vol. 5, pp. 475–490.

Vaysse, K., Heuvelink, G.B.M., Lagacherie, P., 2017. Spatial aggregation of soil property predictions in support of local land management. Soil Use Manage. 33 (2), 299–310. http://dx.doi.org/10.1111/sum.12350.

Viscarra Rossel, R.A., Chen, C., Grundy, M.J., Searle, R., Clifford, D., Campbell, P.H., 2015. The Australian three-dimensional soil grid: Australia's contribution to the GlobalSoilMap project. Soil Res. 53 (8), 845–864. http://dx.doi.org/10.1071/SR14366.

Wadoux, A.M.-C., Heuvelink, G.B., Lark, R.M., Lagacherie, P., Bouma, J., Mulder, V.L., Libohova, Z., Yang, L., McBratney, A.B., 2021. Ten challenges for the future of pedometrics. Geoderma (401), 115155.

Walthert, L., Graf, U., Kammer, A., Luster, J., Pezzotta, D., Zimmermann, S., Hagedorn, F., 2010. Determination of organic and inorganic carbon, HCl. J. Plant Nutr. Soil Sci. 173 (2), 207–216.

Wang, L., Wu, W., Liu, H.-B., 2019. Digital mapping of topsoil pH by random forest with residual kriging (RFRK) in a hilly region. Soil Res. 57 (4), 387. http://dx.doi.org/10.1071/SR18319.

Wilks, D.S., 2011. Statistical Methods in the Atmospheric Sciences, third ed. Academic Press.

Wright, M.N., Ziegler, A., 2017. Ranger: A fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Softw. 77 (1), 1–17. http://dx.doi.org/10.18637/jss.v077.i01.

Zhang, L., Chen, Q., 2005. Predicting bimodal soil–water characteristic curves. J. Geotech. Geoenviron. Eng. 131 (5), 666–670.

Zimmermann, S., Widmer, D., Mathis, B., 2011. Bodenüberwachung der Zentralschweizer Kantone (KABO ZCH): Säurestatus und Versauerungszustand von Waldböden.