Mapping of soil properties at high resolution in Switzerland using boosted geoadditive models

Madlene Nussbaum[1]**, Lorenz Walthert**[2]**, Marielle Fraefel**[2]**, Lucie Greiner**[3]**, and Andreas Papritz**[1]

[1]Institute of Biogeochemistry and Pollutant Dynamics, ETH Zurich, Universitätstrasse 16, CH-8092 Zürich, Switzerland

[2]Swiss Federal Institute for Forest, Snow and Landscape Research (WSL), Zürcherstrasse 111, CH-8903 Birmensdorf, Switzerland

[3]Research Station Agroscope Reckenholz-Taenikon ART, Reckenholzstrasse 191, CH-8046 Zürich, Switzerland

Correspondence to: M. Nussbaum (madlene.nussbaum@env.ethz.ch)

Abstract.

High-resolution maps of soil properties are a prerequisite for assessing soil threats and soil functions and to foster sustainable use of soil resources. For many regions in the world accurate maps of soil properties are missing, but often sparsely sampled (legacy) soil data are available. Soil property data (response) can then be related by digital soil mapping (DSM) to spatially exhaustive environmental data that describe soil forming factors (covariates) to create spatially continuous maps. With air- and spaceborne remote sensing and multi-scale terrain analysis large sets of covariates have become common. Building parsimonious models, amenable to pedological interpretation, is then a challenging task.

We propose a new boosted geoadditive modelling framework (geoGAM) for DSM. A geoGAM models smooth nonlinear relations between responses and single covariates and combines these model terms additively. Residual spatial autocorrelation is captured by a smooth function of spatial coordinates, and nonstationary effects are included by interactions between covariates and smooth spatial functions. The core of fully automated model building for geoGAM is componentwise gradient boosting.

We illustrate the application of the geoGAM framework by using soil data from the Canton of Zurich, Switzerland. We modelled effective cation exchange capacity (ECEC) in forest topsoils as continuous response. For agricultural land we predicted the presence of waterlogged horizons in given soil depths as binary and drainage classes as ordinal responses. For the latter we used proportional odds geoGAM taking the ordering of the response properly into account. Fitted geoGAM contained only few covariates (7 to 17) selected from large sets (333 covariates for forests, 498 for agricultural land). Model sparsity allowed covariate interpretation by partial effects plots. Prediction intervals were computed by model-based bootstrapping for ECEC. Predictive performance of the fitted geoGAM, tested with independent validation data and specific skill scores (SS) for continuous, binary and ordinal responses, compared well with other studies that modelled similar soil properties. SS of 0.23 up to 0.53 (with SS = 1 for perfect predictions and SS = 0 for zero explained variance) were achieved depending on response and type of score. geoGAM combines efficient model building from large sets of covariates with ease of effect interpretation and therefore likely raises the acceptance of DSM products by end-users.

1   Introduction

Soils fulfil many functions important for agriculture, forestry and the management of soil resources and natural hazards. The functionality of soils depends on their properties, hence, accurate and spatially highly resolved maps of basic soil properties such as texture, organic carbon content and pH for specific soil depth are needed for sustainable management of soils (FAO and ITPS, 2015). Unfortunately, such soil property maps are often missing and availability of soil information is very different between nations and continents (Omuto and Nachtergaele, 2013). For areas where spatially referenced, but sparse (legacy) soil data is available, e.g. soil datasets consisting of soil profile data and laboratory measurements, these point data can be linked using digital soil mapping (DSM) techniques (e.g. McBratney et al., 2003; Scull et al., 2003) to spatial information on soil formation factors to generate spatially continuous maps.

In the past, many DSM approaches have been proposed to exploit the correlation between soil properties (response $Y(\mathbf{s})$) and soil forming factors (covariates $\mathbf{x}(\mathbf{s})$). *Linear regression modelling* (LM, e.g. Meersmans et al., 2008; Hengl et al., 2014) and *kriging with external drift* (EDK), its extension for autocorrelated errors (Bourennane et al., 1996; Nussbaum et al., 2014), have been often used. Strengths of LM and EDK are the ease of interpretation of the fitted models (e.g. by partial residual plots, Faraway, 2005, p. 73). This is important for checking whether modelled relations between the target soil property and soil forming factors accord with pedological expertise and for conveying results of DSM analyses to users of such products. LM and EDK capture only linear relations between the covariates and a response. By using interactions between covariates, one can sometimes account for nonlinear relationships, but this quickly becomes unwieldy for a large number of covariates (e.g. above 30). Fitting models to (very) large sets of covariates has become common with the advent of remotely sensed data (Ben-Dor et al., 2009; Mulder et al., 2011) and novel approaches for terrain analysis (Behrens et al., 2010). Model building, i.e. covariate selection, is then a formidable task. Although specialized methods like L2-boosting (Bühlmann and Hothorn, 2007) and lasso (least absolute shrinkage and selection operator, Hastie et al., 2009, chap. 3) are available, they have not often been used for DSM (Nussbaum et al., 2014; Liddicoat et al., 2015; Fitzpatrick et al., 2016). *Generalized linear models* (GLM, e.g. Dobson, 2002) extend linear modelling to binary, nominal (e.g. soil taxonomic units, Hengl et al., 2014; Heung et al., 2016) or ordinal responses (e.g. soil drainage classes, Campling et al., 2002). Although GLM are nonlinear models, the nonlinearly transformed conditional expectation $g(\mathrm{E}[Y(\mathbf{s})|\mathbf{x}(\mathbf{s})]) - g(\cdot)$ is some known link function – still depends linearly on covariates.

Lately, *tree-based machine learning methods* have become popular for DSM: Classification and regression trees (CART, e.g. Liess et al., 2012; Heung et al., 2016), *Cubist*, (e.g.

Henderson et al., 2005; Adhikari et al., 2013; Lacoste et al., 2016) and ensemble tree methods like *random forest* (RF, e.g. Grimm et al., 2008; Wiesmeier et al., 2011) and *boosted regression trees* (BRT, e.g. Moran and Bui, 2002; Martin et al., 2011) were used. All tree-based methods easily account for complex nonlinear relations between responses and covariates. They model continuous and categorical responses (albeit without making a difference between nominal and ordinal responses), inherently deal with incomplete covariate data and allow to model spatially changing (nonstationary) relationships. BRT and RF fit models to large sets of covariates. The structure of the fitted models can be explored by variable importance and partial dependence plots (Hastie et al., 2009, Sect. 10.9, and Martin et al., 2011, for an application). Nevertheless, tree-based ensemble methods remain complex, and results are not as easy to interpret regarding the relevant soil forming factors as results from (G)LM.

Generalized additive models (GAM, e.g. Hastie and Tibshirani, 1990, Chapt. 6) offer a compromise between ease of interpretation and flexibility in modelling nonlinear relationships. GAM expand the (possibly transformed) conditional expectation of a response given covariates as an additive series

$$g\left(\mathrm{E}[Y(\mathbf{s})\,|\,\mathbf{x}(\mathbf{s})]\right) = \nu + f(\mathbf{x}(\mathbf{s})) = \nu + \sum_j f_j(x_j(\mathbf{s})), \ \ (1)$$

where ν is a constant and $f_j(x_j(\mathbf{s}))$ are linear terms or unspecified "smooth" nonlinear functions of single covariates $x_j(\mathbf{s})$ (e.g. smoothing spline, kernel or any other scatterplot smoother) and $g(\cdot)$ is again a link function. GAM extend GLM to account for truly nonlinear relations between Y and $\mathbf{x}$ (and not just for nonlinearities imposed by g), but they limit the complexity of the fitted functions to additive combinations of simple nonlinear terms and thereby avoid the curse of dimensionality (Hastie et al., 2009, Sect. 2.5). For continuous, ordinal and nominal responses, GAM can be readily fitted to large sets of covariates by boosting (Hofner et al., 2014; Hothorn et al., 2015). Boosting handles covariate selection and avoids over-fitting if stopped early (Bühlmann and Hothorn, 2007). Hence, the structure of boosted GAM can be more easily checked and interpreted than RF and BRT models. In the past, GAM have occasionally been used for DSM and only recently became more popular (e.g. Buchanan et al., 2012; Poggio et al., 2013; Poggio and Gimona, 2014; de Brogniez et al., 2015; Sindayihebura et al., 2017).

Besides accurate predictions, sometimes also accurate modelling of *prediction uncertainty* matters for DSM (e.g. for mapping temporal changes of soil carbon and nutrients stocks). *Quantile regression forest* (Meinshausen, 2006), an extension of RF, estimates the quantiles of the distributions $Y(\mathbf{s})|\mathbf{x}(\mathbf{s})$ and provides prediction intervals directly. Prediction intervals can also easily be constructed for predictions by EDK, (G)LM and GAM, as long as the uncertainty arising from model building is ignored. To take the effect of model

building properly into account one resorts best to *bootstrapping* (Davison and Hinkley, 1997, Sect. 6.3.3). Bootstrapping is also useful to model prediction uncertainty for boosted models, which *per se* do not qualify the accuracy of predictions, and to account for all sources of prediction uncertainty of regression kriging approaches (Viscarra Rossel et al., 2014).

In summary, a versatile DSM procedure should

1. model nonlinear relations between $Y(\mathbf{s})$ and $\mathbf{x}(\mathbf{s})$, where responses and covariates may be continous, binary, nominal or ordinal variables,

2. efficiently build models with good predictive performance for large sets of covariates ($p >> 30$),

3. preferably result in parsimonious models with a simple structure that can be easily interpreted and checked for plausibility, and

4. accurately quantify the accuracy of predictions computed from the fitted models.

The objective of our work was to develop a DSM framework that meets requirements 1–4 based on *boosted geoadditive models* (geoGAM), an extension of GAM for spatial data. First, we introduce the modelling framework and describe in detail the model building procedure. Second, we use the method in three DSM case studies in the Canton of Zurich, Switzerland, aiming at different types of responses: Effective cation exchange capacity (ECEC) of forest topsoils (continuous response), presence/absence of morphological features for waterlogging in agricultural soils (binary response), and drainage classes, characterizing prevalence of anoxic conditions, again in agricultural soils (ordinal response). To assess the validity of the modelling results with independent data (obtained by splitting the original dataset into calibration and validations subsets), we used specific criteria that take the nature of the various responses properly into account. These criteria are in common use for forecast verification in atmospheric sciences (e.g. Wilks, 2011), but, to our knowledge, have not been much used for (cross-)validating DSM predictions.

2 geoGAM modelling framework

2.1 Model representation

A generalized additive model (GAM) is based on the following components (Hastie and Tibshirani, 1990, Chapt. 6 and Eq. (1)): i) *Response distribution*: Given $\mathbf{x}(\mathbf{s}) = (x_1(\mathbf{s}), x_2(\mathbf{s}), ..., x_p(\mathbf{s}))^{\mathbf{T}}$, the $Y(\mathbf{s})$ are conditionally independent observations from simple exponential family distributions. ii) *Link function*: $g(\cdot)$ relates the expectation $\mu(\mathbf{x}(\mathbf{s})) = \mathrm{E}[Y(\mathbf{s})|\mathbf{x}(\mathbf{s})]$ of the response distribution to iii) the *additive predictor* $\sum_j f_j(x_j(\mathbf{s}))$.

geoGAM extend GAM by allowing a more complex form of the additive predictor (Kneib et al., 2009; Hothorn et al., 2011): First, one can add a smooth function $f_{\mathbf{s}}(\mathbf{s})$ of the spatial coordinates (smooth spatial surface) to the additive predictor to account for residual autocorrelation. More complex relationships between Y and $\mathbf{x}$ can be modelled by adding terms like $f_j(x_j(\mathbf{s})) \cdot f_k(x_k(\mathbf{s}))$ – capturing the effect of interactions between covariates – and $f_{\mathbf{s}}(\mathbf{s}) \cdot f_j(x_j(\mathbf{s}))$ – accounting for spatially changing dependence between Y and $\mathbf{x}$. Hence, in its full generality, a generalized additive model for spatial data is represented by

$$g(\mu(\mathbf{x}(\mathbf{s}))) = \nu + f(\mathbf{x}(\mathbf{s})) =$$
$$\nu + \underbrace{\sum_u f_{j_u}(x_{j_u}(\mathbf{s})) + \sum_v f_{j_v}(x_{j_v}(\mathbf{s})) \cdot f_{k_v}(x_{k_v}(\mathbf{s}))}_{\text{global marginal and interaction effects}}$$
$$+ \underbrace{\sum_w f_{\mathbf{s}_w}(\mathbf{s}) \cdot f_{j_w}(x_{j_w}(\mathbf{s}))}_{\text{nonstationary effects}} + \underbrace{f_{\mathbf{s}}(\mathbf{s})}_{\text{autocorrelation}} . \quad (2)$$

Kneib et al. (2009) called Eq. (2) a geoaddrive model, a name coined before by Kammann and Wand (2003) for a combination of Eq. (1) with a geostatistical error model.

It remains to specify what response distributions and link functions should be used for the various response types: For (possibly transformed) *continuous* responses one uses often a normal response distribution combined with the identity link $g(\mu(\mathbf{x}(\mathbf{s}))) = \mu(\mathbf{x}(\mathbf{s}))$. For binary data (coded as 0 and 1), one assumes a Bernoulli distribution and uses often a logit link

$$g(\mu(\mathbf{x}(\mathbf{s}))) = \log\left(\frac{\mu(\mathbf{x}(\mathbf{s}))}{1 - \mu(\mathbf{x}(\mathbf{s}))}\right), \quad (3)$$

where

$$\mu(\mathbf{x}(\mathbf{s})) = \mathrm{Prob}[Y(\mathbf{s}) = 1 \,|\, \mathbf{x}(\mathbf{s})] = \frac{\exp(\nu + f(\mathbf{x}(\mathbf{s})))}{1 + \exp(\nu + f(\mathbf{x}(\mathbf{s})))}. \quad (4)$$

For ordinal data, with ordered response levels, $1, 2, \ldots, k$, we used the cumulative logit or proportional odds model (Tutz, 2012, Sect. 9.1). For any given level $r \in (1, 2, \ldots, k)$, the logarithm of the odds of the event $Y(\mathbf{s}) \le r \,|\, \mathbf{x}(\mathbf{s})$ is then modelled by

$$\log\left(\frac{\mathrm{Prob}[Y(\mathbf{s}) \le r \,|\, \mathbf{x}(\mathbf{s})]}{\mathrm{Prob}[Y(\mathbf{s}) > r \,|\, \mathbf{x}(\mathbf{s})]}\right) = \nu_r + f(\mathbf{x}(\mathbf{s})), \quad (5)$$

with ν_r a sequence of level-specific constants satisfying $\nu_1 \le \nu_2 \le \ldots \le \nu_r$. Conversely,

$$\mathrm{Prob}[Y(\mathbf{s}) \le r \,|\, \mathbf{x}(\mathbf{s})] = \frac{\exp(\nu_r + f(\mathbf{x}(\mathbf{s})))}{1 + \exp(\nu_r + f(\mathbf{x}(\mathbf{s})))}. \quad (6)$$

Note that $\mathrm{Prob}[Y(\mathbf{s}) \le r \,|\, \mathbf{x}(\mathbf{s})]$ depends on r only through the constant ν_r. Hence, the ratio of the odds of two events $Y(\mathbf{s}) \le r \,|\, \mathbf{x}(\mathbf{s})$ and $Y(\mathbf{s}) \le r \,|\, \tilde{\mathbf{x}}(\mathbf{s})$ is the same for all r (Tutz, 2012, p. 245).

2.2   Model building (selection of covariates)

To build parsimonious models that can readily be checked for agreement with pedological understanding, we applied a sequence of fully automated steps 1–6. In several of these steps we optimised tuning parameters by 10-fold cross-validation with fixed subsets using either root mean squared error (RMSE, Eq. (12), continuous responses), Brier score (BS, Eq. (16), binary responses) or ranked probability score (RPS, Eq. (18), ordinal responses) as optimisation criteria. Model building aims to optimise accuracy of predictions, hence we did not use equivalent "goodness-of-fit" statistics. To improve the stability of the algorithm continuous covariates were first scaled (by difference of maximum and minimum value) and centred.

1. Boosting (see step 2 below) is more stable and converges more quickly when the effects of categorical covariates (factors) are accounted for as model offset. We therefore used the group lasso (Breheny and Huang, 2015) – an algorithm that likely excludes non-relevant covariates and treats factors as groups – to select important factors for the offset. For ordinal responses (Eq. (6)) we used stepwise proportional odds logistic regression in both directions with BIC (e.g. Faraway, 2005, p. 126) to select the offset covariates because lasso cannot be used for such responses.

2. Next, we selected a subset of relevant factors, continuous covariates and spatial effects by componentwise gradient boosting. Boosting is a slow stagewise additive learning algorithm. It expands $f(\mathbf{x}(\mathbf{s}))$ in a set of base procedures (baselearners) and approximates the additive predictor by a finite sum of them as follows (Bühlmann and Hothorn, 2007):

   (a) Initialize $\hat{f}(\mathbf{x}(\mathbf{s}))^{[m]}$ by offset of step 1 above and set $m = 0$.

   (b) Increase m by 1. Compute the negative gradient vector $\mathbf{U}^{[m]}$ (e.g. residuals) for a loss function $l(\cdot)$.

   (c) Fit all baselearners $f_j(x_j(\mathbf{s})), j = 1, ..., p$ to $\mathbf{U}^{[m]}$ and select the baselearner, say $f_k(x_k(\mathbf{s}))^{[m]}$ that minimizes $l(\cdot)$.

   (d) Update $\hat{f}(\mathbf{x}(\mathbf{s}))^{[m]} = \hat{f}(\mathbf{x}(\mathbf{s}))^{[m-1]} + v \cdot f_k(x_k(\mathbf{s}))^{[m]}$ with step size $v \leq 1$.

   (e) Iterate steps (b) to (d) until $m = m_{\text{stop}}$ (main tuning parameter).

   We used the following settings in above algorithm: As loss functions $l(\cdot)$ we used L_2 for continuous, negative binomial likelihood for binary (Bühlmann and Hothorn, 2007) and proportional odds likelihood for ordinal responses (Schmid et al., 2011). Early stopping of the boosting algorithm was achieved by determining optimal m_{stop} by cross-validation. We used default step length ($v = 0.1$). This is not a sensitive parameter as long as it is clearly below 1 (Hofner et al., 2014). For continuous covariates we used penalized smoothing spline baselearners (Kneib et al., 2009). Factors were treated as linear baselearners. To capture residual autocorrelation we added a bivariate tensor-product P-spline of spatial coordinates (Wood, 2006, pp. 162) to the additive predictor. Spatially varying effects were modelled by baselearners formed by multiplication of continuous covariates with tensor-product P-splines of spatial coordinates (Wood, 2006, pp. 168). Uneven degree of freedom of baselearners biases baselearner selection (Hofner et al., 2011). We therefore penalized each baselearner to 5 degrees of freedom (df). Factors with less than 6 levels ($df < 5$) were aggregated to grouped baselearners. By using an offset, effects of important factors with more than 6 levels were implicitly accounted for without penalization.

3. At m_{stop} (see step 2 above), many included baselearners had very small effects only. We fitted generalized additive models (GAM, Wood, 2011) and included smooth and factor effects only if their effect size e_j of the corresponding baselearner $f_j(x_j(\mathbf{s}))$ was substantial. Effect size e_j of factors was the largest difference between effects of two levels and for continuous covariates it was equal to the maximum contrast of estimated partial effects (after removal of extreme values as in boxplots, Frigge et al., 1989). We iterated through e_j and excluded covariates with e_j smaller than a threshold effect size e_t. Optimal e_t was determined by 10-fold cross-validation of GAM. In these GAM fits smooth effects were penalized by 5 degrees of freedom as imposed by componentwise gradient boosting (step 2 above). The factors selected as offset in step 1 were now included in the GAM.

4. We further reduced the GAM by stepwise removal of covariates by cross-validation. The candidate covariate to drop was chosen by largest p value of F tests for linear terms and approximate F tests (Wood, 2011) for smooth terms.

5. Factor levels with similar estimated effects were merged stepwise again by cross-validation based on largest p values from two sample t-tests of partial residuals.

6. The final model (used to compute spatial predictions) was a parsimonious GAM. Because of step 5, factors had possibly a reduced number of coefficients. Effects of continuous covariates were modelled by smooth functions and – if at all present – spatially structured residual variation (autocorrelation) was represented by a smooth spatial surface. To avoid over-fitting both types of smooth effects were penalized by 5 degrees of freedom (as imposed by step 2).

Model building steps 1 to 6 were implemented in the R package geoGAM (Nussbaum, 2017).

2.3 Predictions and predictive distribution

Soil properties were predicted for new locations $\mathbf{s}_+$ from the final geoGAM by $\tilde{Y}(\mathbf{s}_+) = \hat{f}(\mathbf{x}(\mathbf{s}_+))$. To model the predictive distributions for continuous responses we used a non-parametric, model-based bootstrapping approach (Davison and Hinkley, 1997, pp. 262, 285) as follows:

A. New values of the response were simulated according to $Y(\mathbf{s})^* = \hat{f}(\mathbf{x}(\mathbf{s})) + \epsilon$, where $\hat{f}(\mathbf{x}(\mathbf{s}))$ are the fitted values of the final model and ϵ are errors randomly sampled with replacement from the centred, homoscedastic residuals of the final model (Wood, 2006, p. 129).

B. The geoGAM was fitted to $Y(\mathbf{s})^*$ according to steps 1–6 of Sect. 2.2.

C. Prediction errors were computed according to $\delta_+^* = \hat{f}(\mathbf{x}(\mathbf{s}_+))^* - (\hat{f}(\mathbf{x}(\mathbf{s}_+)) + \epsilon)$, where $\hat{f}(\mathbf{x}(\mathbf{s}_+))^*$ are predicted values at new locations $\mathbf{s}_+$ of the model built with the simulated response $Y(\mathbf{s})^*$ in step B above, and the errors ϵ are again randomly sampled from the centred, homoscedastic residuals of the final model (see step A).

Prediction intervals were computed according to

$$[\hat{f}(\mathbf{x}(\mathbf{s}_+)) - \delta_{+(1-\alpha)}^* ; \hat{f}(\mathbf{x}(\mathbf{s}_+)) - \delta_{+(\alpha)}^*]. \quad (7)$$

where $\delta_{+(\alpha)}^*$ and $\delta_{+(1-\alpha)}^*$ are the α- and $(1-\alpha)$-quantiles of δ_+^*, pooled over all 1000 bootstrap repetitions.

Predictive distributions for binary and ordinal responses were directly obtained from the final geoGAM fit by predicting probabilities of occurrence $\widetilde{\mathrm{Prob}}(Y(\mathbf{s}) = r \,|\, \mathbf{x}(\mathbf{s}))$ (Davison and Hinkley, 1997, p. 358).

3 Case studies - Materials and Methods

3.1 Study regions

We applied the modelling framework to 3 datasets on properties of forest and agricultural soils in the Canton of Zurich in Switzerland (Fig. 1). Forests (ZH forest), as defined by the Swiss topographic landscape model (swissTLM3D, Swisstopo, 2013a), cover an area of 506.5 km^2, or roughly 30 % of the total area of the Canton of Zurich. The spatial extent of the agricultural region near the Lake Greifensee was defined by the availability of imaging spectroscopy data collected by the APEX spectrometer (Schaepman et al., 2015). Agricultural land was defined as the area not covered by any areal features such as settlements or forests extracted from the Swiss topographic landscape model (swissTLM3D, Swisstopo, 2013a). Wetlands, forests, parks or city gardens were excluded, resulting in a study region of 170 km^2.

Data sources: Biogeographical regions © 2001 BAFU / Swiss Boundary, Lakes © 2012 BFS GEOSTAT / Boundries Europe: NUTS © 2010 EuroGeographics
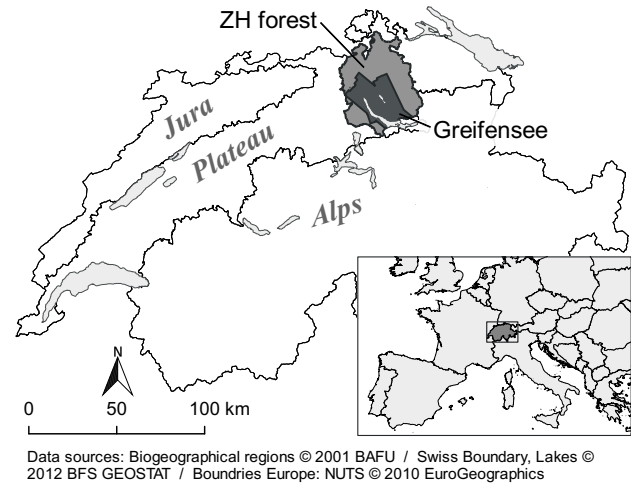
Figure 1. Location of the study regions Greifensee and ZH forest on the Swiss Plateau.

In the Canton of Zurich, forests extend across altitudes ranging from 340 to 1170 m above sea level (a.s.l), and in the Greifensee area elevation ranges from 390 to 840 m a.s.l. (Swisstopo, 2016). The climatic conditions (period 1961–1990, Zimmermann and Kienast, 1999) vary accordingly, with mean annual rainfall between 880–1780 mm for the forested and 1040–1590 mm for the agricultural study region. Mean annual temperatures range between 6.1–9.1 °C and 7.5–9.1 °C, respectively. Two thirds of the forested area is dominated by coniferous trees (FSO, 2000b). Half of the Greifensee study region is covered by crop land and one third by permanent grassland. The remainder are orchards, horticultural areas or mountain pastures (Hotz et al., 2005). In the Canton of Zurich, soils formed mostly from Molasse formations and quaternary sediments dominantly from the last glaciation (Würm). In the north-eastern part, the limestone Jura hills reach into the ZH forest study region (Hantke, 1967).

3.2 Data

3.2.1 Soil data base

We used legacy soil data collected between 1985 and 2014. Data originates from long-term soil monitoring of the Canton of Zurich (KaBo), a soil pollutant survey (Wegelin, 1989), field surveys for creating soil maps of the agricultural land (Jäggli et al., 1998) or soil investigations in the course of different projects by the Swiss Federal Institute for Forest, Snow and Landscape Research (WSL, Walthert et al., 2004). Sites for pollutant surveying were chosen on a regular grid, those for creating soil maps were determined by purposive sampling (Webster and Lark, 2013, p. 86) by field surveyors to best represent soils typical for the given landform. The sites of WSL were chosen by purposive sampling according to the aims of the project. Soil data was therefore quite het-

erogeneous, and tailored harmonisation procedures were required to provide consistent soil datasets. The heterogeneity resulted from several standards of soil description and soil classification, different data keys, different analytical methods and in particular, often missing metadata for a proper interpretation of the datasets. Therefore, we elaborated a general harmonisation scheme that covers main steps required to merge different soil legacy data into one common consistent database (Walthert et al., 2016). Sampling sites were recorded in the field on topographic maps (scale 1:25 000), hence we estimated accuracy of coordinates to about $\pm$ 25 m.

3.2.2   Effective cation exchange capacity (ECEC, forest soils)

After the removal of sites with missing covariate values, we used 1844 topsoil samples from 1348 sites with data on effective cation exchange capacity (ECEC). Most measurements refer to composite samples where aliquots were measured in 20 by 20 m squares from 0–20 cm soil depth. For about 100 sites soil profiles genetic horizons were sampled. ECEC [$mmol_c\ kg^{-1}$] for 0–20 cm was computed from horizon data by

$$\mathrm{ECEC}_{0--20} = \sum_{i=1}^{h} w_i\ \mathrm{ECEC}_i, \qquad (8)$$

where ECEC_i is the value for horizon i, w_i is a weight given by soil density ρ_i and the fraction of the thickness of horizon i within 0–20 cm and h is the number of horizons intersecting the 0–20 cm depth. ρ_i was estimated from soil organic matter (SOM) and/or sampling depth by a pedotransfer function (PTF see Supplement of Nussbaum et al., 2017). Due to a lack of respective data, the volumetric stone content was assumed to be constant.

For most soil samples, ECEC was measured after extraction in an ammonium chloride solution (FAC, 1989; Walthert et al., 2004, 2013). Roughly 5 % of the samples had only measurements of Ca, Mg, K and Al (extracted by ammonium acetate EDTA solution, Lakanen and Erviö, 1971; ELF, 1996; Gasser et al., 2011). For these samples, we estimated ECEC by using a PTF (Nussbaum and Papritz, 2015).

We assigned 293 of 1348 sites (528 samples) to the validation set, which was used to check the predictive performance of the fitted statistical model, and the remaining 1055 sites (1316 samples) were used to calibrate the model. The legacy samples were spatially clustered. To ensure that the validation sites were evenly spread over the study region, the validation sites were selected by weighted random sampling. The weight attributed to a site was proportional to the forested area within its Dirichlet polygon (Dirichlet, 1850).

We found a considerable variation in ECEC values ranging from 17.4 to 780 $mmol_c\ kg^{-1}$ (median 141.1 $mmol_c\ kg^{-1}$,

Table S1 in Supplement). On average, ECEC was slightly larger in the calibration than in the validation set.

3.2.3   Presence of waterlogged soil horizons (agricultural soils)

Waterlogging characteristics were recorded in the field at 962 sites within the Greifensee study region by visual evaluation (Jäggli et al., 1998). Swiss soil classification distinguishes horizon qualifiers *gg* (strongly gleyic, predominantly oxidized) and *r* (anoxic, predominantly reduced) and both are believed to limit plant growth (Jäggli et al., 1998; Müller et al., 2007; Litz, 1998; Danner et al., 2003; Kreuzwieser and Rennberg, 2014).

We constructed binary responses for three soil depths 0–30 cm, 0–50 cm and 0–100 cm. If one of the horizon qualifiers *gg* or *r* was recorded within the interval, we assigned 1 = *presence of waterlogged horizons* and 0 = *absence of waterlogged soil horizons* otherwise.

We chose 198 of 962 sites to form a validation set, again by using weighted random sampling. The remaining 764 sites were used to build and fit the models. In the topsoil (0–30 cm) *gg* or *r* horizon qualifiers were only observed at 13.4 % of the 962 sites. Down to 50 cm about twice as many sites (25.9 %) showed signs of anoxic conditions and down to 1 m already 38.6 % of sites featured an anoxic or gleyic horizon (Table S2 in Supplement).

3.2.4   Drainage classes (agricultural soils)

Swiss soil classification differentiates hydromorphic features of soils in more detail, describing the degree, depth and source of waterlogging by 3 supplementary qualifiers for stagnic, gleyic or anoxic profiles (I, G, R; categorical attributes, Brunner et al., 1997). To reduce complexity of classification, we aggregated these qualifiers to three ordered levels *well drained* (qualifiers I1–I2, G1–G3, R1 or no hydromorphic qualifier), *moderately well drained* (I3–I4, G4) and *poorly drained* (G5–G6, R2–R5).

For validation we used the same 198 sites as for *presence of waterlogged soil horizons*, but only 732 sites were used for model building due to missing supplementary qualifiers. The majority (66.6 %) of the 930 sites were *well drained*, only 12.7 % were classified as *moderately well drained* and 20.7 % as *poorly drained* (Table S3 in Supplement).

3.2.5   Covariates for statistical modelling

To represent local soil formation conditions, we used data from 23 sources (Table 1). For ECEC a total of 333 covariates were used to describe climatic (71 covariates) and topographic conditions (196 covariates). For the agricultural land, we used in addition 180 spectral bands of the APEX spectrometer, spatial information on historic wetlands and

agricultural drainage networks resulting in 498 covariates in total.

3.3 Statistical analysis

We built models for the five responses according to Sect. 2.2 and computed predictions for new locations at nodes of a 20 m-grid. Predictions were post-processed in the following way:

3.3.1 Response transformation

ECEC data in 0–20 cm soil depth was positively skewed (Table S1 in Supplement), hence we fitted the model to the log-transformed data. In full analogy to lognormal kriging (Cressie, 2006, Eq. (20)), the predictions were backtransformed by

$$\mathrm{E}[Y(\mathbf{s}_+)\,|\,\mathbf{x}] = \exp\left(\,\hat{f}(\mathbf{x}(\mathbf{s}_+)) + \frac{1}{2}\hat{\sigma}^2 - \frac{1}{2}\mathrm{Var}[\hat{f}(\mathbf{x}(\mathbf{s}_+)]\,\right) \quad (9)$$

with $\hat{f}(\mathbf{x}(\mathbf{s}_+))$ being the prediction of the log-transformed response, $\hat{\sigma}^2$ the estimated residual variance of the final geoGAM fit and $\mathrm{Var}[\hat{f}(\mathbf{x}(\mathbf{s}_+)]$ the variance of $\hat{f}(\mathbf{x}(\mathbf{s}_+))$ as provided again by the final geoGAM. Limits of prediction intervals were backtransformed by $\exp(\cdot)$ as they are quantiles of the predictive distributions.

3.3.2 Conversion of probabilistic to categorical predictions

For binary and ordinal responses, Eq. (4) and (6) predict probabilities of the respective response levels. To predict the "most likely" outcome one has to apply a threshold to these probabilities. For binary data we predicted *presence of waterlogged horizons* if the probability exceeded the optimal value of the Gilbert skill score (GSS, Sect. 3.3.3) that discriminated *presence* and *absence of waterlogged horizons* best in cross-validation of the final geoGAM. GSS was selected because *absence of waterlogged horizons* was more common than *presence*, especially in topsoil. To ensure consistency of maps for sequential soil depths we assigned *presence of waterlogged horizons* to the lower depth if it was predicted for the depth above.

For ordinal responses we predicted the level to which the median of the probability distribution $\widetilde{\mathrm{Prob}}(Y(\mathbf{s}) \leq r | \mathbf{x}(\mathbf{s}))$ was assigned (Tutz, 2012, p. 475).

3.3.3 Evaluating the predictive performance of the statistical models

The predictive performance of the geoGAM, fitted for the continuous response ECEC, was tested by comparing predictions $\tilde{Y}(\mathbf{s}_i)$ (Eq. (9)) with measurements $Y(\mathbf{s}_i)$ of independent validation sets. Marginal bias and overall accuracy

were assessed by

$$\mathrm{BIAS} = -\frac{1}{n}\sum_{i=1}^{n}(Y(\mathbf{s}_i) - \tilde{Y}(\mathbf{s}_i)), \quad (10)$$

$$\mathrm{robBIAS} = -\mathrm{median}_{1 \leq i \leq n}\left(Y(\mathbf{s}_i) - \tilde{Y}(\mathbf{s}_i)\right), \quad (11)$$

$$\mathrm{RMSE} = \left(\frac{1}{n}\sum_{i=1}^{n}\left(Y(\mathbf{s}_i) - \tilde{Y}(\mathbf{s}_i)\right)^2\right)^{1/2}, \quad (12)$$

$$\mathrm{robRMSE} = \mathrm{MAD}_{1 \leq i \leq n}\left(Y(\mathbf{s}_i) - \tilde{Y}(\mathbf{s}_i)\right), \quad (13)$$

$$\mathrm{SS}_{\mathrm{mse}} = 1 - \frac{\sum_{i=1}^{n}\left(Y(\mathbf{s}_i) - \tilde{Y}(\mathbf{s}_i)\right)^2}{\sum_{i=1}^{n}\left(Y(\mathbf{s}_i) - \frac{1}{n}\sum_{i=1}^{n}Y(\mathbf{s}_i)\right)^2}, \quad (14)$$

where MAD is the median absolute deviation. $\mathrm{SS}_{\mathrm{mse}}$ was defined as mean squared error skill score (Wilks, 2011, p. 359) with the sample mean of the measurements as reference prediction method. Interpretation is similar to R^2 with $\mathrm{SS}_{\mathrm{mse}} = 1$ for perfect predictions and $\mathrm{SS}_{\mathrm{mse}} = 0$ for zero explained variance. $\mathrm{SS}_{\mathrm{mse}}$ becomes negative if the root mean squared error (RMSE) exceeds the standard deviation of the data. To validate the accuracy of the bootstrapped predictive distributions we plotted the empirical distribution function of the probability integral transform (Wilks, 2011, p. 375), which is equivalent to a plot of the coverage of one-sided prediction intervals $(0, \tilde{q}_\alpha(\mathbf{s}))$ against the nominal probabilities α used to construct the quantiles $\tilde{q}_\alpha(\mathbf{s})$.

For binary responses the predictive performance of the fitted geoGAM was evaluated with independent validation data by the Brier skill score (BSS, Wilks, 2011, Eq. (8.37))

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}} \quad (15)$$

where the Brier score (BS) is computed by

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}(y_i - o_i)^2 \quad (16)$$

where n is the number of sites, $y_i = \widetilde{\mathrm{Prob}}[Y(\mathbf{s}_i) = 1 | \mathbf{x}(\mathbf{s}_i)]$ are the predicted probabilities and $o_i = \mathrm{I}(Y(\mathbf{s}_i) = 1)$ the observation. $\mathrm{BS}_{\mathrm{ref}}$ is the BS of a reference prediction where always the more abundant level (*absence of waterlogged horizons*) is predicted. After transforming the predicted probabilities to the binary levels *presence or absence of waterlogged horizons* (Sect. 3.3.2) we further evaluated the bias ratio, Peirce skill score (PSS) and GSS. Bias ratio is the ratio of the number of *presence* predictions to the number of *presence* observations (Wilks, 2011, Eq. (8.10)). PSS is a skill score based on the proportion of correct *presence* and *absence* predictions where the reference predictions are purely random predictions that are constrained to be unbiased (Wilks, 2011, Eq. (8.16)). GSS is a skill score that uses the threat score as accuracy measure (Wilks, 2011, Eq. (8.18)) and again random predictions as reference. Perfect predictions have PSS

Table 1. Overview of geodata and derived covariates, for more information see Supplement of Nussbaum et al. (2017) (r: pixel resolution for raster datasets or scale for vector datasets, a: only available for study region Greifensee (Gr) or ZH forest (Zf), NDVI: normalized differenced vegetation index, TPI: topographic position index, TWI: topographic wetness index, MRVBF: multi-resolution valley bottom flatness).

geodata set	r	a	covariate examples
Soil			physiographical units, historic wetland
Soil overview map (FSO, 2000a)	1:200 000		presence, presence of drainage
Wetlands Wild maps (ALN, 2002)	1:50 000	Gr	networks or soil ameliorations
Wetlands Siegfried maps (Wüst-Galley et al., 2015)	1:25 000	Gr	
Anthropogenic soil interventions (AWEL, 2012)	1:5 000	Gr	
Drainage networks (ALN, 2014b)	1:5 000	Gr	
Parent material			(aggregated) geological units, ice level
Last Glacial Maximum (Swisstopo, 2009)	1:500 000		during last glaciation, information on
Geotechnical map (BFS, 2001)	1:200 000		aquifers
Geological map (ALN, 2014a)	1:50 000		
Groundwater occurrence (AWEL, 2014)	1:25 000	Gr	
Climate			mean annual/monthly temperature,
MeteoSwiss 1961–1990 (Zimmermann and Kienast, 1999)	25/100 m		precipitation, radiation, degree days,
MeteoTest 1975–2010 (Remund et al., 2011)	250 m		NH_3 concentration in air
Air pollutants (BAFU, 2011)	500 m	Zf	
NO_2 immissions (AWEL, 2015)	100 m	Gr	
Vegetation			band ratios, NDVI, 180 hyperspectral
Landsat7 scene (USGS EROS, 2013)	30 m		bands, aggregated vegetation units,
DMC mosaic (DMC, 2015)	22 m		canopy height
SPOT5 mosaic (Mathys and Kellenberger, 2009)	10 m	Zf	
APEX spectrometer mosaics (Schaepman et al., 2015)	2 m	Gr	
Share of coniferous trees (FSO, 2000b)	25 m	Zf	
Vegetation map (Schmider et al., 1993)	1:5 000	Zf	
Species composition data (Brassel and Lischke, 2001)	25 m	Zf	
Digital surface model (Swisstopo, 2011)	2 m	Zf	
Topography			slope, curvature, northness, TPI, TWI,
Digital elevation model (Swisstopo, 2011)	25 m		MRVBF (various radii/resolutions)
Digital terrain model (Swisstopo, 2013b)	2 m		

and GSS equal to 1, for random predictions the scores are equal to 0 and predictions worse than the reference receive negative scores. PSS is truly and GSS asymptotically equitable, meaning that purely random and constant predictions get the same scores (see Wilks, 2011, p. 316 and 321 for details).

For the ordinal response drainage classes we tested the fitted geoGAM by evaluating the ranked probability skill score (RPSS), computed for the independent validation data analogously to BSS by

$$\text{RPSS} = 1 - \frac{\text{RPS}}{\text{RPS}_\text{ref}} \qquad (17)$$

where RPS is the ranked probability score (RPS, Wilks, 2011, Eq. (8.52)) given by

$$\text{RPS} = \sum_{i=1}^{n} \sum_{j=1}^{J} (Y_{i,j} - O_{i,j})^2 \qquad (18)$$

with $Y_{i,j} = \widetilde{\text{Prob}}[Y(\mathbf{s}_i) \leq j \,|\, \mathbf{x}(\mathbf{s}_i)]$ being the predicted cumulative probabilities up to class j and $O_{i,j} = \sum_{r=1}^{j} I(Y(\mathbf{s}_i) = r)$ indicating observed absence (0) or presence (1) up to class j. RPS_ref is the RPS for a reference that predicts always the most abundant class (*well drained*). For predictions of the ordinal outcomes (Sect. 3.3.2) we also computed mean bias ratio from three bias ratios created analogously to the binary case. These two-class settings were achieved by stepwise aggregation of two out of three classes (*well* vs. *moderately well* or *poorly drained*, then *well* or *moderately well* vs. *poorly drained*, Wilks, 2011, p. 319). In addition, PSS was computed in its general form (Wilks, 2011, p. 319) together with the Gerrity score (GS), which applies weights to the joint distribution of predicted and observed classes to consider their ordering and frequency (Wilks, 2011, p. 322).
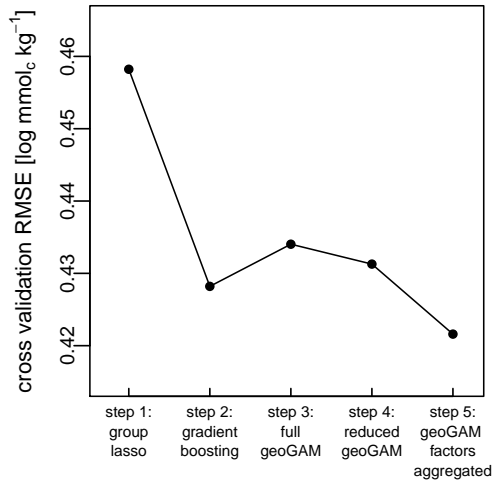
Figure 2. Change of cross-validation root mean squared error (RMSE) in steps 1–5 of model building procedure (Sect. 2.2).
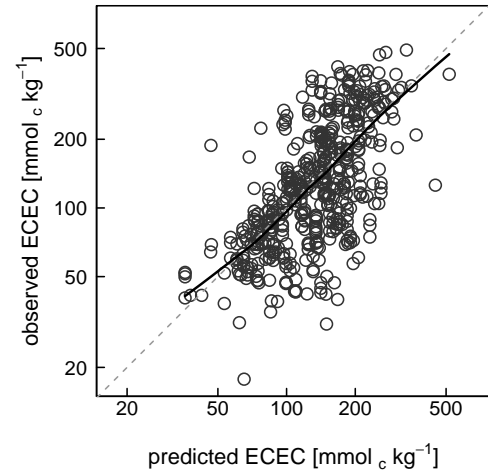
Figure 3. Scatter plot of measured against predicted ECEC in 0–20 cm mineral soil depth, computed with geoGAM (Sect. 4.1.1) for the sites of the validation set (solid line: loess scatter plot smoother).

3.3.4 Software

Terrain attributes were computed by ArcGIS (version 10.2, ESRI, 2010) and SAGA 2.1.4 (version 2.1.4, Conrad et al., 2015). All statistical computations were done in R (version 3.2.2, R Core Team, 2016) using several add-on packages, in particular `grpreg` for group lasso (version 2.8-1, Breheny and Huang, 2015), `MASS` for proportional odds logit regression (version 7.3-43, Venables and Ripley, 2002), `mboost` for componentwise gradient boosting (version 2.5-0, Hothorn et al., 2015), `mgcv` for geoadditive model fits (version 1.8-6, Wood, 2011), `raster` for spatial data processing (version 2.4-15, Hijmans et al., 2015) and `geoGAM` for the model building routine (version 0.1-2, Nussbaum, 2017).

4 Results

4.1 ECEC – case study 1

4.1.1 Models for ECEC in 0–20 cm depth

Figure 2 shows the change of RMSE during model building (10-fold cross-validation). The small root mean squared error (RMSE) of 0.428 log $\mathrm{mmol_c\,kg^{-1}}$ after the gradient boosting step – with coefficients shrunken by the algorithm – could further be reduced (cross-validation RMSE 0.422 log $\mathrm{mmol_c\,kg^{-1}}$) by removing covariates and by factor aggregation. Aggregating factor levels resembles shrinking of coefficients of such covariates.

Starting with 333 covariates model building successfully reduced the number of covariates in the model to 17. The remaining ones characterized geology, vegetation and topography (Table 2). Effective cation exchange capacity (ECEC) depended nonlinearly on nearly all continuous covariates, but

nonlinearities were in general rather weak (Fig. S1 in Supplement). No $f_s(\mathbf{s})$ term was included in the model, because residual autocorrelation was very weak (Fig. S2 in Supplement). Including nonstationary effects in the model would have improved the model only slightly (cross-validation RMSE 0.406 log $\mathrm{mmol_c\,kg^{-1}}$), but would have added considerable complexity to the final model (21 covariates including 8 interactions with $f_s(\mathbf{s})$ terms).

4.1.2 Validation of predicted ECEC with independent data

Predictive performance, as evaluated at 293 independent validation sites, was satisfactory. Figure 3 shows for the validation set measured ECEC in 0–20 cm plotted against the predictions. The solid line of the loess scatterplot smoother (Cleveland, 1979) is close to the 1:1 line indicating absence of conditional bias. This was confirmed by small marginal BIAS measures (Table 3). BIAS^2-to-MSE ratio was small for both log-transformed and original data (1.2 and 0.7 %, respectively). robRMSE (0.411 log $\mathrm{mmol_c\,kg^{-1}}$) was somewhat smaller than RMSE (0.471 log $\mathrm{mmol_c\,kg^{-1}}$) indicating that a few outlying ECEC observations were not particularly well predicted. RMSE computed with backtransformed predictions of the validation set (74.9 $\mathrm{mmol_c\,kg^{-1}}$) was also larger than its robust counterpart robRMSE (55.3 $\mathrm{mmol_c\,kg^{-1}}$). Judged on SS$_{\mathrm{mse}}$ calculated for the independent validation data, the model explained about 40 % of the variance of the log-transformed and 37 % of the variance of the original data (Table 3).

Figure 4 shows somewhat too large coverage for quantiles in the lower tails of the predictive distributions, hence the

Table 2. Covariates contained in final geoGAM for responses ECEC, *presence of waterlogged horizons* and drainage classes. More details on covariate effects can be found in Fig. S1 and S4 to S6 in Supplement (*p*: number of covariates, SD: standard deviation in a moving window, RAD: radius of moving window or parameter of terrain attribute algorithm, r: resolution of elevation model, TPI: topographic position index, TWI: topographic wetness index, MRVBF: multiresolution valley bottom flatness).

| | ECEC 0–20 cm | presence of waterlogged horizons down to | | | drainage class |
		30 cm	50 cm	100 cm	
p	17	7	12	14	11
Legacy soil data	correction factor				
Geology, land use	distance to moraines, aquifer map, overview soil map, geological map, geotechnical map	historic wetlands	historic wetlands, drainage systems map	historic wetlands, drainage systems map, anthropogenic soil disturbance, extent last glaciation, geological map	historic wetlands, drainage systems map, aquifer map
Climate	—	global radiation (r: 250 m), precipitation (r: 250 m)	global radiation (r: 250 m), precipitation (r: 100 m)	dew point temperature (r: 250 m)	precipitation (r: 250 m)
Vegetation	SPOT5 vegetation index (r: 10 m), vegetation map	—	DMC green band (r: 22 m)	—	DMC green band (r: 22 m)
Topography	SD slope (RAD: 20 m, r: 2 m), smooth northness (RAD: 10 m, r: 2 m), ruggedness (RAD: 225 m, r: 25 m), surface convexity (RAD: 450 m, r: 25 m), negative openness (RAD: 2 km, r: 25 m), vertical distance to rivers (r: 25 m)	curvature (r: 25 m), smooth eastness (RAD: 3.6 km, r: 25 m), roughness (RAD: 50 m, r: 2 m), negative openness (RAD: 1 km, r: 2 m)	SD elevation (RAD: 3.6 km, r: 25 m), SD slope (RAD 50 m, r: 2 m), smooth curvature (RAD: 120 m, r: 2 m), negative openness (RAD: 1 km, r: 25 m), TPI (RAD: 50 m, r: 2 m), smooth TWI (RAD 14 m, r: 2 m), MRVBF (r: 25 m)	SD elevation (RAD: 3.6 km, r: 25 m), smooth curvature (RAD: 120 m, r: 2m), smooth eastness (RAD: 3.6 km, r: 25 m), convergence index (RAD: 250 m, r: 25 m), terrain texture (RAD: 60 m, r: 2 m), horizontal distance to rivers (r: 25 m), TWI (RAD: 14 m, r: 2 m), MRVBF (25 m)	SD elevation (RAD: 3.6 km, r: 25 m), terrain texture (RAD: 60 m, 2 m), TPI (RAD: 300 m and 50 m, r: 2 m), smooth TWI (RAD: 14 m, r: 2 m)

Table 3. Validation statistics for (a) log-transformed and (b) back-transformed ECEC 0–20 cm [$mmol_c\ kg^{-1}$] calculated for 528 samples (293 sites) of the independent validation set (definition of statistics see Sect. 3.3.3).

	BIAS	robBIAS	RMSE	robRMSE	SS_{mse}
(a)	0.052	0.006	0.471	0.411	0.407
(b)	6.3	8.9	74.9	55.3	0.365

extent of lower tails of bootstrapped predictive distributions was underestimated. Upper tails of the predictive distributions were modelled accurately as the coverage was close to the nominal probability there. The coverage of symmetric 90 %-prediction intervals was again too small (84.1 %) because the lower tails were too short. The median width of 90 %-prediction intervals was equal to 201.8 $mmol_c\ kg^{-1}$, demonstrating that prediction uncertainty remained substantial, in spite of SS_{mse} of nearly 40 %.
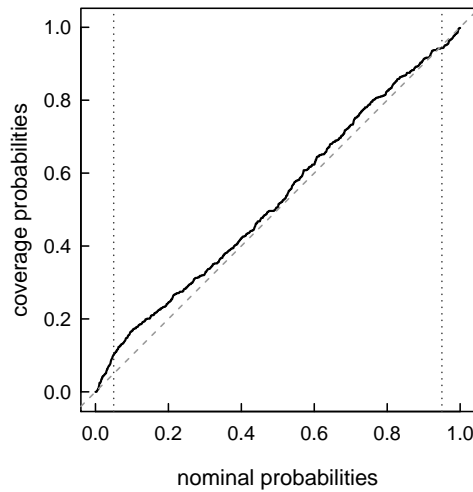
Figure 4. Coverage of one-sided bootstrapped prediction intervals $(0, \tilde{q}_\alpha(\mathbf{s}))$ for 528 ECEC validation samples, plotted against nominal probability α used to construct the upper limit q_α of the prediction intervals (Vertical lines mark the 5 and 95 % probabilities).

4.1.3   Mapping ECEC for ZH forest topsoils

Predictions of ECEC were computed by the final geoGAM for the nodes of a 20 m-grid (Fig. 5). 44 % of the mapped topsoil has large to very large ECEC values. In contrast, 13 % (~66 km^2) of the forest topsoils in the study region are acidic with ECEC below 100 $\mathrm{mmol_c\,kg^{-1}}$. These soils are mostly found in the northern part of the Canton of Zurich. The spatial pattern of the width of 90 %-prediction intervals (Fig. S3 in Supplement) and of the mean predictions (Fig. 5) was very similar (Pearson correlation = 0.981), which follows from the lognormal model that we adopted for this response.

4.2   Presence of waterlogged soil horizons – case study 2

4.2.1   Models for presence of waterlogged horizons

Not surprisingly, the models for *presence of waterlogged horizons* in the three soil depths contained similar covariates, characterizing mostly wet soil conditions such as historic wetland maps, a map of agricultural drainage systems or several climatic covariates (Table 2). The same terrain attributes were repeatedly chosen for the three depths (Figs. S4 to S6 in Supplement). For all three depths model selection resulted in parsimonious sets of only 7 to 14 covariates chosen from a total of 498 covariates. The Brier skill score (BSS), computed using 10-fold cross-validation, increased from 0.350 for the 0–30 cm depth to 0.704 for the 0–100 cm depth suggesting that *presence of waterlogged horizons* can be better modelled when they occur more frequently. Degree of residual spatial autocorrelation on logit-scale was stronger in the

0–30 cm than in 0–100 cm depth (Fig. S2 in Supplement) confirming that the model performed better for the 0–100 cm depth. Adding a $f_s(\mathbf{s})$ term did not improve cross-validated BSS (30 cm: 0.332, 100 cm: 0.688), meaning that a penalized tensor product of spatial coordinates was too smooth to capture short range autocorrelation.

4.2.2   Validation of predicted presence of waterlogged horizons with independent data

Table 4 reports contingency tables for predicted outcomes for *presence of waterlogged horizons* at 198 sites of the validation set. BSS and bias ratio improved again from the 0–30 cm to the 0–100 cm depth. In 0–30 cm depth *presence of waterlogged horizons* were clearly and down to 50 cm slightly over-predicted while down to 100 cm there was no bias. Performance evaluated by percentage correct with the Peirce skill score (PSS) was similar for all three depths (correct predictions being 44 to 50 % more frequent compared to random predictions). Ignoring correct *absence* predictions in Gilbert skill score (GSS), the model predicted the correct level 20–30 % more often than a random prediction scheme. Again, GSS increased with depth and larger chance of waterlogging occurring.

4.2.3   Mapping of presence of waterlogged horizons

Presence of waterlogged horizons in 0–30 cm was predicted for 13.8 % of the area of study region Greifensee (Fig. 6). For 0–50 cm this share increased to 27.3 % and in nearly 40 % of the soils waterlogged horizons were present in 0–100 cm. Waterlogged horizons were mapped in upper soil depths mainly on the larger plains to the North and South of Lake Greifensee. Deeper horizons had waterlogging present mostly in local depressions and comparably smaller valley bottoms in the hilly uplands to the South of the study region.

4.3   Drainge classes – case study 3

4.3.1   Model for drainage classes

The models for the ordinal drainage class data contained about the same covariates as the models for *presence of waterlogged horizons* (Table 2). Most covariates had only very weak nonlinear effects (Fig. S7 in Supplement). Residual spatial autocorrelation was very weak with a short range (Fig. S2 in Supplement) suggesting that the variation was well captured by the geoGAM. 10-fold cross-validation resulted in a ranked probability skill score (RPSS) of 0.588.

4.3.2   Validation of predicted drainage classes with independent data

Table 5 reports the number of correctly classified and misclassified drainage class predictions for the validation set. False predictions were equally distributed above and below
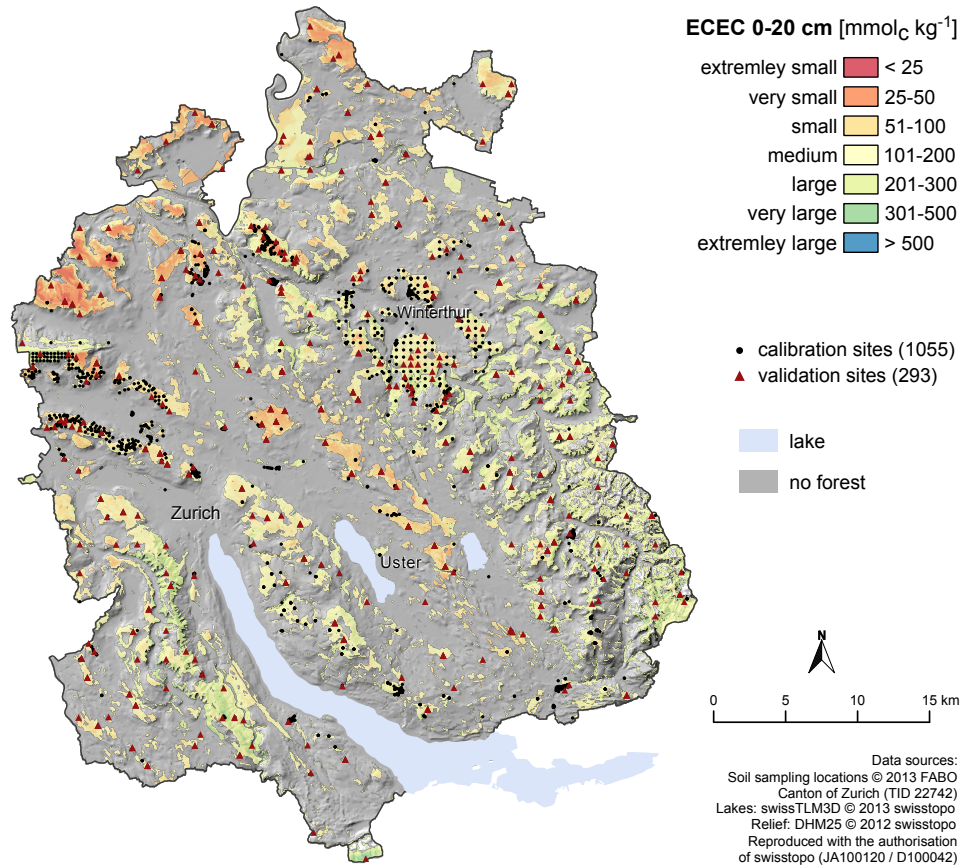
Figure 5. geoGAM predictions of effective cation exchange capacity (ECEC) in 0–20 cm depth of the mineral soil of forests in the Canton of Zurich, Switzerland (computed on a 20 m-grid with final geoGAM with covariates according to Table 2. Black dots are locations used for geoGAM calibration, locations with red triangles were used for model validation, ECEC legend classes according to Walthert et al., 2004).

Table 4. Observed occurrence of waterlogged horizons at three soil depths against predictions by geoGAM for the 198 sites of the validation set. Waterlogged soil horizons were predicted to be present if prediction probabilities were larger than an optimal threshold (30 cm: 0.22, 50 cm: 0.35, 100 cm: 0.51) found by cross-validation with GSS as criteria (#: number of sites per response level, BSS: Brier skill score, bias: bias ratio, PSS: Peirce skill score, GSS: Gilbert Skill score).

waterlogged down to	# predicted	# observed present	absent	BSS	bias	PSS	GSS
30 cm	present	16	27	0.312	1.720	0.484	0.227
	absent	9	146				
50 cm	present	28	25	0.448	1.152	0.444	0.267
	absent	18	127				
100 cm	present	43	22	0.526	1.000	0.496	0.330
	absent	22	111				

the diagonal, hence predictions were unbiased with a mean bias ratio close to 1. Distinguishing *moderately well drained* soils from the other two classes remained difficult as this class had been seldom observed. Overall, the model accuracy was satisfactory, with RPSS of 0.458 being only slightly smaller than cross-validation RPSS. Hence, the geoGAM was clearly better than predicting always the most abundant class *well drained*. Measured by PSS and Gerrity score (GS), the geoGAM was better than random predictions at every second site, for which predictions were computed.

Table 5. Frequency of drainage class levels and predictions of respective outcomes by geoGAM for the 198 sites of the validation set (#: number of sites per response level, RPSS: ranked probability skill score, bias: mean bias ratio, PSS: Peirce skill score, GS: Gerrity score for ordered responses).

# predicted	# observed			RPSS	bias	PSS	GS
	well drained	moderately well drained	poorly drained				
well drained	129	9	9	0.458	0.985	0.477	0.523
moderately well drained	9	9	3				
poorly drained	8	5	17				

4.3.3 Mapping of drainage classes

Drainage classes were again predicted using a 20 m-grid (Fig. 7). 73.2 % of the area of the Greifensee region had *well drained* soils. *Poorly drained* soils were predicted for only 15.6 % of the area. The location of *poorly drained* soils coincides with *presence of waterlogged horizons* in the topsoil (0–30 cm, panel [a] in Fig. 6). The largest contiguous area of *poorly drained* soils was predicted on accumulation plains at the lake inflow to the South of Lake Greifensee. The sites misclassified had TPI values indicating local depressions and had larger erosion accumulation potential (MRVBF) compared to correctly classified sites, thus predicting correct drainage classes in valley bottoms seems more difficult. Misclassified sites of the validation set had on average slightly larger clay and soil organic carbon contents in topsoil.

5 Discussion

5.1 Model building and covariate selection

The model building procedure efficiently selected for all responses parsimonious models with $p \leq 17$ covariates. This corresponds to only 5.8 % of the covariates considered for the effective cation exchange capacity (ECEC) modelling and to 1.4–2.8 % for modelling the binary and ordinal responses describing waterlogging.

The procedure was able to select meaningful covariates, which reveal the influence of soil forming factors on the response variable, without any prior knowledge about the importance of a particular covariate. No pre-processing of covariates was necessary, e.g. such as reducing the dimensionality of the covariate set to deal with multi-collinearity. Especially for terrain covariates this is important. Elevation data are often available in several resolutions, and various algorithms can be used to calculate e.g. curvature or topographic wetness indices (TWI), which all likely produce slightly different results. In addition, radii for computing e.g. topographic position indices (TPI) have to be specified, and it is often not a priori clear how these should be chosen (Behrens et al., 2010; Miller et al., 2015). Therefore, different algorithms and a range of parameter values are used to create terrain covariates and the model building process selects the most suit-

able covariate to model a particular soil property. Meanwhile, none of the 180 APEX bands available for the Greifensee region was chosen for the final models. Most likely, meaningful preprocessing – e.g. based on bare soil areas – could improve the usefulness of such covariates (Diek et al., 2016). Since we used continuous reflectance signals, including vegetated and sparsely vegetated areas, the remotely sensed signal might not have expressed too well direct relationships to actual soil properties.

5.2 Model structure

Parsimonious models lend themselves to a verification of fitted effects from a pedological perspective. Yet, due to multi-collinearities in the covariate set, effects of selected covariates could be substituted by effects of other covariates (Behrens et al., 2014).

Although Johnson et al. (2000) did not find strong relationships between terrain and ECEC, six terrain attributes were selected. Covariates representing geology were important, too, with e.g. ECEC changing as a function of the distance to two types of moraines. Also, vegetation provided information on ECEC in the topsoil because a vegetation index (difference of near infrared to red reflectance) and a vegetation map were included. Larger values of ECEC were modelled for plant communities that are characteristic for nutrient-rich soils. The factor distinguishing the origin of soil data either from direct measurement or pedotransfer function (PTF, legacy data correction, Sect. 3.2.2, Fig. S1 in Supplement) was further relevant in the ECEC model.

For modelling drainage classes and *presence of waterlogged horizons* plausible covariates were selected (Figs. S4 to S7 in Supplement). Most covariates were terrain attributes derived from the digital elevation model (DEM). This is in accordance with Campling et al. (2002) who found topography important in general and Lemercier et al. (2012) who showed that a topographic wetness index was among the most important covariates. Local depression at various scales (concave curvature, basins in TPI, sites with accumulation by erosion, terrain wetness) increased the probability for *poorly drained* soils and *presence of waterlogged horizons*. More variable terrain (standard deviation of elevation) also increased waterlogging probability. Climate covariates also seemed to be
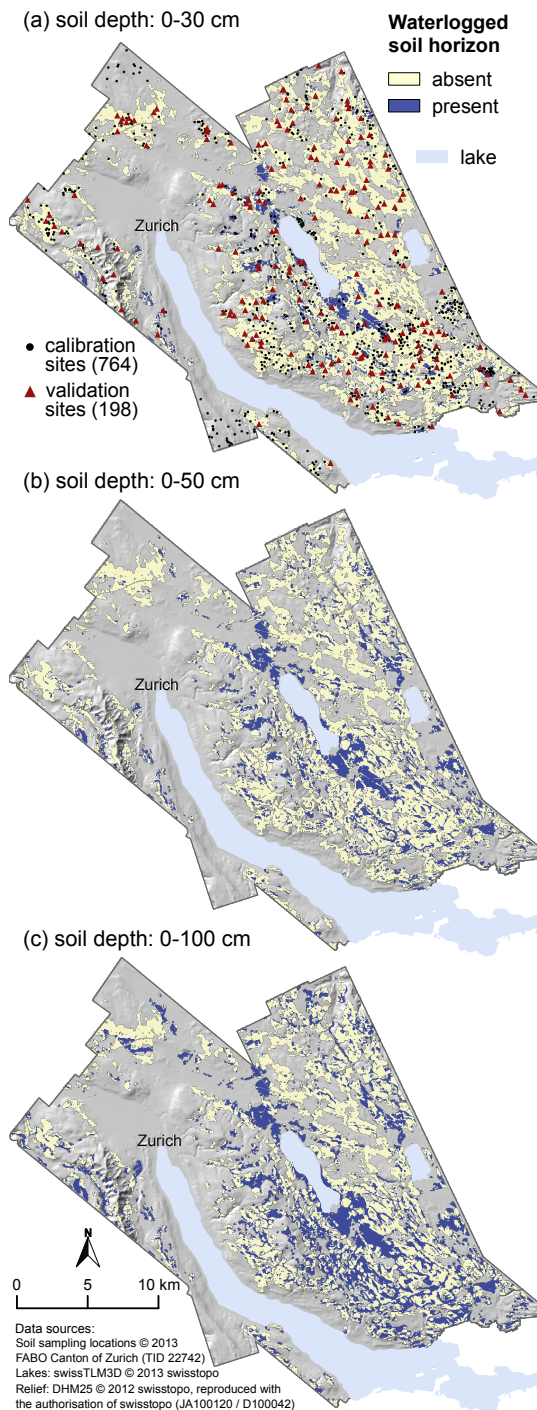
Figure 6. geoGAM predictions of presence of waterlogged horizons between surface and 3 soil depths (a: 0–30, b: 0–50, c: 0–100 cm) for the agricultural land in the Greifensee study region (computed on a 20 m-grid with final geoGAM with covariates according to Table 2, smoothed for better display with focal mean with radius of 3 pixels = 60 m). Black dots in panel (a) are locations used for geoGAM calibration, locations with red triangles were used for model validation.
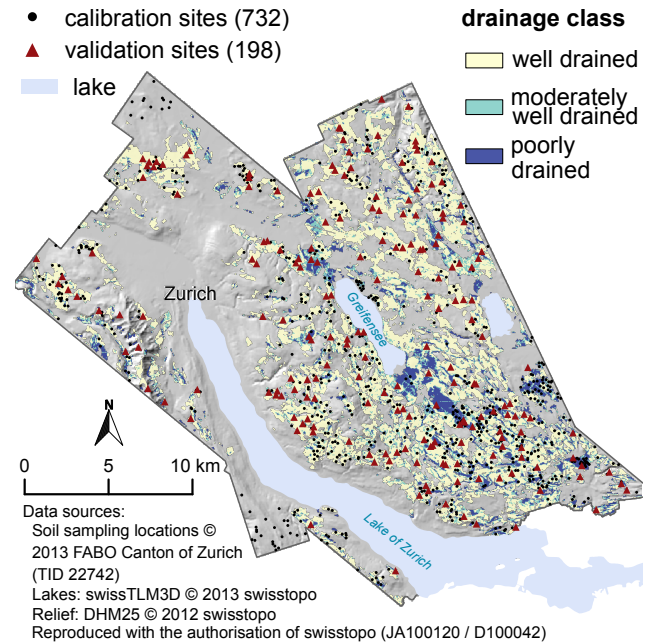
Figure 7. geoGAM predictions of drainage classes for the agricultural land in the Greifensee study region (computed on a 20 m-grid with final geoGAM with covariates according to Table 2, smoothed for better display with focal mean with radius of 3 pixels = 60 m). Black dots are locations used for geoGAM calibration, locations with red triangles were used for model validation.

important. Rainfall pattern in summer (June, July), spring dew point temperature and global radiation (March, April) correlated most strongly with *presence of waterlogged horizons*. Information on human activities related to waterlogged soil amelioration were included in all four models. Maps of historic wetlands and areas with drainage systems were most often chosen in combination. Geology was also partly relevant (*presence of waterlogged horizons* in 0–100 cm soil depth and drainage classes).

Overall, nonlinearities in effects were small for drainage classes and *presence of waterlogged horizons*. Estimated degrees of freedom (EDF, Wood, 2006, pp. 170) were generally smaller than 1.5, with some continuous effects even being close to 1 EDF. In contrast, most nonlinear effects of the model for ECEC had EDF around 1.7–1.8 with northness consuming even 2.0 EDF. The large area of the study region and the response being a chemical property that depends on various combinations of soil forming factors evidently required the use of a more complex model.

5.3 Predictive performance of fitted models

For the final models, cross-validation statistics were similar to results obtained for the independent validation data. Through repeated cross-validation on the same subsets the

cross-validation statistics can be considered as conservative *goodness-of-fit* statistics. Hence, we conclude that geoGAM did not over-fit the calibration data.

Independently validated model accuracy was satisfactory for ECEC in the present study with (SS_{mse} 0.37). Compared to the few available studies the quality of our maps of ECEC was intermediate. Building a separate model for forest soil ECEC for a dataset with about 2.1 sites per km^2 seem to produce much better results than the study reported by Vaysse and Lagacherie (2015) who found very poor model performance for ECEC ($R^2 = 0$, equivalently computed as SS_{mse}) for a dataset with 0.04 sites per km^2 and a study region with multiple land uses. Mulder et al. (2016) achieved somewhat better results ($R^2 = 0.24$, details on computation not given) for mapping topsoil ECEC for whole France. Hengl et al. (2017) mapped ECEC with a large dataset globally and obtained a cross-validation R^2 of 0.65 (computed as SS_{mse}). Viscarra Rossel et al. (2015, Supplement) reported R^2 of 0.79 (computed as SS_{mse}) for topsoil ECEC for Australia. In Hengl et al.'s and Viscarra Rossel et al.'s studies ECEC varied much more than in our study, and this likely explains the better quality of the predictions.

Our models for presence of wet soils reached similar accuracy as reported in other studies. Zhao et al. (2013, Table 1) reported that 64 to 87 % of the sites were correctly classified (percentage correct, PC) in four studies that modelled three drainage class levels. Three studies with up to seven drainage levels achieved PC of 52 to 78%, and Zhao et al. (2013) themselves had 36 % of correctly classified sites. Kidd et al. (2014) found PC of 53 % and 55 % for two study regions, and Lemercier et al. (2012) reported PC of 52 % for a four-level drainage response. The presented models (Table 4 and 5) are about as good with PC of 78 % to 82 % for predicting *presence of waterlogged horizons* and PC of 78 % for predicting the three drainage class levels.

Nevertheless, PC is trivial to *hedge* (Jolliffe and Stephenson, 2012, pp. 46), and comparisons should be made only with care. Better performance measures are PSS and Cohen's kappa (κ), also called Heidke skill score (Wilks, 2011, pp. 347). Campling et al. (2002) reported a κ of 0.705, Kidd et al. (2014) κ's of 0.27 and 0.31 for the two study regions, Lemercier et al. (2012) a κ of 0.27 and Peng et al. (2003) found κ of 0.59 for predictions of three drainage levels. κ's computed for the models of this study ranged between 0.37 and 0.5 for modelling the *presence of waterlogged horizons* and was 0.48 for predicting the three levels of drainage class. Unequal distribution of the three drainage classes in the study region (majority of soils were *well drained*) were reflected in the smaller value of κ compared to PC.

5.4 Spatial structure of predicted maps

The spatial distribution of ECEC as shown by Fig. 5 aligns well with pedological knowledge about soils in the Canton of Zurich. The smallest ECEC ($< 50 \, mmol_c \, kg^{-1}$) was mapped

in the northeast of the study region. The last glaciation (Swisstopo, 2009) did not reach as far north and, as a consequence, strongly weathered soils on old fluvioglacial gravel-rich sediments developed in this part of the study region. Soils not covered by ice during the last glaciation have comparably larger ECEC if they formed on Molasse.

As expected the spatial patterns for the *presence of waterlogged soil horizons* and the drainage classes were very similar (Fig. 6 and 7). Especially soils on plains to the North and South of Lake Greifensee are often *poorly drained*, although at many locations agricultural drainage networks were installed in the past.

6   Summary and conclusion

Effectively building predictive models for digital soil mapping (DSM) becomes crucial if many soil properties are to be mapped. Selecting only a small set of relevant covariates renders interpretation of the fitted models easier and allows to check whether modelled relations accord with pedological understanding. Parsimonious, interpretable DSM models are likely more readily accepted by end-users than complex black-box models. Moreover, model selection out of a large number of covariates describing soil forming factors helps to improve knowledge about relationships at larger scales. In this sense, it is also important, that the modelling approach provides information about covariates which are not relevant for a certain response, e.g. the large number of APEX bands for *persence of waterlogged horizons* and drainage classes.

We developed a model building framework for generalized additive models for spatial data (geoGAM) and applied the framework to legacy soil data from the Canton of Zurich (Switzerland). We found that geoGAM

- consistently modelled continuous, binary and ordinal responses, hence, allow DSM of measured soil properties and soil classification data using one common approach,

- selected, given the large numbers of covariates, adequately small sets of pedogenetically meaningful covariates without any prior knowledge about their importance and without prior reduction of the covariate sets,

- required minimal user interaction for model building, which facilitates future map updates as new soil data or new covariates become available,

- allowed easy interpretation of effects of the included covariates by partial residual plots,

- modelled predictive distributions for continuous responses by a bootstrapping approach, thereby taking uncertainty of model building into account,

- did not over-fit the calibration data in our applications, and

– predicted soil properties with similar accuracy as other approaches did in other digital soil mapping studies, when tested with an independent validation set.

To further assess usefulness of geoGAM for DSM future work should focus on comparisons of predictive accuracy with commonly used statistical methods (e.g. geostatistics or tree-based machine learning techniques) on the same soil datasets. Nussbaum et al. (2017) published a first such study.

Code availability. The geoGAM model building procedure was published as R package geoGAM (Nussbaum, 2017).

Data availability. The soil data were used under a non-public data licence (Canton of Zurich, contract number TID 22742; WSL) and could not be published.

Author contributions. A. Papritz proposed the application of componentwise gradient boosting with smooth baselearners for DSM. M. Nussbaum implemented the framework and adapted it to the needs of DSM. L. Walthert harmonized the soil data with collaborators, and M. Fraefel computed multi-scale terrain attributes. L. Greiner defined the responses *presence of waterlogged soil horizons* and *drainage classes* from Swiss soil classification data. M. Nussbaum prepared the manuscript with major input from A. Papritz and further contributions from all co-authors.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We thank the Swiss National Science Foundation SNSF for funding this work in the frame of the National Research Program "Sustainable Use of Soil as a Resource" (NRP 68)" and "Swiss Earth Observatory Network" (SEON) for funding aerial surveys with APEX. Special thanks go to WSL and the soil protection agency of the Canton of Zurich for sharing their soil data with us. Furthermore, we would like to thank Thorsten Hothorn for advice on model selection and boosting.

References

Adhikari, K., Kheir, R., Greve, M., Bøcher, P., Malone, B., Minasny, B., McBratney, A., and Greve, M.: High-resolution 3-D mapping of soil texture in Denmark, Soil Sci Soc Am J, 77, 860–876, doi:10.2136/sssaj2012.0275, 2013.

ALN: Historische Feuchtgebiete der Wildkarte 1850. Amt für Landschaft und Natur des Kantons Zürich, http://www.aln.zh.ch/internet/baudirektion/aln/de/naturschutz/naturschutzdaten/geodaten.html, last access 29.03.2017, 2002.

ALN: Geologische Karte des Kantons Zürich nach Hantke et. al 1967, GIS-ZH Nr. 41. Amt für Landschaft und Natur des Kantons Zürich, http://www.gis.zh.ch/Dokus/Geolion/gds_41.pdf, last access: 15.02.2015, 2014a.

ALN: Meliorationskataster des Kantons Zürich, GIS-ZH Nr. 148. Amt für Landschaft und Natur des Kantons Zürich., http://www.geolion.zh.ch/geodatensatz/show?nbid=387, last access 29.03.2017, 2014b.

AWEL: Hinweisflächen für anthropogene Böden, GIS-ZH Nr. 260. Amt für Abfall, Wasser, Energie und Luft des Kanton Zürich, http://www.geolion.zh.ch/geodatensatz/show?nbid=985, last access 29.03.2017, 2012.

AWEL: Grundwasservorkommen, GIS-ZH Nr. 327. Amt für Abfall, Wasser, Energie und Luft des Kanton Zürich, http://www.geolion.zh.ch/geodatensatz/show?nbid=723, last access 29.03.2017, 2014.

AWEL: NO₂-Immissionen, GIS-ZH Nr. 82, Amt für Abfall, Wasser, Energie und Luft des Kanton Zürich, http://geolion.zh.ch/geodatensatz/show?nbid=783, last access 29.03.2017, 2015.

BAFU: Luftbelastung: Karten Jahreswerte, Ammoniak und Stickstoffdeposition, Jahresmittel 2007 (modelliert durch METEOTEST), http://www.bafu.admin.ch/luft/luftbelas-tung/schadstoffkarten, last access 15.02.2015, 2011.

Behrens, T., Schmidt, K., Zhu, A. X., and Scholten, T.: The ConMap approach for terrain-based digital soil mapping, Eur J Soil Sci, 61, 133–143, doi:10.1111/j.1365-2389.2009.01205.x, 2010.

Behrens, T., Schmidt, K., Ramirez-Lopez, L., Gallant, J., Zhu, A.-X., and Scholten, T.: Hyper-scale digital soil mapping and soil formation analysis, Geoderma, 213, 578–588, doi:10.1016/j.geoderma.2013.07.031, 2014.

Ben-Dor, E., Chabrillat, S., Demattê, J. A. M., Taylor, G. R., Hill, J., Whiting, M. L., and Sommer, S.: Using imaging spectroscopy to study soil properties, Remote Sens Environ, 113, S38–S55, doi:10.1016/j.rse.2008.09.019, 2009.

BFS: GEOSTAT Benützerhandbuch, Bundesamt für Statistik, Bern, 2001.

Bourennane, H., King, D., Chéry, P., and Bruand, A.: Improving the kriging of a soil variable using slope gradient as external drift, Eur J Soil Sci, 47, 473–483, doi:10.1111/j.1365-2389.1996.tb01847.x, 1996.

Brassel, P. and Lischke, H., eds.: Swiss National Forest Inventory: Methods and models of the second assessment, Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Birmensdorf, 2001.

Breheny, P. and Huang, J.: Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors, Stat Comput, 25, 173–187, doi:10.1007/s11222-013-9424-2, 2015.

Brunner, J., Jäggli, F., Nievergelt, J., and Peyer, K.: Kartieren und Beurteilen von Landwirtschaftsböden, FAL Schriftenreihe 24, Eidgenössische Forschungsanstalt für Agrarökologie und Landbau, Zürich-Reckenholz (FAL), 1997.

Buchanan, S., Triantafilis, J., Odeh, I. O. A., and Subansinghe, R.: Digital soil mapping of compositional particle-size fractions using proximal and remotely sensed ancillary data, Geophysics, 77, WB201–WB211, doi:10.1190/geo2012-0053.1, 2012.

Bühlmann, P. and Hothorn, T.: Boosting algorithms: Regularization, prediction and model fitting, Stat Sci, 22, 477–505, doi:10.1214/07-sts242, 2007.

Campling, P., Gobin, A., and Feyen, J.: Logistic modeling to spatially predict the probability of soil drainage classes, Soil Sci Soc Am J, 66, 1390–1401, doi:10.2136/sssaj2002.1390, 2002.

Cleveland, W. S.: Robust Locally Weighted Regression and Smoothing Scatterplots, J Am Stat Assoc, 74, 829–836, doi:10.2307/2286407, 1979.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., and Böhner, J.: System for automated geoscientific analyses (SAGA) v. 2.1.4, Geosci Model Dev, 8, 1991–2007, doi:10.5194/gmd-8-1991-2015, http://www.geosci-model-dev.net/8/1991/2015/, 2015.

Cressie, N.: Block Kriging for Lognormal Spatial Processes, Math Geol, 38, 413–443, doi:10.1007/s11004-005-9022-8, 2006.

Danner, C., Hensold, C., Blum, P., Weidenhammer, S., Aussendorf, M., Kraft, M., Weidenbacher, A., Holleis, P., and Kölling, C.: Das Schutzgut Boden in der Planung, Bewertung natürlicher Bodenfunktionen und Umsetzung in Planungs- und Genehmigungsverfahren, Bayerisches Landesamt für Umweltschutz, Bayerisches Geologisches Landesamt, http://www.lfu.bayern.de/boden/bodenfunktionen/ertragsfaehigkeit/doc/arbeitshilfe_boden.pdf, last access 29.03.2017, 2003.

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

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

Diek, S., Schaepman, M., and de Jong, R.: Creating multi-temporal composites of airborne imaging spectroscopy data in support of digital soil mapping, Remote Sensing, 8, doi:10.3390/rs8110906, 2016.

Dirichlet, G. L.: Über die Reduction der positiven quadratischen Formen mit drei unbestimmten ganzen Zahlen, Journal für die reine und angewandte Mathematik, 40, 209–227, doi:10.1017/cbo9781139237345.005, http://eudml.org/doc/147457, 1850.

DMC: Disaster Monitoring Constellation International Imaging, http://www.dmcii.com, last access: 03.02.2015, 2015.

Dobson, A. J.: An Introduction to GeneralIzed Linear Models, Chapman & Hall/CRC, Boca Raton, 2002.

ELF: Schweizerische Referenzmethoden der Forschungsanstalten Agroscope – Boden- und Substratuntersuchungen zur Düngeberatung, Loseblattordner E1.011.d 1, Forschungsanstalten Agroscope ART und ACW, Zürich und Changins, Ausgabe 1996 mit Änderungen von 1997 bis 2009, Version 2015, Methode "AAE-10", 1996.

ESRI: ArcGIS Desktop: Release 10, ESRI Environmental Systems Research Institute, Redlands, California, USA., www.esri.com, last access 29.03.2017, 2010.

FAC: Methoden für Bodenuntersuchungen, no. 5 in Schriftenreihe der FAC, Liebefeld, Eidgenössische Forschungsanstalt für Agrikulturchemie und Umwelthygiene (FAC), 1989.

FAO and ITPS: Status of the World's Soil Resources (SWSR), Main report, Food and Agriculture Organization of the United Nations and Intergovernmental Technical Panel on Soils, Rome, Italy, 2015.

Faraway, J. J.: Linear Models with R, vol. 63 of *Texts in Statistical Science*, Chapman & Hall/CRC, Boca Raton, 2005.

Fitzpatrick, B. R., Lamb, D. W., and Mengersen, K.: Ultrahigh Dimensional Variable Selection for Interpolation of Point Referenced Spatial Data: A Digital Soil Mapping Case Study, PLoS One, 11, 1–19, doi:10.1371/journal.pone.0162489, 2016.

Frigge, M., Hoaglin, D. C., and Iglewicz, B.: Some implementations of the boxplot, The American Statistician, 43, 50–54, doi:10.2307/2685173, 1989.

FSO: Swiss soil suitability map. BFS GEOSTAT. Swiss Federal Statistical Office, http://www.bfs.admin.ch/bfs/portal/de/index/dienstleistungen/geostat/datenbeschreibung/digitale_bodeneignungskarte.html, last access 15.02.2015, 2000a.

FSO: Tree composition of Swiss forests. BFS GEOSTAT. Swiss Federal Statistical Office, http://www.bfs.admin.ch/bfs/portal/de/index/dienstleistungen/geostat/datenbeschreibung/waldmischungsgrad.html, last access 15.02.2015, 2000b.

Gasser, U., Gubler, A., Hincapié, I., Karagiannis, D.-A., Schwierz, C., and Zimmermann, S.: Bestimmung der Austauschereigenschaften von Waldböden: Kostenoptimierung, Bulletin Bodenkundliche Gesellschaft der Schweiz, 32, 51–52, 2011.

Grimm, R., Behrens, T., Märker, M., and Elsenbeer, H.: Soil organic carbon concentrations and stocks on Barro Colorado Island – Digital soil mapping using Random Forests analysis, Geoderma, 146, 102–113, doi:10.1016/j.geoderma.2008.05.008, 2008.

Hantke, R. u.: Geologische Karte des Kantons Zürich und seiner Nachbargebiete, Kommissionsverlag Leemann, Zürich, Sonderdruck aus *Vierteljahrsschrift der Naturforschenden Gesellschaft in Zürich*, 112(2): 91–122, 1967.

Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning; Data Mining, Inference and Prediction, Springer, New York, 2 edn., 2009.

Hastie, T. J. and Tibshirani, R. J.: Generalized Additive Models, vol. 43 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London, 1990.

Henderson, B. L., Bui, E. N., Moran, C. J., and Simon, D. A. P.: Australia-wide predictions of soil properties using decision trees, Geoderma, 124, 383–398, doi:10.1016/j.geoderma.2004.06.007, 2005.

Hengl, T., de Jesus, J. M., MacMillan, R. A., Batjes, N. H., Heuvelink, G. B. M., Ribeiro, E., and Samuel-Rosa, A.: SoilGrids1km – Global Soil Information Based on Automated Mapping, PLoS One, 9, doi:10.1371/journal.pone.0105992, 2014.

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., and Kempen, B.: SoilGrids250m: Global gridded soil information based on machine learning, PLoS One, 12, 1–40, doi:10.1371/journal.pone.0169748, 2017.

Heung, B., Ho, H. C., Zhang, J., Knudby, A., Bulmer, C. E., and Schmidt, M. G.: An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping, Geoderma, 265, 62–77, doi:http://dx.doi.org/10.1016/j.geoderma.2015.11.014, 2016.

Hijmans, R. J., van Etten, J., Cheng, J., Mattiuzzi, M., Sumner, M., Greenberg, J. A., Lamigueiro, O. P., Bevan, A., Racine, E. B., and Shortridge, A.: raster: Geographic Data Analysis and Modeling, R package versoin 2.4-15, http://CRAN.R-project.org/package=raster, last access 29.03.2017, 2015.

Hofner, B., Hothorn, T., Kneib, T., and Schmid, M.: A Framework for Unbiased Model Selection Based on Boosting, J Comput Graph Stat, 20, 956–971, doi:10.1198/jcgs.2011.09220, 2011.

Hofner, B., Mayr, A., Robinzonov, N., and Schmid, M.: Model-based boosting in R: A hands-on tutorial using the R package mboost, Computation Stat, 29, 3–35, doi:10.1007/s00180-012-0382-5, 2014.

Hothorn, T., Müller, J., Schröder, B., Kneib, T., and Brandl, R.: Decomposing environmental, spatial, and spatiotemporal components of species distributions, Ecol Monogr, 81, 329–347, 2011.

Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M., and Hofner, B.: mboost: Model-Based Boosting, http://CRAN.R-project.org/package=mboost, R package version R package version 2.4-2, last access 29.03.2017, 2015.

Hotz, M.-C., Weibel, F., Ringgenberg, B., Beyeler, A., Finger, A., Humbel, R., and Sager, J.: Arealstatistik Schweiz Zahlen – Fakten – Analysen, Bericht, Bundesamt für Statistik (BFS), Neuchâtel, 2005.

Jäggli, F., Peyer, K., Pazeller, A., and Schwab, P.: Grundlagenbericht zur Bodenkartierung des Kantons Zürich, Tech. rep., Volkswirtschaftsdirektion des Kantons Zürich und Eidg. Forschungsanstalt für Agrarökologie und Landbau Zürich Reckenholz FAL, 1998.

Johnson, C. E., Ruiz-Méndez, J. J., and Lawrence, G. B.: Forest soil chemistry and terrain attributes in a Catskills watershed, Soil Sci Soc Am J, 64, 1804–1814, doi:10.2136/sssaj2000.6451804x, 2000.

Jolliffe, I. T. and Stephenson, D. B., eds.: Forecast verification: A practitioner's guide in atmospheric science, Wiley-Blackwell, Chichester, 2 edn., 2012.

Kammann, E. E. and Wand, M. P.: Geoadditive models, J Roy Stat Soc C-App, 52, 1–18, doi:10.1111/1467-9876.00385, 2003.

Kidd, D. B., Malone, B. P., McBratney, A. B., Minasny, B., and Webb, M. A.: Digital mapping of a soil drainage index for irrigated enterprise suitability in Tasmania, Australia, Soil Res, 52, 107–119, doi:10.1071/SR13100, 2014.

Kneib, T., Hothorn, T., and Tutz, G.: Variable selection and model choice in geoadditive regression models, Biometrics, 65, 626–634, doi:10.1111/j.1541-0420.2008.01112.x, 2009.

Kreuzwieser, J. and Rennberg, H.: Molecular and physiological responses of trees to waterlogging stress, Plant Cell Environ, 37, 2245–2259, doi:10.1111/pce.12310, 2014.

Lacoste, M., Mulder, V., de Forges, A. R., Martin, M., and Arrouays, D.: Evaluating large-extent spatial modeling approaches: A case study for soil depth for France, Geoderma Regional, 7, 137–152, doi:10.1016/j.geodrs.2016.02.006, 2016.

Lakanen, E. and Erviö, R.: A comparison of eight extractants for the determination of plant available micronutrients in soils, Acta Agralia Fennica, 123, 223–232, 1971.

Lemercier, B., Lacoste, M., Loum, M., and Walter, C.: Extrapolation at regional scale of local soil knowledge using boosted classification trees: A two-step approach, Geoderma, 171–172, 75–84, doi:10.1016/j.geoderma.2011.03.010, 2012.

Liddicoat, C., Maschmedt, D., Clifford, D., Searle, R., Herrmann, T., Macdonald, L., and Baldock, J.: Predictive mapping of soil organic carbon stocks in South Australia's agricultural zone, Soil Res, 53, 956–973, doi:10.1071/SR15100, 2015.

Liess, M., Glaser, B., and Huwe, B.: Uncertainty in the spatial prediction of soil texture. Comparison of regression

tree and Random Forest models, Geoderma, 170, 70–79, doi:http://dx.doi.org/10.1016/j.geoderma.2011.10.010, 2012.

Litz, N.: Schutz vor Organika, in: Handbuch der Bodenkunde, edited by Blume, H.-P., vol. 5, chap. 7.6.6, p. 28, Wiley-VCH, Landsberg, 1998.

Martin, M. P., Wattenbach, M., Smith, P., Meersmans, J., Jolivet, C., Boulonne, L., and Arrouays, D.: Spatial distribution of soil organic carbon stocks in France, Biogeosciences, 8, 1053–1065, doi:10.5194/bg-8-1053-2011, 2011.

Mathys, L. and Kellenberger, T.: Spot5 RadcorMosaic of Switzerland, Tech. rep., National Point of Contact for Satellite Images NPOC: Swisstopo; Remote Sensing Laboratories, University of Zurich, Zurich, 2009.

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

Meersmans, J., De Ridder, F., Canters, F., De Baets, S., and Van Molle, M.: A multiple regression approach to assess the spatial distribution of Soil Organic Carbon (SOC) at the regional scale (Flanders, Belgium), Geoderma, 143, 1–13, doi:10.1016/j.geoderma.2007.08.025, 2008.

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

Miller, B. A., Koszinski, S., Wehrhan, M., and Sommer, M.: Impact of multi-scale predictor selection for modeling soil properties, Geoderma, 239–240, 97–106, doi:10.1016/j.geoderma.2014.09.018, 2015.

Moran, C. J. and Bui, E. N.: Spatial data mining for enhanced soil map modelling, Int J Geogr Inf Sci, 16, 533–549, doi:10.1080/13658810210138715, 2002.

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

Mulder, V. L., de Bruin, S., Schaepman, M. E., and Mayr, T. R.: The use of remote sensing in soil and terrain mapping – A review, Geoderma, 162, 1–19, doi:10.1016/j.geoderma.2010.12.018, 2011.

Müller, L., Schinder, U .and Behrendt, A., Eulenstein, F., and Dannowski, R.: The Muencheberg Soil Quality Rating (SQR): Field manual for detecting and assessing properties and limitations of soils for cropping and grazing, Report, Leibniz-Zentrum für Agrarlandschaftsforschung (ZALF), Müncheberg, Germany, 2007.

Nussbaum, M.: geoGAM: Select Sparse Geoadditive Models for Spatial Prediction, https://CRAN.R-project.org/package=geoGAM, R package version 0.1-2, last access 29.03.2017, 2017.

Nussbaum, M. and Papritz, A.: Transferfunktionen Nährstoffmesswerte, Bericht, ETH Zürich, Soil and Terrestrial Environmental Physics, doi:http://dx.doi.org/10.3929/ethz-a-010810702, Version 2, mit kl. Änderung 27. Nov. 2016, 2015.

Nussbaum, M., Papritz, A., Baltensweiler, A., and Walthert, L.: Estimating soil organic carbon stocks of Swiss forest soils by robust external-drift kriging, Geosci Model Dev, 7, 1197–1210, doi:10.5194/gmd-7-1197-2014, http://www.geosci-model-dev.net/7/1197/2014/, 2014.

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M. E., and Papritz, A.: Evaluation of digital soil mapping approaches with large sets of environmental

covariates, Soil Discuss, 2017, 1–32, doi:10.5194/soil-2017-14, https://www.soil-discuss.net/soil-2017-14/, 2017.

Omuto, C. and Nachtergaele, F .and Vargas Rojas, R.: State of the Art Report on Global and Regional Soil Information : Where are we? Where to go?, Tech. rep., Food and Agriculture Organization of the United Nations, Rome, 2013.

Peng, W., Wheeler, D., Bell, J., and Krusemark, M.: Delineating patterns of soil drainage class on bare soils using remote sensing analyses, Geoderma, 115, 261–279, doi:10.1016/S0016-7061(03)00066-1, 2003.

Poggio, L. and Gimona, A.: National scale 3D modelling of soil organic carbon stocks with uncertainty propagation – An example from Scotland, Geoderma, 232–234, 284–299, doi:10.1016/j.geoderma.2014.05.004, 2014.

Poggio, L., Gimona, A., and Brewer, M.: Regional scale mapping of soil properties and their uncertainty with a large number of satellite-derived covariates, Geoderma, 209–210, 1–14, doi:10.1016/j.geoderma.2013.05.029, 2013.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/, last access: 29.03.2017, 2016.

Remund, J., Frehner, M., Walthert, L., Kägi, M., and Rihm, B.: Schätzung standortspezifischer Trockenstressrisiken in Schweizer Wäldern, 2011.

Schaepman, M., Jehle, M., Hueni, A., D'Odorico, P., Damm, A., Weyermann, J., Schneider, F., Laurent, V., Popp, C., Seidel, F., Lenhard, K., Gege, P., Küchler, C., Brazile, J., Kohler, P., Vos, L., Meuleman, K., Meynart, R., Schläpfer, D., and Itten, K.: Advanced radiometry measurements and Earth science applications with the Airborne Prism Experiment (APEX), Remote Sens Environ, 158, 207–219, doi:10.1016/j.rse.2014.11.014, 2015.

Schmid, M., Hothorn, T., Maloney, K. O., Weller, D. E., and Potapov, S.: Geoadditive regression modeling of stream biological condition, Environ Ecol Stat, 18, 709–733, doi:10.1007/s10651-010-0158-4, 2011.

Schmider, P., Küper, M., Tschander, B., and Käser, B.: Die Waldstandorte im Kanton Zürich Waldgesellschaften, Waldbau Naturkunde, vdf Verlag der Fachvereine an den schweizerischen Hochschulen und Techniken, Zürich, 1993.

Scull, P., Franklin, J., Chadwick, O. A., and McArthur, D.: Predictive Soil Mapping: A review, Prog Phys Geogr, 27, 171–197, doi:10.1191/0309133303pp366ra, 2003.

Sindayihebura, A., Ottoy, S., Dondeyne, S., Meirvenne, M. V., and Orshoven, J. V.: Comparing digital soil mapping techniques for organic carbon and clay content: Case study in Burundi's central plateaus, CATENA, 156, 161–175, doi:http://dx.doi.org/10.1016/j.catena.2017.04.003, 2017.

Swisstopo: Switzerland during the Last Glacial Maximum 1:500 000, http://www.swisstopo.admin.ch/internet/swisstopo/en/home/products/maps/geology/geomaps/LGM-map500.html, last access: 07.06.2016, 2009.

Swisstopo: Höhenmodelle, http://www.swisstopo.admin.ch/internet/swisstopo/de/home/products/height.html, last access: 07.06.2016, 2011.

Swisstopo: swissTLM3D: Topographic Landscape Model 3D. Version 1.1, http://www.swisstopo.admin.ch/internet/swisstopo/de/home/products/landscape/swissTLM3D.html, last access: 08.03.2016, 2013a.

Swisstopo: swissAlti3D. Das hoch aufgelöste Terrainmodell der Schweiz, http://www.swisstopo.admin.ch/internet/swisstopo/de/home/products/height/swissALTI3D.html, last access: 07.06.2016, 2013b.

Swisstopo: swissBoundaries3D, http://www.swisstopo.admin.ch/internet/swisstopo/de/home/products/landscape/swissBOUNDARIES3D.html, last access: 08.03.2016, 2016.

Tutz: Regression for Categorical Data, Cambridge University Press, doi:10.1017/cbo9780511842061, 2012.

USGS EROS: USGS Land Remote Sensing Program, Landsat 7 Scene 01.09.2013. U.S. Geological Survey's Earth Resources Observation and Science Center, 2013.

Vaysse, K. and Lagacherie, P.: Evaluating Digital Soil Mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France), Geoderma Regional, 4, 20–30, doi:10.1016/j.geodrs.2014.11.003, 2015.

Venables, W. N. and Ripley, B. D.: Modern applied statistics with S-PLUS, Springer-Verlag, New York, 4 edn., 2002.

Viscarra Rossel, R., Webster, R., and Kidd, D.: Mapping gamma radiation and its uncertainty from weathering products in a Tasmanian landscape with a proximal sensor and random forest kriging, Earth Surf Process Landf, 39, 735–748, doi:10.1002/esp.3476, 2014.

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

Walthert, L., Zimmermann, S., Blaser, P., Luster, J., and Lüscher, P.: Waldböden der Schweiz. Band 1. Grundlagen und Region Jura, Eidg. Forschungsanstalt WSL and Hep Verlag, Birmensdorf and Bern, 2004.

Walthert, L., Pannatier, E. G., and Meier, E. S.: Shortage of nutrients and excess of toxic elements in soils limit the distribution of soil-sensitive tree species in temperate forests, For Ecol Manage, 297, 94–107, doi:10.1016/j.foreco.2013.02.008, 2013.

Walthert, L., Bridler, L., Keller, A., Lussi, M., and Grob, U.: Harmonisierung von Bodendaten im Projekt "Predictive mapping of soil properties for the evaluation of soil functions at regional scale (PMSoil)" des Nationalen Forschungsprogramms Boden (NFP 68), Bericht, Eidgenössische Forschungsanstalt WSL und Agroscope Reckenholz, Birmensdorf und Zürich, doi:10.3929/ethz-a-010801994, 2016.

Webster, R. and Lark, R.: Field Sampling for Environmental Science and Management, Environmental science/statistics, Routledge, https://books.google.ch/books?id=7Xz5u05QC4AC, 2013.

Wegelin, T.: Schadstoffbelastung des Bodens im Kanton Zürich Resultate des kantonalen Bodenrasternetzes, Bericht, Amt für Gewässerschutz und Wasserbau Fachstelle Bodenschutz, Zürich, 1989.

Wiesmeier, M., Barthold, F., Blank, B., and Kögel-Knabner, I.: Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem, Plant Soil, 340, 7–24, doi:10.1007/s11104-010-0425-z, 2011.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, Academic Press, 3 edn., 2011.

Wood, S. N.: Generalized Additive Models: An Introduction with R, Chapman and Hall/CRC, 2006.

Wood, S. N.: Fast stable restricted maximum likelihood and
    marginal likelihood estimation of semiparametric generalized
    linear models, J Roy Stat Soc B, 73, 3–36, doi:10.1111/j.1467-
    9868.2010.00749.x, 2011.
5 Wüst-Galley, C., Grünig, A., and Leifeld, J.: Locating organic soils
    for the Swiss greenhouse gas inventory, Agroscope Science 26,
    Agroscope,   Zurich,   https://www.bafu.admin.ch/dam/bafu/
    en/dokumente/klima/klima-climatereporting-referenzen-cp2/
    wuest-galley_c_gruenigaleifeldj2015.pdf.download.pdf,   last
10   access: 29.03.2017, 2015.
    Zhao, Z., Irfan, A. M., and Fan-Rui, M.: Model prediction of soil
    drainage classes over a large area using a limited number of field
    samples: A case study in the province of Nova Scotia, Canada,
    Can J Soil Sci, 93, 73–83, doi:10.4141/cjss2011-095, 2013.
15 Zimmermann, N. E. and Kienast, F.: Predictive mapping of alpine
    grasslands in Switzerland: Species versus community approach,
    J Veg Sci, 10, 469–482, doi:10.2307/3237182, 1999.