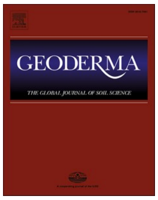
Geoderma 410 (2022) 115659

Contents lists available at ScienceDirect

Geoderma

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



Tier 4 maps of soil pH at 25 m resolution for the Netherlands

Anatol Helfenstein[a,b,*], Vera L. Mulder[a], Gerard B.M. Heuvelink[a], Joop P. Okx[b]

[a] Soil Geography and Landscape Group, Wageningen University, PO Box 47, 6700 AA Wageningen, The Netherlands
[b] Soil, Water and Land Use Team, Wageningen Environmental Research, Droevendaalsesteeg 3, 6708 RC Wageningen, The Netherlands

ARTICLE INFO

Handling Editor: Kristin Piikki

Keywords:
GlobalSoilMap
The Netherlands
Soil pH
Quantile regression forest
Statistical validation
Prediction uncertainty

ABSTRACT

Accurate and high resolution spatial soil information is essential for efficient and sustainable land use, management and conservation. Since the establishment of digital soil mapping (DSM) and the goals set by the *GlobalSoilMap* (*GSM*) working group, great advances have been made to attain spatial soil information worldwide. Highly populated areas such as the Netherlands demand multi-functional land use, for which information of key soil properties such as pH is essential to make decisions. We a) provide soil pH prediction maps at six standard depth layers between 0 m to 2 m for the Netherlands at 25 m resolution, whereby the calibrated Quantile Regression Forest (QRF) model allows for prediction at any desired depth, and b) determine map accuracy using various statistical validation strategies and evaluation of prediction uncertainty. This study is unique among *GSM* products by including design-based inference of a probability sample as an external accuracy assessment and providing Tier 4 maps with spatially explicit accuracy thresholds for end-users based on *GSM* specifications.

QRF models were tuned and calibrated using 15 338 soil observations between 0 m and 2 m depth from 4230 locations and 195 covariates representing the soil-forming factors. The following statistical validation strategies were used for external accuracy assessment of map quality: out-of-bag, location-grouped 10-fold cross-validation, an independent validation set (5677 observations, 1367 locations) and a stratified random sample of the independent validation set separated by depth layer. Mean error (ME), root mean squared error (RMSE), model efficiency coefficient (MEC) and the prediction interval coverage probability (PICP) were calculated in all four strategies. In addition, the 90th prediction intervals were used to categorize each 25 m pixel into “none”, A, AA or AAA quality as a measure of the internal accuracy assessment.

We obtained large differences depending on the four external accuracy assessment strategies and depth layer (ME = −0.08–0.20, RMSE = 0.41–0.83, MEC = 0.64–0.90, PICP of PI90 = 0.80–0.94). Design-based inference (LSK-SRS) was most indicative of map accuracy based on sampling theory (ME = 0.09–0.17, RMSE = 0.7–0.79, MEC = 0.73–0.82). QRF prediction uncertainty was slightly overestimated. Less than 10% of pixels were designated with AA and AAA and therefore we recommend future studies to also test the achievability of high quality thresholds for Tier 4 *GSM* maps. We believe these 3D soil pH maps at 25 m resolution are useful for a variety of end users and that our workflow can be applied elsewhere and for other soil properties to further diminish the gap of missing spatial soil information.

1. Introduction

Soil is a vital part of the natural environment and essential for global ecosystem services, including production of food and fiber, water infiltration, climate regulation, and maintaining biodiversity. Decision makers therefore require accurate spatial soil information to ensure that the soil and land are being used, managed and conserved in an efficient and sustainable way. Digital soil mapping (DSM) is often used to attain spatially explicit soil information. DSM is the computer-assisted production of soil type and soil property maps, using statistical models to infer the relationship between a response, the soil type or soil property, and the predictors, the spatially exhaustive environmental explanatory variables (McBratney et al., 2003; Scull et al., 2003). Usually, the predictors, also termed covariates, are directly or indirectly related to the main soil forming factors: climate, organisms, relief or topography, parent material and time (Dokuchaev, 1899; Jenny, 1941).

---

[*] Corresponding author.
E-mail address: anatol.helfenstein@wur.nl (A. Helfenstein).

https://doi.org/10.1016/j.geoderma.2021.115659
Received 6 September 2021; Received in revised form 3 December 2021; Accepted 8 December 2021

The *GlobalSoilMap* (*GSM*) working group of the International Union of Soil Sciences (IUSS) was formed across eight geographic "nodes" around the world to strive for a common goal: a high-resolution spatial soil information system of selected soil properties and their uncertainties at six standard depths for the entire world (Arrouays et al., 2014). In order to achieve this common goal, both top-down and bottom-up approaches have been implemented. *SoilGrids*, perhaps the most prominent top-down approach, provides global maps of key *GSM* soil properties developed by ISRIC – World Soil Information, which have since its initial release (Hengl et al., 2014) been improved and updated twice (Hengl et al., 2017; Poggio et al., 2021). Other examples of global soil maps include Global Gridded Surfaces of Selected Soil Characteristics [IGBP-DIS;] (Group, 2000), WISE30sec (Batjes, 2016), S-World (Stoorvogel et al., 2017) and SoilKsatDB for soil saturated hydraulic conductivity (Gupta et al., 2020).

In parallel, bottom-up approaches to create soil information systems (SIS) at regional, national and continental scales have also been implemented. A few examples of countries with SIS using DSM techniques include Denmark (Adhikari et al., 2014), the United States (Hempel et al., 2014), Nigeria (Akpa et al., 2014), Australia (Rossel et al., 2015), France (Mulder et al., 2016; Mulder et al., 2016), Scotland (Poggio et al., 2017) and more recently Brazil (Gomes et al., 2019), China (Liang et al., 2019; Liu et al., 2020) and India (Dharumarajan et al., 2019; Dharumarajan et al., 2020). A global soil organic carbon map (GSOC) was also created based on national soil organic carbon (SOC) maps from 110 countries (Brus et al., 2017; FAO, 2018), which are maintained by the Food and Agriculture Organization of the United Nations (FAO, 2017).

The first publication of the spatial distribution of soil properties in the Netherlands dates back to the 19th century (Felix, 1995). Systematic soil mapping became institutionalized with the establishment of the Dutch Soil Survey institute (StiBoKa) in 1945 (Hartemink et al., 2013). From 1950 to 1995, StiBoKa conducted conventional soil surveys (Buringh et al., 1962) and produced national maps of soil types at a 1:50 000 scale. A review of the history of soil mapping in the Netherlands and its different phases including the first decade of the 21st century was conducted by Hartemink et al. (2013). Various studies compared different (geo) statistical methods and developed prototypes of qualitative and quantitative soil property maps for the Netherlands using the data collected by StiBoKa (Brus et al., 2007; Brus et al., 2009; Kempen et al., 2014). In addition, a variety of DSM techniques were used to update the Dutch soil maps, with a focus on soil organic matter (SOM) and peatland regions (Kempen et al., 2009; Kempen et al., 2011; Kempen et al., 2012). More recently, SOM was estimated at a national scale using a soil type and binary land use map (arable land or grassland), at a resolution of 250 m at four fixed depths (Van den Berg et al., 2017). On the basis of this SOM map, a Dutch contribution to the Global Soil Organic Carbon (GSOC) map was also delivered for the topsoil (0 cm to 30 cm), which was spatially aggregated to 1 km resolution (Walvoort and Hoogland, 2017). Nevertheless, there is an increasing demand for accurate, 3D, and high resolution information of key soil properties for the Netherlands. This is especially important for highly populated and relatively small countries such as the Netherlands (land area $=$ 33 481 km^2) because land use decisions are often made on a field scale, e.g. per agricultural parcel.

Since the establishment of DSM as a research field, the main focus has been on implementing new methods to improve the predictive performance of soil maps. Today, pedometricians can make use of increasing amounts of available spatial data as well as an extensive toolkit of geostatistical and machine learning approaches combined with a powerful computational infrastructure. However, considerably less effort has been invested in providing appropriate measures of the quality of soil maps. This is essential for DSM products to be adopted by a broader community, for future research guidance and most importantly, to ensure that the quantified accuracy is suitable to fulfill the map's purpose (Arrouays et al., 2020).

The quality of maps can be evaluated using internal and/or external accuracy assessment measures. One way to quantify internal, or model-based accuracy assessment is using the prediction uncertainty. In this regard, Quantile Regression Forest [QRF;] (Meinshausen, 2006) models are advantageous within the DSM toolkit not only due to their predictive performance, but also for their ability to quantify prediction uncertainty. Ensemble decision tree models such as Random Forest [RF;] (Breiman, 2001) and QRF have repeatedly outperformed other machine learning and non-machine learning approaches in DSM applications [e.g.] (Hengl et al., 2015; Nussbaum et al., 2017; Keskin et al., 2019). In addition, QRF delivers a probability distribution of the soil property at each prediction location, rather than a single (mean) prediction as with RF (Meinshausen, 2006). To the best of our knowledge, this makes it unique among other machine learning approaches in that the algorithm inherently also gives an indication of the prediction uncertainty. This may be a reason for the increasing use of QRF in DSM in recent years [e.g.] (Vaysse et al., 2017; Lagacherie et al., 2019; Lagacherie et al., 2020; Dharumarajan et al., 2020; Poggio et al., 2021).

Another advantage of QRF prediction uncertainty is that it can be incorporated in the concept of accuracy thresholds for Tier 4 *GSM* products. In order to coordinate and guarantee a minimum quality for soil maps, specifications were made regarding the spatial entity, soil properties, date, uncertainty, validation, documentation and reproducibility of *GSM* products (Arrouays et al., Jan. 2014; Arrouays et al., 2014). Increasing specifications have to be fulfilled with increasing quality of a DSM product, which are organized into so-called *Tiers* (Arrouays et al., 2015). Tier 4 products, which have the strictest requirements, specify three levels of accuracy thresholds (A, AA and AAA) depending on the soil property and depth layer (Appendix A, Table A1), although note that also none of these three levels can be met. These levels specify that prediction uncertainties should be within certain ranges, with an increasingly narrow and therefore accurate range from A to AAA. This may provide a powerful measure to communicate the quality of a soil map to end users for a specific purpose and region. Expressing the uncertainty of predictions in a meaningful way for end users was described as one of the ten major challenges for pedometricians (Wadoux et al., 2021). However, these accuracy thresholds have to our knowledge not yet been used in DSM studies.

Internal accuracy assessment using QRF has the advantage that prediction uncertainty and their respective *GSM* accuracy thresholds are spatially explicit. However, the disadvantage is that these are based on the model structure and model assumptions. Therefore, there is also a need for model-free evaluation of the map's accuracy, i.e. external accuracy assessment.

For assessing the external accuracy, the vast majority of DSM studies use statistical validation methods (Wadoux et al., 2020; Piikki et al., 2021). This usually involves data-splitting, either using a single split involving a calibration and validation set, or repeating this multiple times during n-fold cross-validation (CV). These validation methods are a form of external accuracy assessment because the independent and separate observations not used in model calibration are compared to the predictions at the observation locations (Chatfield, 1995). However, if the validation locations are not selected using a probability sampling design, then the accuracy assessment may be biased (Brus et al., 2011; Brus, 2014; Brus, 2019). In summary, external accuracy assessment without a probability sample gives an indication of the accuracy at independent locations, but these locations may not be indicative of the map itself. Therefore, Brus et al. (2011) conclude that, when evaluating map quality, a probability sample and associated design-based statistical inference should be used for the external accuracy assessment whenever possible. This classical sampling theory method is statistically sound and has been extensively described in statistics (Cochran, 1977) and environmental science (de Gruijter et al., 2006; Gregoire and Valentine, 2007). However, in two recent systematic reviews, Wadoux et al. (2020) reported that only two out of 150 studies used an additional probability sample for validation (Subburayalu and Slater, 2013; Lacoste et al., 2014) and Piikki et al. (2021) reported that only 13% of 188 studies

Table 1
Descriptive statistics of soil pH [KCl] for calibration (PFB) and validation (LSK) data.

Response	Dataset	Locations	Observations	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Skewness
Soil pH [KCl]	Calibration (PFB)	4230	15338	0.90	4.20	4.80	5.20	6.10	9.00	0.52
	Validation (LSK)	1367	5677	1.90	4.40	5.20	5.53	7.00	8.20	0.26

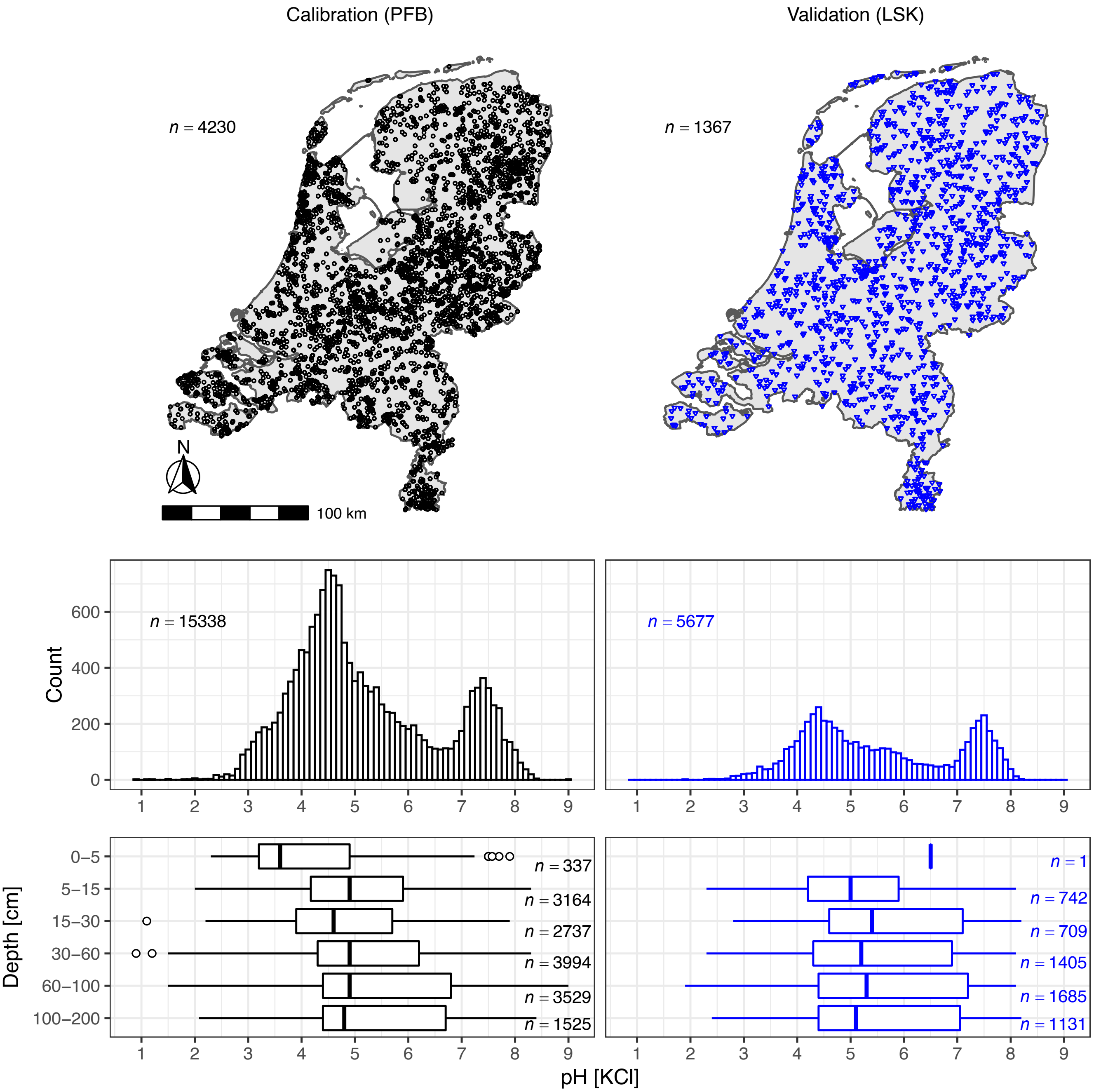


Fig. 1. Soil pH [KCl] sampling locations, histograms and boxplots grouped by depth layer of calibration (PFB; left in black) and validation (LSK; right in blue) data. Observations were grouped into depth layers using the midpoint of each sampled soil horizon.

used probability sampling. This is most likely because probability samples are often not available due to time and cost restraints (Domburg et al., 1997; Hartemink et al., 2008; Hartemink et al., 2010).

Given the strong demand for high-resolution 3D soil information in the Netherlands and the need to properly assess map quality, this study has two main objectives. Firstly, we aim to contribute to the *GSM* project by providing soil pH prediction maps for the Netherlands at 25 m resolution, at any desired depth between 0 m and 2 m, using QRF. We chose to focus on soil pH because it is an indispensable soil property to assess soil processes and fertility: it not only provides information on acidity and alkalinity, but is also an indication of nutrient availability, metal dissolution and (micro-) biological activity (Miller and Kissel, 2010; Weil and Brady, 2017). Secondly, we aim to quantify map accuracy using a) spatially explicit QRF prediction uncertainty and

Table 2
Summary of the main covariates used grouped by soil forming factor. For the complete list of covariates, see Supplement S2, Table S1.

Soil forming factor	Description	Timespan/ version	Source
Climate	Long-term mean, minimum & maximum temperature	1981–2010	KNMI (2020)
Climate	Long-term mean precipitation	1981–2010	KNMI (2020)
Organism	Land use (historical): “Historisch Grondgebruik Nederland” (HGN)	1900, 1960, 1970, 1980	Alterra (2004)
Organism	Land use (recent): “Landelijk Grondgebruiksbestand Nederland” (LGN)	1986–2019	WENR (2020); Hazeu et al., 2020
Organism	Land use (recent): “Bestand Bodemgebruik” (BBG), Top10NL	1993, 1996–2019	CBS (2015); BRT, 2020; BRT, 2021
Organism	Copernicus land monitoring: CORINE Land Cover (CLC), Riparian land cover, water and wetness index, grassland vs. non-grassland, tree cover density	1986, 2000, 2006, 2012, 2018	EEA (2018); Thunnissen and van Middelaar, 1995; Hazeu and de Wit, 2004; EEA, 2007
Organism	Nature land cover maps	1988, 1990, 2003, 2004, 2013	Bakker et al. (1989); Kramer and Clement, 2015; Sanders and Prins, 2001
Organism	Agricultural crop parcels (“BRP Gewaspercelen”)	2005–2019	EZK (2019)
Organism	Agricultural management type, ammonia & nitrogen emissions, manure application	1993, 2018, 2019	BIJ12 (2019); RIVM, 2020
Organism	Water drainage classes, areas behind dikes or not		Maas et al. (2019)
Organism	Vegetation maps: forest classified by age, recreational use, tree species, tree height		de Vries (1992); Clement, 2001
Topography	DEM: “Actueel Hoogtebestand Nederland” (AHN) & hillshade	AHN1 (1997–2004), AHN2 (2007–2012), AHN3 (2014–2019)	AHN (2021)
Topography	AHN2 derivatives: slope, profile curvature, deviation from mean, openness, Topographic Wetness Index (TWI), Multiresolution Valley Bottom Flatness (VBF), valley depth	AHN2 (2007–2012)	AHN (2021)
Parent material	Geomorphology based on geomorphological classes, genesis, form, formation begin and end and relief	2004, 2008, 2019	Koomen and Maas (2004); Maas et al., 2019
Parent material	Physical geography and groundwater maps	2013, 2004	EZK (2013); KRW, 2004
Parent material	(Paleo-) geographical maps	9000–250 B.C., 100–1850 A.D.	Vos (2015); Vos et al., 2020

respective *GSM* Tier 4 accuracy thresholds and b) statistical validation strategies.

2. Materials and methods

2.1. Soil point datasets

We used 21 015 pH measurements, or observations, from 5 597 locations between 0 m to 2 m depth excluding the O horizon or humus layer (Table 1, Fig. 1). Excluding built-up (urban) and water surface area, this approximately yields an average density of 1 soil sampling location per 5 km^2. All observations were retrieved from the Dutch soil database, or “Bodemkundig Informatie Systeem” [BIS;] (IenM, TNO, 2017; TNO, 2020). We chose to use pH measurements conducted in KCl suspension (pH [KCl]) as opposed to the internationally more frequently used H_2O or $CaCl_2$ suspension methods because KCl suspension was the preferred measurement method in the Netherlands from 1950 to 2000 (Supplement S1, Fig. S1). There were $<$ 1000 measurements available using the other methods and we refrained from converting between methods since this may introduce substantial uncertainty.

For model calibration, we used 15 338 pH [KCl] measurements from 4230 locations (Table 1, Fig. 1). At these locations, profile descriptions, or “Profielbeschrijving” (PFB), were made, soil samples were collected from each horizon between 1953 and 2012 and measured in the lab (Supplement S1, Fig. S1). This dataset was specifically chosen for model calibration because it constitutes the majority of soil pH data in the Netherlands. The somewhat clustered locations cover all regions of the Netherlands with the exception of southwestern Flevoland (Fig. 1). The pH calibration data follow a bimodal distribution with the majority of values between 4 and 5 and a smaller peak around 7.5 (Fig. 1). Bimodal distributions for soil pH are common and in the case of the Netherlands can be attributed to the dominating Pleistocene sandy soils vs. Holocene clay soils. Grouped into the *GSM* depth layers by the respective midpoints of the sampled layers, the median pH values are around 4.5 to 5 across all depth layers. Three unexpectedly low values were measured between 15 cm and 60 cm depth (Fig. 1), but there was insufficient evidence for them to be classified as outliers and removed.

The separate and independent validation data were gathered during the “Landelijke Steekproef Kaarteenheden” (LSK) between 1993 and 2000 (5677 measurements from 1367 locations; Supplement S1, Fig. S1). Soil sampling locations were determined in the LSK campaign using a probability sample, more specifically a stratified simple random sample (SRS), wherein 94 strata were defined based on soil type and groundwater class (Finke et al., 2001; Visschers et al., 2007). As with the calibration (PFB) data, observations were made for each soil horizon, which indicates that it is only a SRS in 2D space. This has implications for the statistical validation (see Section 2.6). The validation set also has a bimodal distribution, although the relative difference between the two peaks is much smaller than for the calibration set (Fig. 1). Consequently and in contrast to the calibration data, the overall median as well as the median of each grouped depth layer is above 5 (Table 1, Fig. 1). Note that there are considerably fewer observations (n = 251) in the validation set with a midpoint between 15 cm to 30 cm compared to the other depth layers.

2.2. Covariate selection

The covariates (total number P = 195) were chosen specifically to represent the soil forming factors. The covariates are summarized in Table 2 and a complete list is included in Supplement S2, Table S1.

As indicators of the soil forming factor climate, we used the long-term mean, minimum and maximum temperature and precipitation between 1981 and 2010 from the Royal Netherlands Meteorological Institute [Table 2;] (KNMI, 2020).

The majority of covariates used in this study are historical, agricultural or natural land use and vegetation maps relating to the soil forming factor “organism” (Table 2). We specifically chose a large number of these maps because there is profound anthropogenic influence and high land use intensity in the Netherlands. Approximately 82% of the land surface in the Netherlands is agricultural, urban or infrastructure (Hazeu et al., 2020). Multiple versions covering different time spans were included.

We used the national digital elevation model (DEM) of the Netherlands, or “Actueel Hoogtebestand Nederland” (AHN), and commonly used DEM derivatives for the soil forming factor topography (Table 2). The standard deviation of the systematic as well as random error of AHN 2 and 3 is ± 5 cm (AHN, 2021). With such high accuracy, we considered topographic covariates to be informative not only for hilly regions but also for the large majority of the Netherlands that is relatively flat. We computed the following commonly used DEM derivatives: slope, profile curvature, deviation from the mean value within a local neighborhood, positive and negative openness, Topographic Wetness Index (TWI), Multiresolution Valley Bottom Flatness [MrVBF;] (Gallant and Dowling, 2003) and valley depth (Wood, 1996; Wood, 2009). The deviation from the mean value was computed within a radius of 11 cells (275 m) to account for local changes in topography. AHN 2 was used to obtain these derivatives because it has a higher accuracy than AHN 1 and because AHN 3 has not been thoroughly validated yet (AHN, 2021). A hillshade from the AHN 2 was also downloaded and used.

We used geomorphological, paleo- and physical geography maps as indicators of parent material (Table 2). The parent material for soils in the Netherlands consists almost exclusively of geologically young material from fluvial and coastal lowlands of the Holocene age as part of the Rhine-Meuse delta (60%) as well as Pleistocene sand (Van der Meulen et al., 2013). In this sense for the Netherlands there are no lithology or bedrock maps commonly used in DSM studies in other parts of the world.

For many covariates, multiple versions from different years were included to account for changes in soil forming factors over time. In addition, several of the covariates were based on each other. For example, “Landelijk Grondgebruiksbestand Nederland” (LGN) uses “Bestand Bodemgebruik” (BBG) and “Top10NL” data. This indicates that many of the covariates are highly correlated. Ensemble decision tree models are robust against highly correlated data; it does not cause an overfit or decrease prediction accuracy. However, it is important to note that the higher the number of correlated covariates, the lower the relative importance of each will become, which leads to a distorted variable importance measure (Strobl et al., 2007; Kuhn and Johnson, 2013). For this study, we did not refrain from using many highly correlated covariates because we deemed prediction accuracy more important than model interpretability based on variable importance measures.

2.3. Covariate preprocessing

All covariates were first visually explored for inconsistencies. Rasters were exported at a target resolution of 25 m because this matches the resolution of the LGN land use maps (WENR, 2020; Hazeu et al., 2020) and allows for land use decisions at a fine resolution, e.g. within a small agricultural parcel.

The first step of covariate preparation and preprocessing was to project all covariates to the Amersfoort or RD New coordinate reference system (EPSG:28992). Next, all covariates were resampled to a common origin, extent and resolution. In this step, continuous covariates were resampled using the cubic spline method whereas categorical covariates were resampled using the nearest neighbor method. During reprojection and resampling, the AHN2 was used as reference and the AHN2 “no-data” layer was used as a mask (water and buildings).

Many of the categorical covariates were reclassified because some classes did not occur at observation locations. For example, the detailed classes ($n = 15$) of different cereals in crop rotation covariates (“BRP Gewaspercelen”) were aggregated into one general cereals class (Supplement S2, Table S1).

We stacked the covariates and extracted values at all calibration locations by overlaying them with the covariate stack, resulting in a regression matrix used for model tuning and calibration. Sampling depth information was also included as a predictor in the regression matrix. Including depth along with spatial covariates in a so-called “3D” modelling approach has been used before (Akpa et al., 2014; Filippi et al., 2019; Filippi et al., 2020; Hengl et al., 2017; Ramcharan et al., 2018; Zhang et al., 2020) and is compared in detail to so-called 2D and 2.5D approaches in Ma et al. (2021). More specifically, we included the midpoint of each sampled layer or horizon, as well as the upper and lower boundary to also account for horizon thickness. In summary, we chose to include depth information so that predictions can easily be made at any chosen depth (user specific) and as a means to account for changes in soil pH over depth.

2.4. Model tuning and calibration

For model tuning, calibration, and prediction, it is important to differentiate between mean and median predictions when using QRF. During calibration, for each node in each tree, RF keeps only the mean of the observations that fall into each node. In contrast, QRF keeps the value of all observations in each node (Meinshausen, 2006). Based on this, the fitted QRF can then be used to yield a cumulative probability distribution (i.e., quantiles of the distribution) of the response pH at every sampled location and depth during prediction. In predictive modelling, users are generally interested in the best possible predictions that are closest to the “truth”. It thus makes sense to go for the expected value, i.e. the conditional mean. If the median is used, as e.g. retrieved from the 0.50 quantile in QRF, predictions may be biased if the response is not symmetrically distributed. In addition, median predictions are not additive, if e.g. soil organic carbon stocks need to be calculated from a soil organic carbon map. However, the advantage of using the median is that it is more robust to outliers. For model tuning, we grew RF (not QRF) models with the goal of optimizing hyper-parameters for mean predictions; therefore, there was no need to keep the value of all observations in each node as in QRF (Meinshausen, 2006), which greatly decreased computation time and did not change the tuning results. However, for the final model calibration, a QRF was fitted so that predictions could thereafter be made for both the mean and quantiles (including median).

Model tuning for RF was performed using a location-grouped 10-fold CV wherein all PFB observations from the same location were grouped, abbreviated hereafter as PFB-CV. This means that all observations from the same soil profile were either part of the hold-in or hold-out fold across each of the 10 folds. We tested all combinations (full cartesian grid search) of the following hyper-parameters (Boehmke and Greenwell, 2020):

- **Number of trees in the forest (*ntree*):** 100, 150, 200, 250, 500 (`ranger` default), 750, 1000
- **Number of covariates to consider at any given split (*mtry*):** $\sqrt{P}$ (`ranger` default) and 25%, 33.3% (`randomForest` default) and 40% of P, i.e. 14, 49, 65, 78
- **Complexity of each tree (*minimal nodesize*):** 1, 3 and 5
- **Sampling with replacement (*replace*)**: TRUE (sample with replacement) and FALSE (sample without replacement)
- **Fraction of observations to sample (*sample.fraction*):** 0.5, 0.63 and 0.8 (based on recommendations from Boehmke and Greenwell (2020); this only applies if *replace* = FALSE)

The final set of hyper-parameters was chosen based on the lowest root mean squared error (RMSE) across the 10-fold CV. When the increase in RMSE was below 0.1%, the model with fewer trees was chosen to reduce computation time. Besides the commonly tuned *ntree*, *mtry* and

Table 3
Five strategies to evaluate map quality, based on an internal or external (i.e. statistical validation) accuracy assessment. ME, RMSE, MEC and their respective CI95s in LSK-SRS were calculated using probability sampling theory for SRS (Eqs. (5)–(11)). *PI90 and PICP of LSK and LSK-SRS are identical, respectively, since the same observations are compared to the respective PIs. **Accuracy metrics in LSK-SRS can only be computed for separate layers in order to adhere to the probability sampling design, whereas they can also be computed using observations at all depths for the other statistical validation strategies.

Accuracy assessment	Abbreviation	Description	Statistical validation	Dataset	Accuracy metrics	2D space	Depth**
Internal	–	Tier 4 *GSM* accuracy thresholds	–	PI90 of predictions	**None** (PI90 > 3.0 pH units) **A** (PI90 ⩽3.0 pH units) **AA** (PI90 ⩽2.0 pH units) **AAA** (PI90 ⩽1.0 pH units)	explicit (25 m pixels)	User specific
External	PFB-OOB	Out-of-bag	non-design-based	PFB	ME, RMSE, MEC, PI90, PICP	point locations	All, layers
External	PFB-CV	Location-grouped 10-fold CV	non-design-based	PFB	ME, RMSE, MEC, PI90, PICP	point locations	All, layers
External	LSK	Independent validation	non-design-based	LSK	ME, RMSE, MEC, PI90*, PICP*	point locations	All, layers
External	LSK-SRS	SRS of independent validation	design-based	LSK	ME, RMSE, MEC, CI95, PI90*, PICP*	strata weighed	Layers

nodesize hyper-parameters, we also tested different values related to the sampling scheme. Sampling with replacement can lead to biased variable split selection when there are many categorical covariates with varying numbers of levels (Janitza et al., 2016; Strobl et al., 2007). Hence, we tested sampling without replacement because we had many categories that were not balanced, hoping to achieve a less biased use of all levels across the trees in the forest. In addition, decreasing the sample fraction size of observations leads to more diverse trees and thus lowers between-tree correlation, which can increase the prediction accuracy, especially if there are a few dominating covariates (Boehmke and Greenwell, 2020). The splitting rule used during tree construction (*splitrule*) was held constant at the default value of selecting the split at each node that minimizes the variance of the response.

The final QRF used for model predictions was fitted using all soil observations in the calibration set ($n = 15338$), covariates including depth indications ($P = 195$) and the final set of optimized hyper-parameters. Permutation was used to assess relative variable importance during model fitting. In this method, the mean squared error (MSE) is compared to the MSE after permuting the values of a covariate, yielding a difference in MSE per covariate. These MSE differences are normalized by the standard deviation of the MSE differences over all covariates (Breiman, 2001).

2.5. Maps of predicted soil pH, uncertainty and accuracy thresholds

The calibrated QRF were used to derive the mean, median (0.50 quantile; $q_{0.50}$), 0.05 quantile ($q_{0.05}$) and 0.95 quantile ($q_{0.95}$) at every 25 m pixel and each standard depth layer specified by *GSM* (0 cm to 5 cm, 5 cm to 15 cm, 15 cm to 30 cm, 30 cm to 60 cm, 60 cm to 100 cm and 100 cm to 200 cm) over the Netherlands. Predictions were made at the same support as the observations, i.e. at point support at the center of each pixel and the specified depth increment. Support is defined as the area or volume over which a measurement or prediction is made [Section 4.8] (Webster and Oliver, 2007).

In addition, spatially explicit 90% prediction intervals (PI90) were obtained at every 25 m pixel as a measure of prediction uncertainty as follows:

$$PI90 = q_{0.95} - q_{0.05} \tag{1}$$

As an additional measure of map quality using internal accuracy assessment, we used the PI90 to designate every 25 m pixel at every predicted depth layer into one of four thresholds: none, A, AA and AAA (Table 3). These accuracy thresholds are specified by *GSM* Tier 4 products (Arrouays et al., 2015) and do not vary over depth in the case of soil pH (Appendix A, Table A1). From “none” to AAA, the PI90 (uncertainty) of a given prediction gradually decreases, indicating that users can be very certain about predictions at AAA locations and least certain about predictions at “none” locations. For example, for a AAA pixel 9 out of 10 times the true value is less than ±0.5 pH units from the mean prediction, and less than ±1.5 pH units from the mean prediction for a A pixel.

2.6. Evaluation of map accuracy using statistical validation

2.6.1. Non-design-based inference

We also evaluated map quality using external accuracy assessment in the form of statistical validation strategies (Table 3). Firstly, we used the out-of-bag (OOB) observations, in other words the PFB observations not selected during bootstrapping when QRF is calibrated [PFB-OOB;] (Breiman, 2001). This is commonly used in various disciplines to assess accuracy of RF or other ensemble decision tree models. Secondly, we used location-grouped 10-fold CV (PFB-CV; see Section 2.4). Compared to PFB-OOB, this method was chosen because it prevents observations from the same location in being both in the hold-in and hold-out set, wherein the hold-in samples are used for model calibration and the hold-out for model validation. PFB-OOB and PFB-CV both only make use of the PFB calibration dataset (see Section 2.1). Thirdly, we used the LSK dataset as an independent validation set. The probability sampling design of the LSK cannot easily be utilized when considering all depths because there are multiple observations from different depth layers at the same locations. This means that in 3D, it cannot be considered a SRS. We nevertheless included this strategy because we wanted to investigate whether there are substantial differences between a non-design-based vs. design-based (see Section 2.6.2) inference of LSK.

To obtain commonly used accuracy metrics, mean predictions at all depths were used to calculate residuals and estimate from them the mean error (ME or bias), the RMSE and the model efficiency coefficient (MEC):

$$\widehat{ME} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i) \tag{2}$$

$$\widehat{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2} \tag{3}$$

$$\widehat{MEC} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \tag{4}$$

where n is the number of observations, Y_i and $\widehat{Y}_i$ are the ith observation and prediction, respectively, at a certain location and depth, and $\overline{Y}$ is the mean of all observations. The MEC was originally used in hydrological modelling (Nash et al., 1970) and is also referred to as the mean squared

Table 4

Metadata of the LSK-SRS method per depth layer, including number of removed observations, number of averaged observation pairs, percentage of the strata that were removed from the total ($H = 94$), which strata were removed and the percentage of the Netherlands covered. *We refer to the strata codes from Finke et al. (2001), Appendix A.

Depth layer	Observations		Strata		% NL coverage
	# removed	# averaged pairs	% removed	Removed*	
0 cm to 5 cm	–	–	–	–	–
5 cm to 15 cm	0	0	6.38	1904, 1910, 1915, 2007, 2108, 2114	98.51
15 cm to 30 cm	0	0	7.45	1913, 1914, 1917, 2102, 2116, 2117, 2901	95.66
30 cm to 60 cm	73	13	0	–	100
60 cm to 100 cm	417	56	0	–	100
100 cm to 200 cm	222	4	9.57	1502, 1503, 1504, 1505, 1915, 2201, 2401, 2601, 2701	97.15

error skill score in other disciplines such as meteorology (Wilks, 2011). In addition, all quantiles from 0 to 1 were predicted at all depths at all observation locations for statistical validation to obtain the PI90 as well as the prediction interval coverage probability (PICP) of all PIs. The PICP is the proportion of observations that fall into the corresponding prediction interval (Papadopoulos et al., 2001). If the model is able to accurately quantify the uncertainty, then the percentage of observations within a PI should be close to the PICP.

2.6.2. Design-based inference

In order to conduct a design-based inference of map accuracy using the SRS probability sample, we grouped the LSK observations into the *GSM* depth layers so that there was at most one observation at each location (LSK-SRS) for every depth layer. Observations were grouped into *GSM* depth layers by allocating each observation to the layer in which the midpoint of the sampled soil horizon lies. This means that some locations had no observations for that particular depth layer while other layers had more than one. For locations where there were more than one observation per depth layer, the observation was chosen whose midpoint was closest to the midpoint of the *GSM* depth layer. This meant that not every observation was used for LSK-SRS. If the distances were identical, then the observations and predictions were averaged (mean). If there were no observations for an entire stratum for a particular depth layer, then that stratum was removed from the analysis. The number of observations that were left out or averaged as well the left-out strata and the percentage of land they constituted were reported for each depth layer (Table 4). The uppermost *GSM* depth layer cannot be validated using LSK (both design- and non-design based inference) because there is only 1 observation from 0 cm to 5 cm (Fig. 1 and Table 4).

For each depth layer (except 0 cm to 5 cm), the estimates of ME, RMSE and MEC (Eqs. (2)–(4)) were adjusted for LSK-SRS according to probability sampling theory. In addition, the lower and upper 97.5% confidence limits, which together give the 95% confidence intervals (CI95) of these metrics were also computed according to sampling theory [Section 7.2.4] (de Gruijter et al., 2006). The estimated mean error ($\widehat{ME}$), the associated estimation error variance and the lower and upper confidence limits were computed as follows:

$$\widehat{ME} = \sum_{h=1}^{H}\left(w_h \cdot \frac{1}{n_h}\sum_{i=1}^{n_h}\left(Y_{hi} - \widehat{Y}_{hi}\right)\right) \tag{5}$$

$$Var\left(\widehat{ME} - ME\right) = \sum_{h=1}^{H}\left(w_h^2 \cdot \frac{1}{n_h(n_h-1)} \cdot \sum_{i=1}^{n_h}\left(Y_{hi} - \widehat{Y}_{hi} - \left(\frac{1}{n_h}\sum_{i=1}^{n_h}\left(Y_{hi} - \widehat{Y}_{hi}\right)\right)\right)^2\right) \tag{6}$$

$$lower\ \&\ upper\ CL = \widehat{ME} \pm qt(0.95,\ n-H) \cdot \sqrt{Var\left(\widehat{ME} - ME\right)} \tag{7}$$

where H is the total number of strata, n_h is the number of observations in stratum h ($h = 1, \ldots, H$), w_h is the stratum weight, which equals the stratum area A_h divided by the total area A, Y_{hi} and $\widehat{Y}_{hi}$ are the ith observation and prediction in stratum h, respectively, and $qt(0.95, n-H)$ is the 0.95 quantile with $n-H$ degrees of freedom.

The estimated mean squared error ($\widehat{MSE}$), its estimation error variance and respective lower and upper confidence limits were computed in a similar manner:

$$\widehat{MSE} = \sum_{h=1}^{H}\left(w_h \cdot \frac{1}{n_h}\sum_{i=1}^{n_h}\left(Y_{hi} - \widehat{Y}_{hi}\right)^2\right) \tag{8}$$

$$Var\left(MSE - \widehat{MSE}\right) = \sum_{h=1}^{H}\left(w_h^2 \cdot \frac{1}{n_h(n_h-1)} \cdot \sum_{i=1}^{n_h}\left(\left(Y_{hi} - \widehat{Y}_{hi}\right)^2 - \left(\frac{1}{n_h}\sum_{i=1}^{n_h}\left(Y_{hi} - \widehat{Y}_{hi}\right)^2\right)\right)^2\right) \tag{9}$$

$$lower\ \&\ upper\ CL = \widehat{MSE} \pm qt(0.95,\ n-H) \cdot \sqrt{Var\left(MSE - \widehat{MSE}\right)} \tag{10}$$

The RMSE and its respective CI95 were obtained by simply taking the square root of the $\widehat{MSE}$ and its lower and upper confidence limits. ME and RMSE and respective CI95 metrics are in units of the response variable (pH [KCl]).

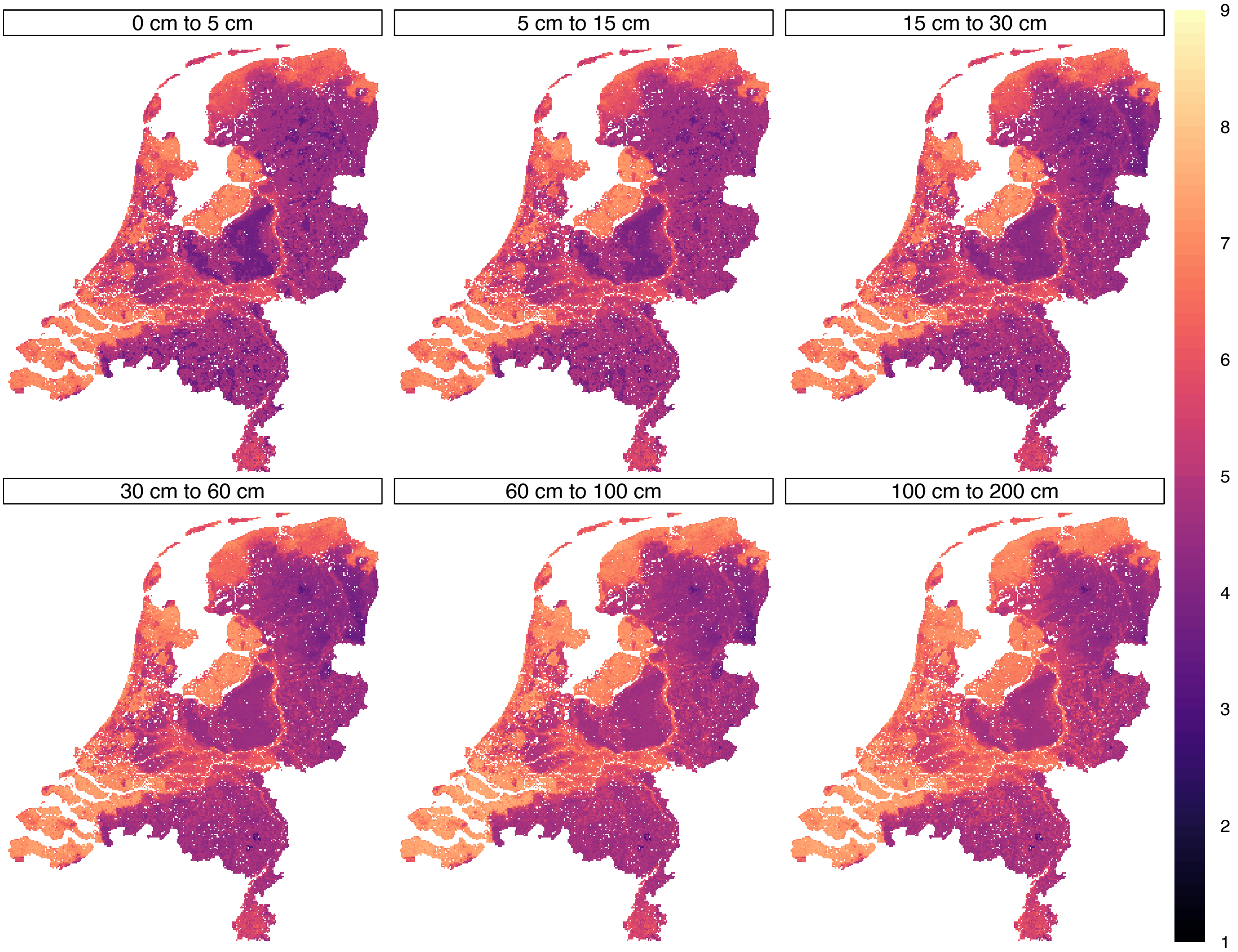


Fig. 2. Soil pH [KCl] predictions (mean) for every 25 m pixel over the Netherlands for the six depth layers specified by *GSM*.

The estimate of the model efficiency coefficient ($\widehat{MEC}$) and its CI95 for LSK-SRS were calculated as follows:

$$\widehat{MEC} = 1 - \frac{\widehat{MSE}}{\widehat{Var(Y)}} \tag{11}$$

where $\widehat{Var(Y)}$ is defined in Eq. 7.16 in de Gruijter et al. (2006) as:

$$\widehat{Var(Y)} = \widehat{\overline{Y}^2_{st}} - (\widehat{\overline{Y}}_{st})^2 + \widehat{V}(\widehat{\overline{Y}}_{st}) \tag{12}$$

where

$$\widehat{\overline{Y}^2_{st}} = \sum_{h=1}^{H}\left(w_h \cdot \frac{1}{n_h}\sum_{i=1}^{n_h} Y^2_{hi}\right) \tag{13}$$

$$\widehat{\overline{Y}}_{st} = \sum_{h=1}^{H}\left(w_h \cdot \frac{1}{n_h}\sum_{i=1}^{n_h} Y_{hi}\right) \tag{14}$$

$$\widehat{V}\left(\widehat{\overline{Y}}_{st}\right) = \sum_{h=1}^{H}\left(w^2_h \cdot \frac{1}{n_h(n_h-1)}\sum_{i=1}^{n_h}\left(Y_{hi} - \frac{1}{n_h}\sum_{i=1}^{n_h} Y_{hi}\right)^2\right) \tag{15}$$

The CI95 of the MEC was computed by taking a bootstrap sample from all observations per stratum 1000 times and then retrieving the 0.025 and 0.975 quantile of the distribution. For strata with just one observation per depth layer ($n_h = 1$), a within-stratum variance cannot be calculated in Eqs. 6, 9 and 15 because a minimum of two observations are required. For these strata, we took the average of the within-stratum variances of all strata with two or more observations.

2.7. *Software and computational framework*

The computational framework was entirely based on open source software and performed on a Ubuntu 20.04.1 operating system (OS) with 48 cores and 126 GB working memory (RAM). QGIS (version 3.16.3) was used for covariate and soil prediction map visualization (QGIS Development Team, 2021). All scripts, metadata, reclassification tables (the original covariate values, a description of each class, the reclassified value and description of the reclassified class) of the categorical covariates and model outputs (soil pH and their associated uncertainty and accuracy threshold maps) are openly accessible (see code and data availability below).

Resampling, reclassification of categorical covariates and masking covariates for buildings and water bodies was done using the GDAL (version 3.1.3) functions `gdalwarp`, `gdal_calc` and `gdal_translate`, respectively (GDAL/OGR contributors, 2020). Reclassification of categorical covariates was automated as much as possible. First, a table was exported from R with all the values within a raster. A short description of the original value (e.g. 3) as well as a reclassified value (e.g. 2) were manually added where necessary (e.g. “barley” and “cereals”). Lastly, each reclassification table was imported back into R, converted into a string format necessary for GDAL’s `gdal_calc`

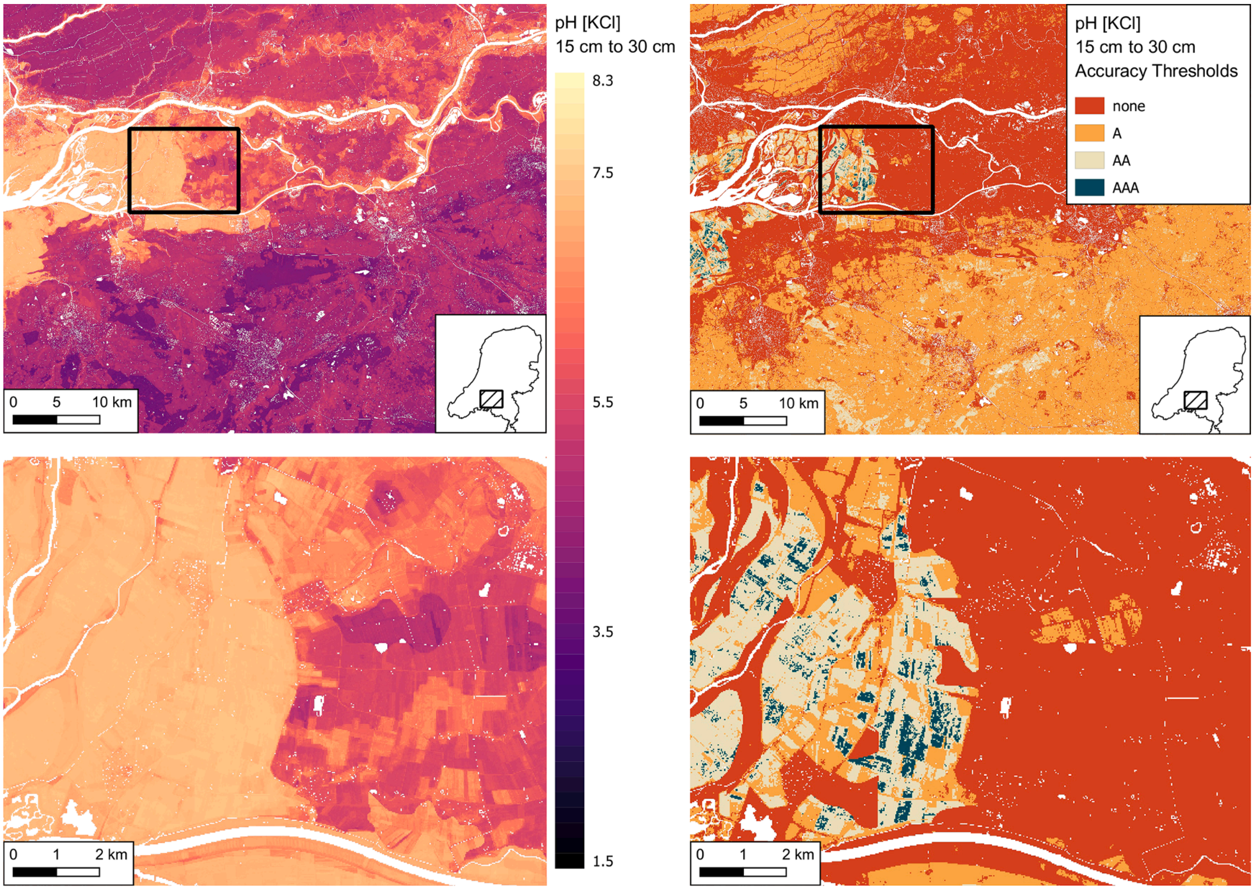


Fig. 3. Two-stage zoom-in of mean predictions of pH [KCl] (left) and corresponding accuracy thresholds (right) for the depth layer 15 cm to 30 cm in the southern part of the Netherlands.

function and reclassified accordingly. DEM derivatives were calculated using SAGA-GIS [version 7.3.0;] (Conrad et al., 2015). Covariate preprocessing steps using GDAL and SAGA-GIS were run on the OS as suggested in Hengl and and MacMillan (2019) but parallelized in R [version 4.0.3;] (R Core Team, 2020) using the `doParallel` (Wallig et al., 2020) and `foreach` packages (Wallig et al., 2020). GDAL and SAGA-GIS were specifically chosen for these steps because it massively decreased computation time compared to using similar functions in R using the `raster` (Hijmans, 2020) or `terra` packages (Hijmans, 2021). All other covariate preprocessing steps including extracting covariate values at calibration locations were done in R using the `raster` or `terra` packages.

All model tuning, calibration and evaluation using statistical analysis was done in R. The indices necessary for the location-grouped 10-fold CV were made using the `CAST` package (Meyer, 2021). The remaining model tuning and selection of hyper-parameters were done using the `caret` package (Kuhn, 2019; Kuhn, 2020). We used the `ranger` package (Wright and Ziegler, 2017) with the option “quantreg” to grow

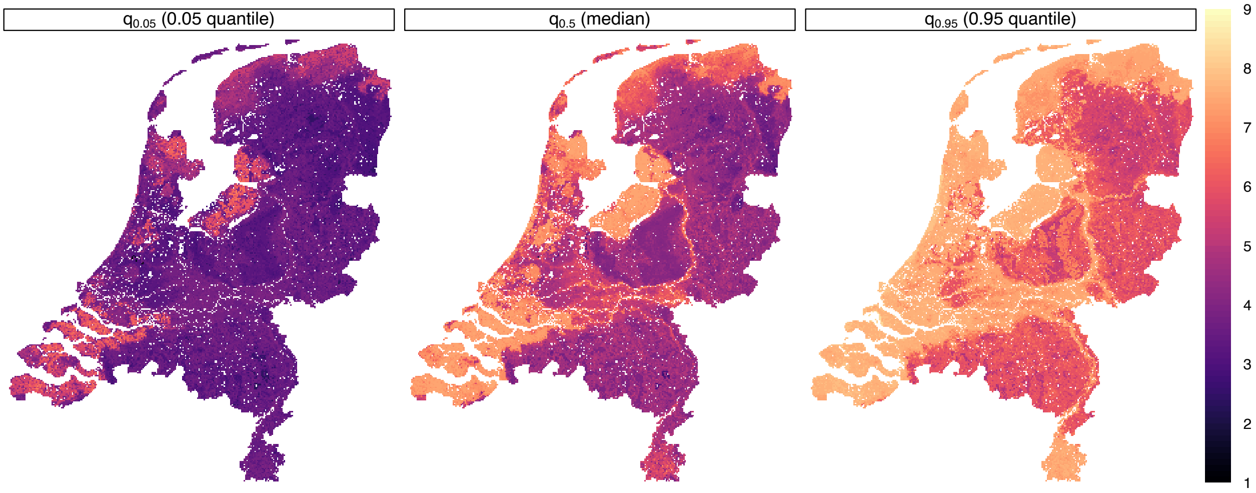


Fig. 4. $q_{0.05}$ (left), median (middle) and $q_{0.95}$ (right) pH [KCl] for every 25 m pixel over the Netherlands for the depth layer 15 cm to 30 cm.

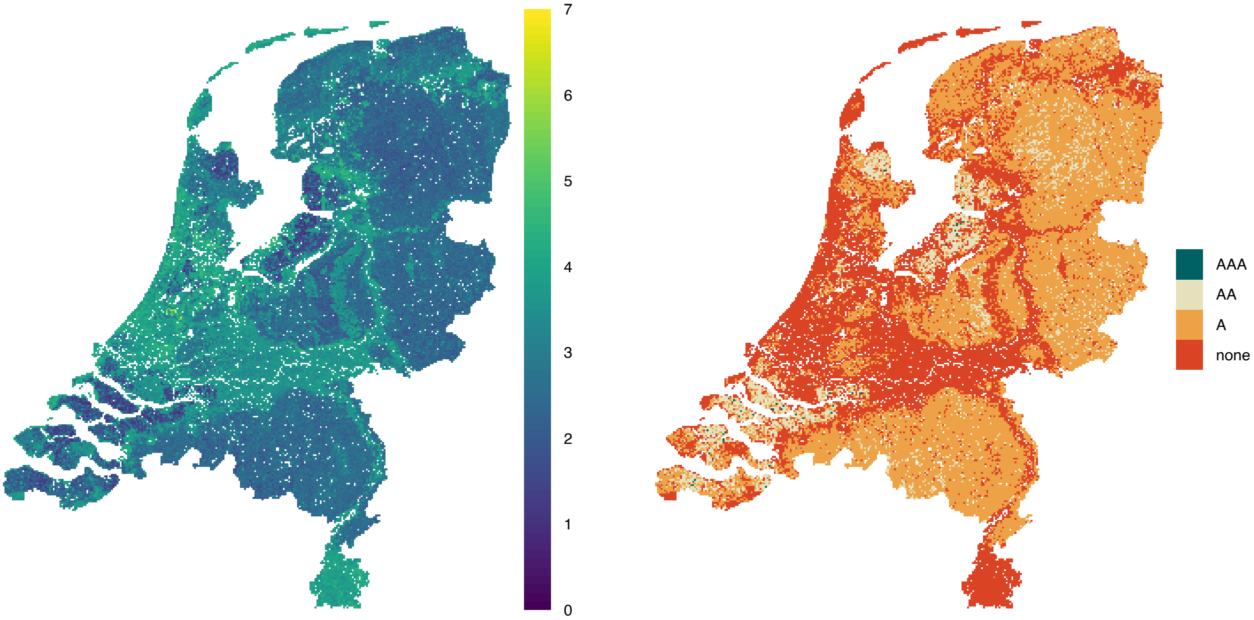


Fig. 5. The PI90 (left) and corresponding accuracy thresholds (none, A, AA and AAA; right) for the depth layer 15 cm to 30 cm.

a QRF and without it to grow a RF (for tuning). For predictions, the option “quantiles” was used to predict quantiles while the option “response” was used to predict the mean. A combination of the `ranger` and `terra` packages was used for predicting at all locations and depths. Finally, prediction maps were visualized using the `rasterVis` package (Lamigueiro and Hijmans, 2021). The complete computational

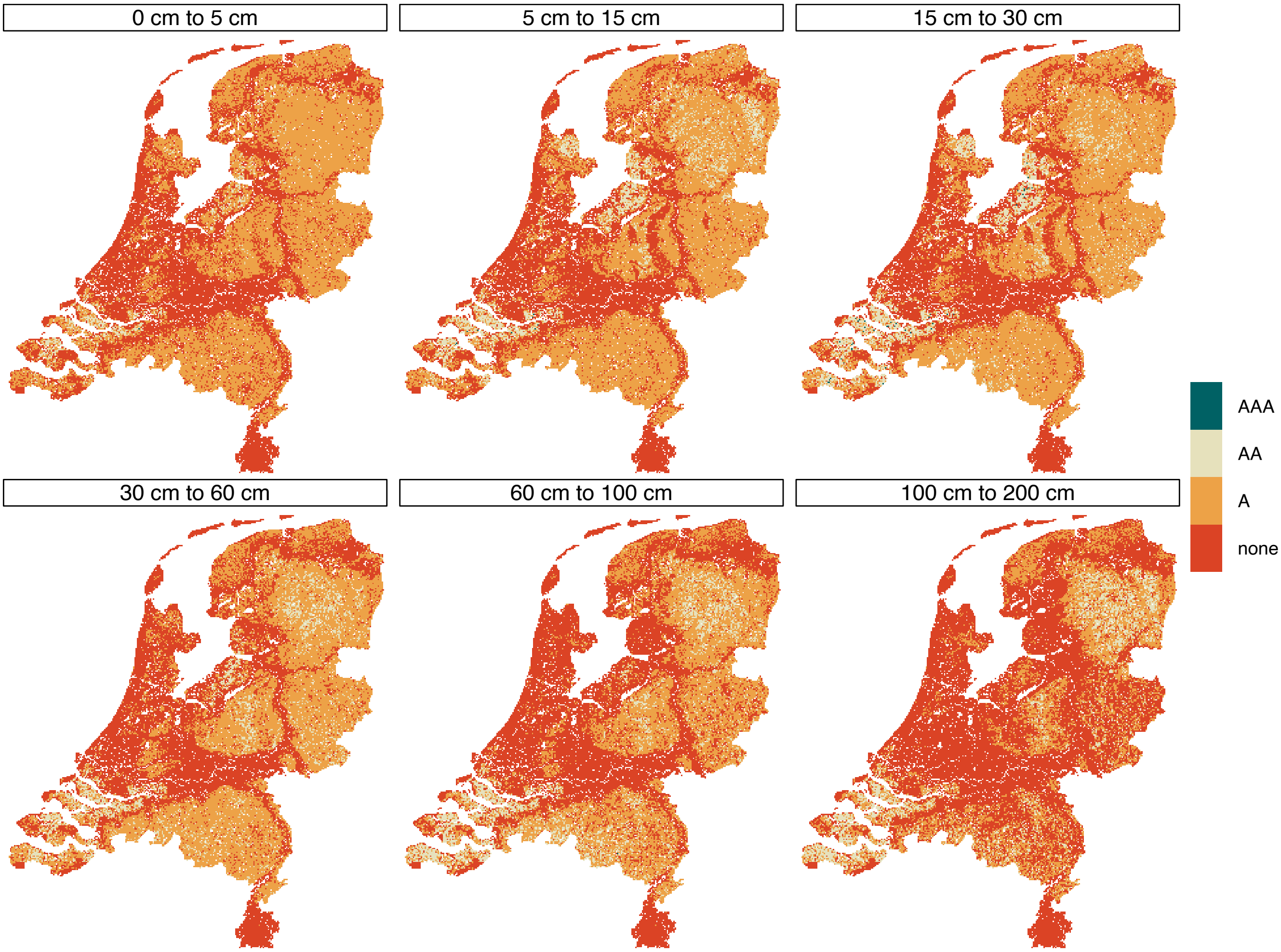


Fig. 6. pH [KCl] accuracy thresholds (none, A, AA and AAA) for the six depth layers specified by *GSM*.

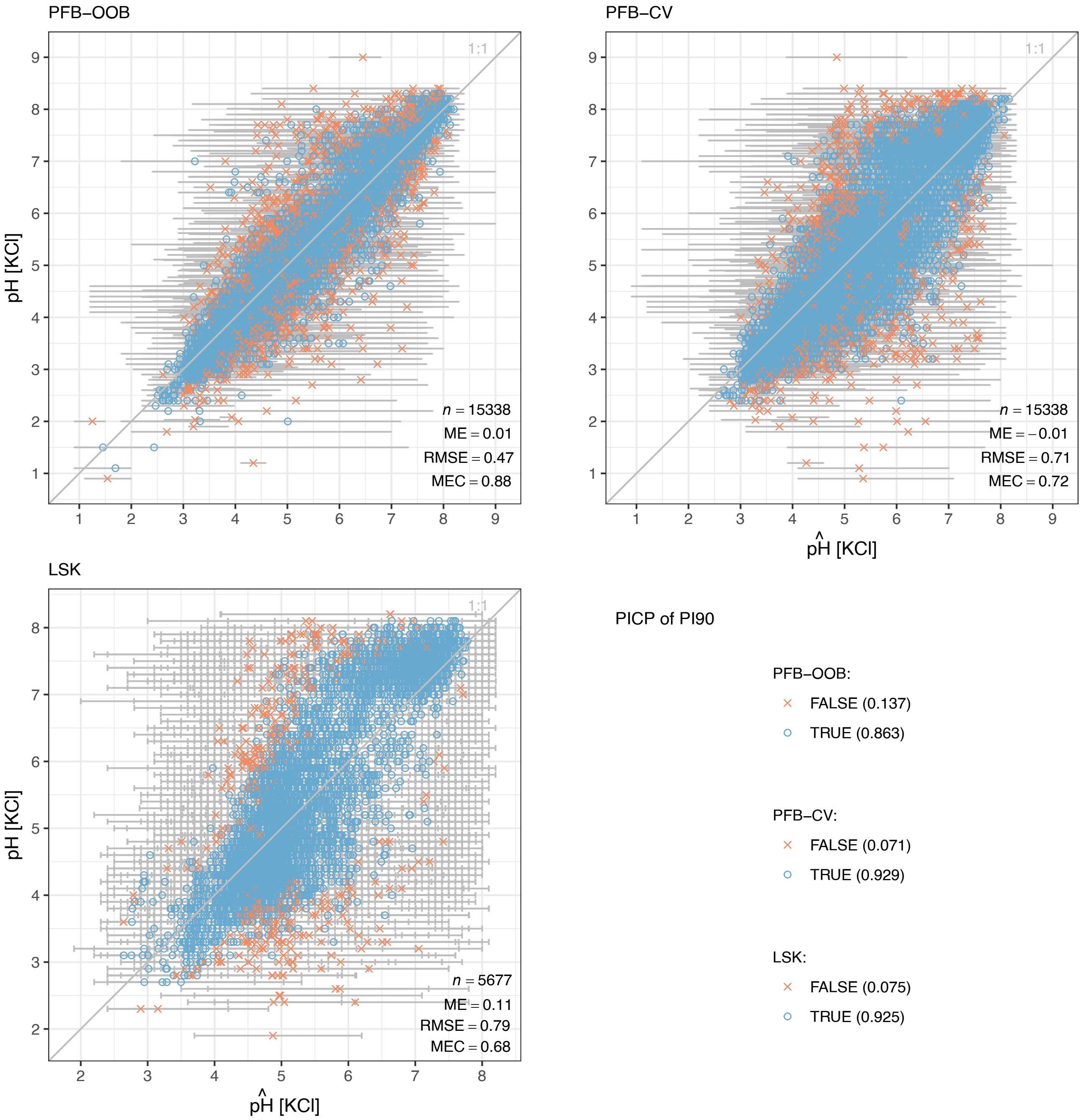


Fig. 7. Accuracy plot (predicted vs. observed) and metrics of soil pH [KCl] at all depths for PFB-OOB, PFB-CV and LSK. The horizontal grey error bars are the PI90. Blue circles indicate observations within the PI90 (i.e. error bars cross the 1:1 line) and orange crosses indicate observations outside the PI90 (i.e. error bars do not cross the 1:1 line). The PICP of PI90 specifies the percentage of the observations inside (TRUE) or outside (FALSE) the PI90. Notice the different scale of the axes for LSK compared to PFB-OOB and PFB-CV.

workflow for 24 pH maps (mean and three quantiles for six depth layers) took approximately 688 CPU-hours and included covariate preprocessing (96 CPU-hours), model tuning and calibration (232 CPU-hours) and prediction (360 CPU-hours).

3. Results

3.1. Model tuning, calibration and variable importance

Out of the 336 possibilities of hyper-parameter combinations tested, we chose *ntrees* = 500, *mtry* = 49, minimal *nodesize* = 1 and a sampling scheme without replacement with a sample fraction of 0.8. This set of hyper-parameters resulted in the lowest RMSE (0.713) out of all combinations across the location-grouped 10-fold CV. Increasing the number of trees above 500 decreased the RMSE by less than 0.1 %. The range of RMSE values obtained from all hyper-parameter combinations was 0.712–0.732. Slightly improved performance with higher numbers of trees and 25 % of the total covariates to consider at any given split (*mtry*) align with the general recommendations of using ensemble decision tree models. Sampling without replacement with sample fractions of 0.8 or lower generally led to lower RMSE values. These results can be explained by the high number of categorical covariates with large differences in the number of classes.

The most important variables of the final model calibration based on

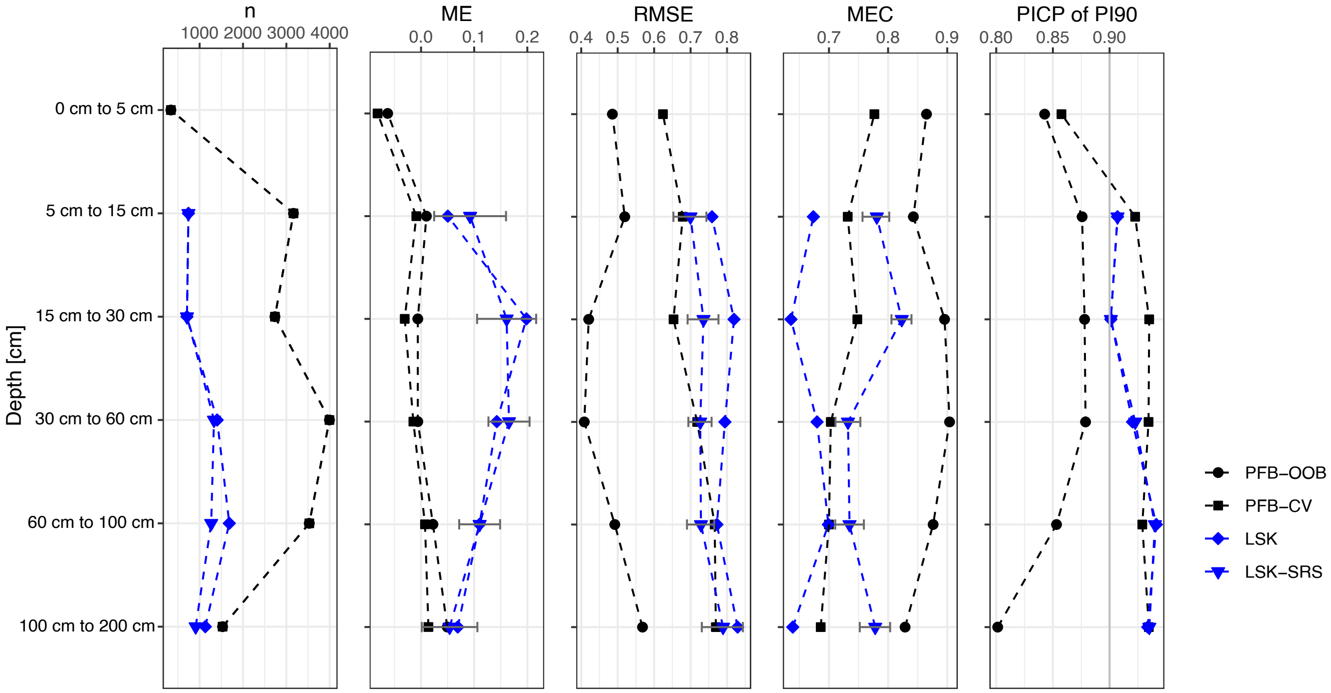


Fig. 8. Total number of observations (n), ME, RMSE, MEC and PICP of PI90 of the different strategies (PFB-OOB, PFB-CV, LSK and LSK-SRS) over depth. Dashed lines do not represent actual data and are only for visual guidance. Gray error bars indicated the CI95 of LSK-SRS accuracy metrics.

permutation were physical geographical maps, followed by geomorphological maps, the AHN (DEM), forest type, land use and temperature maps (Supplement S3, Fig. S2). However, these variable importance measures are not reliable due to high correlation between a large proportion of the covariates (Strobl et al., 2007; Kuhn and Johnson, 2013).

3.2. Soil pH maps: mean predictions

Mean prediction maps of soil pH at 25 m resolution varied across the different *GSM* depth layers (Fig. 2). High pH values indicating alkaline soils were found in the marine clay regions, for example in the Southwest (Zeeland) and the regions where land was reclaimed, or “polders” (e.g. Flevoland). Low pH indicating acidic soils were found in sandy areas, e.g. the glacial moraines of the Saalien ice age such as the Utrechtse Heuvelrug and Veluwe regions. In the very south of the Netherlands (Limburg), the model predicted neutral or slightly alkaline soils. This is the only region in the Netherlands that contains calcareous sediments. There was also a distinct pattern along rivers, such as along the Rhine, Maas and IJssel River valleys. Here, the pH was also neutral, reflecting the riverine clays and sediments being deposited along the river banks.

The spatial patterns of the mean predictions over all depth layers suggested that with increasing depth, vegetation and land use played a smaller role (Fig. 2). QRF was able to detect the large effect that forested and peatland regions had on soil pH for the uppermost soil layers, revealing acidic conditions in general. With increasing depth, vegetation and land use appeared to become less important and the spatial patterns at these depths resemble geomorphological and parent material indicators. There were a few exceptions to this general pattern in the deepest soil layer (100 cm to 200 cm) that showed distinct local patterns of low pH values, which might be attributed to regions with thick peat layers. Predictions at a high resolution revealed differences in soil pH between and within small agricultural parcels (Fig. 3, left).

3.3. Soil pH maps: quantiles, PI90 and accuracy thresholds

The maps of quantiles, PI90 and corresponding accuracy thresholds were regarded as a spatially explicit internal accuracy assessment since it quantifies the prediction uncertainty of the calibrated QRF model. For example, between 15 cm to 30 cm, $q_{0.05}$ showed low (acidic) values almost throughout the Netherlands except for the marine clay regions (Fig. 4). $q_{0.95}$ revealed high (alkaline) values of soil pH except for Pleistocene “coversand” regions, which showed values around 6. This pattern was evident across all depth layers (Supplement S4, Fig. S3). In contrast, the loamy riverine and peat areas in the Netherlands were

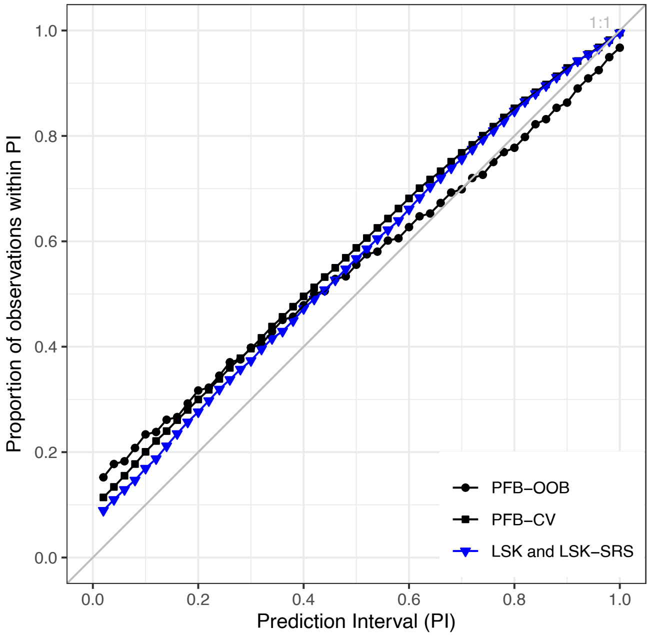


Fig. 9. PICP of all observations for the different strategies (PFB-OOB, PFB-CV, LSK and LSK-SRS). The PICP of LSK and LSK-SRS are identical because the observations and predictions are the same. Lines connecting the points do not represent actual data and are only for visual guidance.

contrasted by much higher uncertainty (Fig. 5 and Supplement S4, Fig. S3).

The PI90 and accuracy thresholds thereof indicated large uncertainty of QRF predictions (PI90 > 2.0, i.e quality “none” or “A”) for the majority of the Netherlands (Figs. 5 & 6). Depending on the depth layer, the percentage of pixels for each accuracy threshold ranged between 39.5% to 64.7% for “none”, 26.5% to 53% for A, 2.0% to 9.0% for AA and $<$ 0.2% for AAA. In general, areas with marine clay soils or sandy soils showed smaller PI90 and better accuracy thresholds. AAA quality was only achieved for a few pixels in the marine clay soils in the depth layers 0 cm to 5 cm, 5 cm to 15 cm and especially 15 cm to 30 cm. With increasing depth, a larger part of the Netherlands did not even achieve the lowest threshold A. However, many sandy soil regions improved from A to AA quality with increasing depth (Fig. 6). When zooming in, the maps of accuracy thresholds also revealed differences between and within agricultural parcels (Fig. 3, right). Areas with high variation in accuracy thresholds were often not the same areas as areas with high variation in pH predictions.

3.4. Evaluation of map accuracy using statistical validation

3.4.1. Non-design-based inference

The external accuracy assessment of soil pH maps using non-design-based statistical validation techniques revealed different results between PFB-OOB, PFB-CV and LSK (Figs. 7–9). The accuracy plots and metrics of mean predictions over all depth layers combined indicated the best performance using PFB-OOB (ME $= 0.01$ pH, RMSE $= 0.47$ pH, MEC $=$ 0.88), followed by PFB-CV (ME $= -0.01$ pH, RMSE $= 0.71$ pH, MEC $=$ 0.72) and the worst performance using LSK (ME $= 0.11$ pH, RMSE $=$ 0.79 pH, MEC $= 0.68$; Fig. 7). Statistical validation using PFB-OOB suggested much higher map accuracy than both PFB-CV and LSK, where residuals were larger, RMSE higher and MEC lower. However, we do not recommend the reader to choose the statistical validation method based on apparent performance, as the metrics may not necessarily indicate the “true” map accuracy (see Section 4.1). Both strategies that used the PFB dataset had a ME around zero, indicating an unbiased map, whereas the strategy using the independent validation set (LSK) indicated that QRF systematically under-predicted pH by 0.11 units. For PFB-CV and LSK, observations in the low pH ranges were generally predicted too high, while observations in the high pH ranges were generally predicted too low.

We also found a clear discrepancy between PFB-OOB, PFB-CV and LSK based on accuracy metrics of mean predictions over depth (Fig. 8). At PFB locations, the model slightly overpredicted soil pH for the first 5 cm and slightly underpredicted soil pH from 60 cm to 200 cm but was unbiased for the depth layers in between. Statistical validation using LSK resulted in positive ME values for all depth layers and the bias was highest for the depth layers 15 to 30 cm (ME $= 0.20$), followed by 30 cm to 60 cm (ME $= 0.14$) and 60 cm to 100 cm (ME $= 0.11$). In contrast to ME results, RMSE values of PFB-OOB and PFB-CV indicated different results for all depth layers. PFB-OOB RMSE results indicated the best model performance at depths 15 cm to 60 cm (RMSE ≈ 0.4). PFB-CV also indicated lower accuracy in depth layers below 60 cm (RMSE $= 0.77$), but even for the upper 5 cm revealed relatively poor results compared to PFB-OOB (RMSE $= 0.62$ vs. RMSE $= 0.49$). In comparison to PFB-OOB and PFB-CV, RMSE values using the LSK were higher for all depth layers but only varied slightly over depth between 0.76 and 0.83. MEC values of PFB-OOB indicated best model fit for predicting at 15 cm to 60 cm (0.90) and lowest for 100 cm to 200 cm (0.83). MEC values using PFB-CV were highest for 0 cm to 5 cm (0.78) and gradually decreased to 0.69 in the deepest layer. As with ME and RMSE, LSK also indicated relatively poor results of map accuracy over depth based on lower MEC values at all depths between 0.64 and 0.70. Accuracy evaluation using the LSK dataset could not be made for 0 cm to 5 cm, since there was only one observation with a midpoint within this depth layer (Fig. 8).

The PFB-CV and LSK results of the PICP of PI90 (Figs. 7 and 8) and all other PIs (Fig. 9) indicated that QRF prediction uncertainty was slightly overestimated. The PICP of PI90 for PFB-CV and LSK were between 0.925 and 0.929 over all depths (Fig. 7) and above 0.90 for most depth layers (Fig. 8). Only the uncertainty of the depth layer 0 cm to 5 cm for PFB-CV was underestimated. The prediction uncertainty based on PFB-CV and LSK was also overestimated for all remaining PIs (Fig. 9). In contrast, the PICP of PI90 for PFB-OOB revealed a clear underestimation of the prediction uncertainty (0.863; Fig. 7), especially at depth layers 0 cm to 5 cm and 60 cm to 200 cm (Fig. 8). Based on the PFB-OOB evaluation, the 0.70 PI of QRF was the most accurate, wherein lower PIs overestimated the prediction uncertainty and higher PIs underestimated the prediction uncertainty (Fig. 9).

3.4.2. Design- vs. non-design-based inference

The statistical validation of the design-based inference (LSK-SRS) resulted in ME values between 0.09 and 0.17, RMSE values between 0.70 and 0.79 and MEC values between 0.73 and 0.82, depending on the depth layer (Fig. 8)). LSK-SRS indicated higher map accuracy compared to LSK (ME $= 0.05$–0.14, RMSE $= 0.76$–0.83 and MEC $= 0.64$–0.70). The small differences in the number of observations (n) used in the statistical validation for LSK and LSK-SRS were because only one observation was used per depth layer for LSK-SRS (Section 2.6.2, Table 4). Similarly to LSK, evaluation of the map using LSK-SRS also showed biased results, following a similar pattern over depth. The ME values of LSK were within the CI95 of the ME of LSK-SRS. RMSE values also followed the pattern of LSK over depth. However, the values indicated a higher map accuracy and were closer to the PFB-CV metrics. The RMSE values of LSK were outside the CI95 range except for the depth interval from 100 cm to 200 cm, but the RMSE metrics of PFB-CV were mostly within this range. The MEC metrics and their respective CI95s of LSK-SRS indicated better mapping accuracy results than PFB-CV and LSK. The CI95 did not overlap with MEC metrics from other approaches. Overall, the CI95 were narrow for both RMSE and MEC metrics, indicating a high certainty of RMSE and MEC values. The PICP metrics of LSK and LSK-SRS were identical because the observations and quantile predictions were the same (Figs. 8 and 9). In summary, metrics across all four strategies (PFB-OOB, PFB-CV, LSK and LSK-SRS) varied by respective minimum and maximum values over all depths between ME $=$ -0.03 and 0.20, RMSE $= 0.42$ and 0.82, MEC $= 0.64$ and 0.90 and PICP $= 0.80$ and 0.93.

4. Discussion

4.1. Map accuracy using statistical validation strategies

The large differences depending on the external accuracy assessment strategy used (PFB-OOB, PFB-CV, LSK or LSK-SRS; Table 3) emphasize that map accuracy depends largely on the statistical validation approach. Different approaches can yield substantially different indications of the same map’s quality. Hence, the statistical validation strategy needs to be chosen carefully. Based on sampling theory (Cochran, 1977; de Gruijter et al., 2006; Gregoire and Valentine, 2007), maps should be validated with a design-based approach using a probability sample whenever possible (Brus et al., 2011). Therefore, LSK-SRS may be regarded as the best estimate of the “true” map accuracy in our study and is further advantageous because the CI95 also quantifies the accuracy of the estimated metrics (Fig. 8). Nonetheless, we acknowledge that even the uncertainty of design-based metrics (CI95) are themselves imperfect and prone to uncertainty, as numerical experiments using pseudo values have shown (Lagacherie et al., 2019).

However, a slight disadvantage of LSK-SRS in our study is that some observations had to be removed or averaged and some depth layers had no observations for specific strata (Table 4). The depth layer 0 cm to 5 cm was not evaluated and the metrics of depth layers 5 cm to 15 cm, 15 cm to 30 cm and 100 cm to 200 cm only pertain to between 95.66 % and 98.51 % of the Netherlands. This may be avoided in other studies by considering the soil in 3D space when planning the sampling design and deciding which target depth layers to map beforehand. Such an approach may be planned in a similar way as for design-based inference of spatio-temporal models, where the probability sample needs to include both locations and time (Brus, 2014). In the case of the Netherlands, the LSK sampling campaign was planned before the standard *GSM* depth layers were defined (Finke et al., 2001; Visschers et al., 2007).

Both LSK and LSK-SRS strategies suggest that soil pH maps are positively biased, i.e. systematically under-predicted pH, with ME = 0.05 to 0.17 (Fig. 8). Depth layers between 15 cm to 60 cm showed the largest bias. This may be due to the difference in distribution of the PFB (calibration) vs. LSK (validation) data (Fig. 1). The relatively large peak in observations around 4.5 pH in PFB in comparison to the peaks for LSK results in a lower average in the calibration data. Ensemble decision trees tend to predict the average well while performance decreases towards the tails of the distribution. This is most likely due to averaging of trees in the forest [e.g.] (Hengl et al., 2018). Hence, the calibrated QRF using PFB data possibly led to overall biased predictions at LSK locations. The values of ME for LSK-SRS according to sampling theory also indicate that the predictions are not only biased at validation locations, but for all of the Netherlands. Such a bias may be avoided by using a representative dataset of all of the Netherlands (e.g. probability sample) also for calibration, not only for validation. On one hand, another possible reason for the bias may be that PFB and LSK data originate from different time periods (Supplement S1, Fig. S1). On the other hand, field and lab methods and protocols remained the same for both PFB and LSK data, so a bias solely due to the year of soil sampling and analysis is unlikely.

We found that QRF was able to detect changes in pH over depth, as indicated by the qualitative evaluation of the spatial patterns (Fig. 2 and Section 3.2), and that accuracy slightly decreased over depth, as indicated by the design-based inference (Fig. 8). RMSE increased from 0.70 (5 cm to 15 cm) to 0.73–0.74 (15 cm to 100 cm) to 0.79 (100 cm to 200 cm). However, the MEC and ME did not indicate lower accuracy over depth. Performance often decreases with depth in DSM studies due to fewer observations and fewer covariates available indicative of soil conditions at lower depths (Keskin et al., 2018).

Using LSK-SRS as a reference, PFB-CV was most indicative of map quality out of the non-design-based inference strategies. RMSE, MEC and PICP values of PFB-CV were closest to those of LSK-SRS (Figs. 8 and 9). OOB validation for ensemble decision tree models without grouping soil profile locations overestimated map accuracy. This is supported by Meyer et al. (2018), who show that without leaving out all observations from entire locations, model accuracy metrics are overly optimistic. In contrast, validation using the independent dataset (LSK) without design-based inference was too pessimistic over all depth layers based on RMSE and MEC values (Fig. 8). This may be because when ignoring the probability sample, observations from small (niche) strata, where the predictive performance is likely worse, are oversampled and frequently occurring strata, where predictive performance is likely better, are undersampled in comparison to their relative occurrence in the study area. In summary, using either PFB-OOB or not accounting for the LSK sample design both resulted in misleading map accuracy metrics.

For other studies that do not have the resources to validate using a probability sample, the location-grouped *k*-fold CV used here (PFB-CV) may be further refined to estimate map accuracy. Random *k*-fold CV, even when grouped by location, potentially still leads to biased estimates since the data are often clustered or unevenly distributed (Brenning, 2012; Schratz, 2019). When data are clustered or unevenly distributed, data dense areas are weighed more than sparsely sampled areas in random *k*-fold CV. Thus, the CV indicates how accurate the model is at predicting the sampled data, but not necessarily the area. To overcome these challenges, we recommend to use weighted CV: a form of random *k*-fold CV where the dataset is resampled into multiple datasets based on point density (Van Ebbenhorst Tengbergen, 2021). Weighted CV was not tested here due to the availability of a probability sample to perform design-based inference (LSK-SRS). We do not recommend spatial partitioning, i.e. spatial CV, or the use of buffers in CV [e.g.] (Brenning, 2005; Brenning, 2012; Le Rest et al., 2014; Pohjankukka et al., 2017; Roberts et al., 2017; Ploton et al., 2020; Hengl et al., 2021), as these are not theoretically sound and are likely systematically and potentially severely over-pessimistic (Wadoux et al., 2021).

4.2. Comparison to other soil pH maps

The main improvements of our soil pH maps for the Netherlands compared to other *GSM* products are a better estimation of the “true” map accuracy using design-based inference and the high spatial resolution (25 m). Here, we compare our design-based LSK-SRS accuracy metrics with those of other studies. However, it is important to note that other studies did not use probability samples and design-based inference for validation and so results are not directly comparable. Accuracy metrics of studies using a non-probability sample may be considered overly optimistic in the case of clustered data (Wadoux et al., 2021).

Our maps have similar patterns as previous soil pH maps of the Netherlands. Although an initial map for soil pH from 0 cm to 25 cm was made for the Netherlands using a soil type map and co-kriging, mean predictions are difficult to compare because no statistical validation was done (Brus et al., 2007; Brus et al., 2009). Nevertheless, qualitative evaluation of the spatial patterns reveals strong resemblance of mean predictions [Fig. 4;][p. 19] (Brus et al., 2009).

For Denmark, another Northern European country similar in both size and soil variability to the Netherlands, Adhikari et al. (2014) used a hybrid model consisting of Cubist followed by local point kriging of the residuals and validated maps using 25% of samples. Results were only reported for the depth layer with the best performance (5 cm to 15 cm), for which an R^2 of 0.46 and RMSE of 0.61 were achieved. Our MEC value, which is comparable to R^2, is better (0.78) while our RMSE is higher (0.70) for the same depth layer using design-based inference.

On average, our results for map accuracy using design-based inference are also comparable to other recent *GSM* products that used ensemble decision trees (RF or QRF) to model soil pH. For example, Chen et al. (2019) predicted topsoil (0 cm to 20 cm) pH for China using RF and assessed the accuracy with a random 10-fold CV (not grouped by location). We attained comparable results for depths 5 cm to 30 cm in terms of RMSE (0.70 to 0.74) compared to their study (0.72), but our MEC values were higher than their R^2 values (MEC = 0.78–0.82 vs. 0.71).

We also compared our results to the recent SoilGrids version 2.0 (Poggio et al., 2021), for which we compared global metrics as well as prediction performance in the Netherlands. SoilGrids 2.0 also used QRF but assessed map accuracy using a CV procedure based on spatial stratification. When comparing global metrics, we attained better results for depths 5 cm to 100 cm (RMSE = 0.70 to 0.74; MEC = 0.73–0.82) compared to their values over all depths [RMSE = 0.77 pH (water), MEC = 0.66–0.69;] (Poggio et al., 2021), although for the depth layer 100 cm to 200 cm, our RMSE was higher (0.79). We attained identical PICP of PI90 results as SoilGrids 2.0 for depths 5 cm to 15 cm and 15 cm to 30 cm (0.91 and 0.90, respectively), but slightly poorer results for 30 cm to 200 cm (0.92 to 0.94 vs. 0.89 to 0.91). However, we achieved much higher performance, for example for the depth layer 5 cm to 15 cm (ME = 0.09, RMSE = 0.70, MEC = 0.78, PICP of PI90 = 0.91) compared to SoilGrids (ME = −0.74, RMSE = 1.23, MEC = 0.34, PICP of PI90 = 0.71) when evaluating prediction performance in the Netherlands using LSK-SRS design-based inference. These magnitudes of differences in predictive performance between our maps and SoilGrids for the Netherlands were consistent when comparing all depth layers. These results align with Mulder et al. (2016) and Chen et al. (2019), who compared their national products with the older version of SoilGrids (Hengl et al., 2017) and also achieved higher accuracy.

4.3. Uncertainty using QRF

The PICP obtained from the QRF quantiles reveal that PI's are generally too large and hence prediction uncertainty was overestimated. For applications, this means that users can be slightly more certain than indicated because the PI90 in reality appears to cover more than 90% of the observations (Figs. 8 and 9). Our maps of $q_{0.05}$ and $q_{0.95}$ for depths 0 cm to 30 cm (e.g. Fig. 4) match those of [p. 20] Brus et al. (2009), who used co-kriging based on a soil type map of the Netherlands. This suggests that in this case, the uncertainty quantification using QRF was similar to the kriging variance. However, other studies have often found large differences in the spatial distribution of the PI when comparing QRF and kriging (Vaysse et al., 2017; Baake, 2018; Szatmári et al., 2019).

A modelling approach using QRF can make use of the flexibility and predictive performance of machine learning while still attaining an estimate of prediction uncertainty in a general context. However, the limitations of prediction methods such as QRF are that they do not deliver knowledge of the different sources of this uncertainty. Recent studies have developed approaches to either quantify uncertainty of data used as model inputs in DSM, such as the measurement errors of soil observations (Van Leeuwen et al., 2021) or covariates, or how these errors can be incorporated in machine learning algorithms such as RF (van der Westhuizen et al., 2021).

There are many possible sources that may have contributed to the QRF prediction uncertainty, such as the inability of the covariates to explain all soil pH variation, lab measurement errors and the temporal variation of soil pH over time. We used legacy data from 1953–2012 (Supplement S1, Fig. S1), but ignored time even though soil pH may have changed over the decades. Even within one year, soil pH varies with season and soil moisture content, with higher pH values associated with wetter soils and winter conditions and lower pH values with drier soils and summer conditions (Miller and Kissel, 2010; Robinson et al., 2017). However, differences that might be expected due to soil pH temporal variation (e.g. 0.5 pH units) are only a major source of uncertainty in areas for which the PI90 is very low (e.g. AAA pixels). Therefore, such temporal variation is smaller than the vast majority of uncertainty quantified here. In agreement with Arrouays et al. (2017), there is a need to address soil measurement age and time in future DSM studies. Accounting for time may potentially improve prediction accuracy by removing this source of uncertainty. Moreover, it potentially estimates how pH has changed over time for different parts of the Netherlands.

4.4. Accuracy thresholds for Tier 4 GSM maps and user applications

Using the *GSM* accuracy thresholds for Tier 4 products for the PI90, the large majority of the Netherlands was designated A or "none" quality. We believe that there are several reasons for this. Firstly, as indicated by the PICP of PI90 (Figs. 7–9), the uncertainty quantification using QRF was overestimated, meaning that overall, slightly better accuracy threshold designation can be expected (e.g. less "none" and more AAA). Secondly, the accuracy thresholds are also dependent on the spatial support. Uncertainty of predictions at block support is typically smaller than for point predictions because within-block variation is averaged out. The degree of uncertainty reduction depends on the degree of within-block spatial variation and therefore uncertainty reduction by spatial aggregation can only be computed if the spatial correlation is included in the model, e.g. as in Szatmári et al. (2021). Thus, we expect more AAA areas with increasing spatial support. Thirdly and most importantly, the achievability of AA or AAA accuracy thresholds are largely dependent on the size and variability of soil observations in the study area. Higher accuracy thresholds can generally be achieved in study areas where there is less variation while mostly lower accuracy thresholds are achieved in areas where there is a large variation. If users require AA or AAA accuracy for their intended use, we recommend to conduct a local or regional mapping study where there is less variation (e.g. sandy soils) and to increase the sampling density.

Based on our results, AA and AAA thresholds are difficult to achieve for national maps and we are curious whether other countries will obtain similar results. We think it is important that accuracy thresholds remain ambitious because thresholds below "none" would imply such a high uncertainty that it would most likely be meaningless for user applications. We hope that our results will lead to a discussion that includes end-users about the uncertainty ranges of the *GSM* accuracy thresholds.

We believe that accuracy thresholds as used here have several advantages. Unlike statistical validation, which can only make use of observations at sampled locations, they are spatially explicit and can be designated to each pixel, as is also the case for other uncertainty measures (Heuvelink, 2014; Heuvelink, 2018). Moreover, we believe accuracy thresholds are easier to communicate with end-users than other widely used uncertainty metrics, e.g. PI90. A user merely has to know the quality required for their specific application and then look at the map of four possible thresholds. Note that maps are not only useful where there are high quality pixels; many users may only require e.g. A quality. National-scale measures, legislation or projects that are based on soil information can easily be applied specifically to areas above a certain threshold. For agricultural applications for example, our maps may potentially be useful in order to consider uncertainty for liming recommendations (Libohova et al., 2019). Lark et al. (2014) used a similar idea to communicate to users where critical trace element values might approach agronomically important thresholds.

Our modelling framework is convenient for a variety of users that require spatially explicit soil pH information and associated uncertainty with quantified accuracy at 25 m resolution for any desired depth anywhere in the Netherlands. If users are interested in the overall pH map accuracy for the Netherlands, we recommend to use the LSK-SRS (and PFB-CV if between 0 cm to 5 cm) metrics. If users are interested in a small target area within the Netherlands or require spatially explicit accuracy measures, then we recommend the use of PI90 and accuracy threshold maps. Given that the accuracy is within an acceptable range (for a given target area), the high resolution maps may be used for local and small-scale land use planning and management. In this regard, we hope that these soil pH maps are useful for the Dutch Ministries for Agriculture, Nature and Food Quality, Economic Affairs and Climate Policy, Infrastructure and Water Management, governmental waterboards, as well as farmers, researchers from different fields and nonprofit organisations. The reproducible, efficient and flexible computational workflow may also make it attractive to generate future maps of other target soil properties for the European Joint Program (*EJP*) on agricultural soil management of the European Union (Keesstra et al., 2021) or the Global Soil Partnership (*GSP*) of the Food and Agriculture Organization of the United Nations (FAO).

5. Conclusion

This study contributed to the *GSM* project by providing soil pH prediction maps for the Netherlands at 25 m resolution, at six standard depth layers (0 cm to 5 cm, 5 cm to 15 cm, 15 cm to 30 cm, 30 cm to 60 cm, 60 cm to 100 cm and 100 cm to 200 cm), yet the calibrated model allows prediction at any user-required depth. We compared non-design-based to design-based external accuracy assessment strategies using ME, RMSE, MEC, PI90 and PICP metrics. Among these statistical validation methods, the probability sample available in the Netherlands presented a unique opportunity for accuracy assessment using design-based inference (LSK-SRS). Consequently, we were able to provide unbiased estimates of the "true" map quality and quantify the accuracy of these estimates with confidence intervals. We used a robust, reproducible and data-driven DSM workflow that uses QRF to quantify spatially explicit uncertainty as an internal accuracy assessment. In addition, these are to our knowledge the first Tier 4 *GSM* maps, since they also provide spatially explicit accuracy thresholds as quality rankings. We attained A and "none" quality accuracy thresholds for the large majority of the Netherlands and therefore, we call upon future studies to also test whether the highest Tier 4 *GSM* quality rankings are difficult to achieve for national scale soil maps. We hope that our soil pH maps are useful for national agencies and initiatives and expect that with modest modification our workflow can be applied to other soil properties and other areas in the world to meet the increasing demand for spatial soil information.

Code and data availability

The code used to produce the results of this paper is available here: https://github.com/anatol-helfenstein/BIS-3D. The soil pH maps for six depth layers (0 cm to 5 cm, 5 cm to 15 cm, 15 cm to 30 cm, 30 cm to 60 cm, 60 cm to 100 cm and 100 cm to 200 cm) at 25 m resolution for the Netherlands, including mean predictions, $q_{0.05}$, $q_{0.50}$ (median), $q_{0.95}$, PI90 ($q_{0.95} - q_{0.05}$) estimates and accuracy threshold maps are openly accessible: https://doi.org/10.4121/16451739.v1 (Helfenstein et al., 2021). Soil data are available from the Dutch National Key Registry of the Subsurface (BRO, in Dutch) at https://bodemdata.nl/ or https://basisregistratieondergrond.nl/. More detailed data can be obtained from the BIS database from Wageningen Environmental Research: https://www.wur.nl/nl/Onderzoek-Resultaten/Onderzoeksinstituten/Environmental-Research/Faciliteiten-tools/Bodemkundig-Informatie-Systeem-BIS-Nederland.htm.

Video supplement

A video explaining the importance of these soil pH maps, this project and spatial soil information in general can be found here: https://www.youtube.com/watch?v=ENCYUnqc-wo

Author contributions

All authors contributed to the scientific research and writing of this paper. A.H. prepared the data and wrote all R scripts and sections of the paper. V.L.M., G.B.M.H. and J.P.O. provided valuable inputs and suggestions during all stages of the research, including conceptualisation of the modelling framework, interpretation of results and improving the writing.

Declaration of Competing Interest

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

Acknowledgement

We thank Dick J. Brus (Wageningen University) for providing help on calculating CI for model metrics of a probability sampling design, Laura Poggio (ISRIC - World Soil Information) for providing a tweaked function to retrieve mean predictions from QRF and Tijn L. van Orsouw (Wageningen University) for providing precipitation and temperature covariates. We would also like to thank Mirjam J.D. Hack-ten Broeke, Dennis J.J. Walvoort, Martin Knotters, Fokke Brouwer, Kees C. Teuling, Dorothee van Tol-Leenders and Willy J.M. de Groot of the Soil, Water and Land Use group (Wageningen Research) and Hugo L.E. de Groot (Wageningen Research) for their expert knowledge on Dutch soil information and the database (BIS). This project (WOT-04–013-010) was financed by the Dutch Ministry of Agriculture, Nature and Food Quality: https://research.wur.nl/en/projects/soil-property-mapping-wot-04-013-010.

Appendix A. *GSM* accuracy thresholds

Table A1.

Table A1
These tabular data are referred to in Arrouays et al. (2015) but to the best of our knowledge has not been published. The specifications for the response of this study, soil pH (multiplied by 10), is shown in bold.

Soil property	Unit	A	AA	AAA
		Depth 0 −5		
Depth to rock	cm	Mean ± 50% (Mean)	Mean ± 30% (Mean)	Mean ± 15% (Mean)
Plant Exploitable (Effective) Depth	cm	Mean ± 50% (Mean)	Mean ± 30% (Mean)	Mean ± 15% (Mean)
SOC	g/kg	Mean ± 60% (Mean)	Mean ± 35% (Mean)	Mean ± 15% (Mean)
pH x 10		**(±) 15**	**(±) 10**	**(±) 5**
Clay	g/kg	Mean ± 40% (Mean)	Mean ± 25% (Mean)	Mean ± 15% (Mean)
Silt	g/kg	Mean ± 40% (Mean)	Mean ± 25% (Mean)	Mean ± 15% (Mean)
Sand	g/kg	Mean ± 40% (Mean)	Mean ± 25% (Mean)	Mean ± 15% (Mean)
Coarse fragments	m^3/m^3	Mean ± 35% (Mean)	Mean ± 25% (Mean)	Mean ± 15% (Mean)
ECEC	$mmol_c/kg$	Mean ± 40% (Mean)	Mean ± 30% (Mean)	Mean ± 15% (Mean)
		Depth 5 −15		
Depth to rock	cm	Mean ± 50% (Mean)	Mean ± 30% (Mean)	Mean ± 15% (Mean)
Plant Exploitable (Effective) Depth	cm	Mean ± 50% (Mean)	Mean ± 30% (Mean)	Mean ± 15% (Mean)
SOC	g/kg	Mean ± 50% (Mean)	Mean ± 30% (Mean)	Mean ± 15% (Mean)
pH x 10		**(±) 15**	**(±) 10**	**(±) 5**
Clay	g/kg	Mean ± 40% (Mean)	Mean ± 25% (Mean)	Mean ± 15% (Mean)
Silt	g/kg	Mean ± 40% (Mean)	Mean ± 25% (Mean)	Mean ± 15% (Mean)
Sand	g/kg	Mean ± 40% (Mean)	Mean ± 25% (Mean)	Mean ± 15% (Mean)
Coarse fragments	m^3/m^3	Mean ± 40% (Mean)	Mean ± 30% (Mean)	Mean ± 15% (Mean)
ECEC	$mmol_c/kg$	Mean ± 40% (Mean)	Mean ± 30% (Mean)	Mean ± 15% (Mean)

(*continued on next page*)

Table A1 (*continued*)

Soil property	Unit	A	AA	AAA
Depth 0 −5				
Depth 15 −30				
Depth to rock	cm	Mean ± 50% (Mean)	Mean ± 30% (Mean)	Mean ± 15% (Mean)
Plant Exploitable (Effective) Depth	cm	Mean ± 50% (Mean)	Mean ± 30% (Mean)	Mean ± 15% (Mean)
SOC	g/kg	Mean ± 50% (Mean)	Mean ± 30% (Mean)	Mean ± 15% (Mean)
pH x 10		**(±) 15**	**(±) 10**	**(±) 5**
Clay	g/kg	Mean ± 40% (Mean)	Mean ± 25% (Mean)	Mean ± 15% (Mean)
Silt	g/kg	Mean ± 40% (Mean)	Mean ± 25% (Mean)	Mean ± 15% (Mean)
Sand	g/kg	Mean ± 40% (Mean)	Mean ± 25% (Mean)	Mean ± 15% (Mean)
Coarse fragments	m^3/m^3	Mean ± 40% (Mean)	Mean ± 30% (Mean)	Mean ± 15% (Mean)
ECEC	mmol$_c$/kg	Mean ± 40% (Mean)	Mean ± 30% (Mean)	Mean ± 15% (Mean)
Depth 30 −60				
Depth to rock	cm	Mean ± 50% (Mean)	Mean ± 30% (Mean)	Mean ± 15% (Mean)
Plant Exploitable (Effective) Depth	cm	Mean ± 50% (Mean)	Mean ± 30% (Mean)	Mean ± 15% (Mean)
SOC	g/kg	Mean ± 40% (Mean)	Mean ± 25% (Mean)	Mean ± 15% (Mean)
pH x 10		**(±) 15**	**(±) 10**	**(±) 5**
Clay	g/kg	Mean ± 40% (Mean)	Mean ± 25% (Mean)	Mean ± 15% (Mean)
Silt	g/kg	Mean ± 40% (Mean)	Mean ± 25% (Mean)	Mean ± 15% (Mean)
Sand	g/kg	Mean ± 40% (Mean)	Mean ± 25% (Mean)	Mean ± 15% (Mean)
Coarse fragments	m^3/m^3	Mean ± 40% (Mean)	Mean ± 30% (Mean)	Mean ± 15% (Mean)
ECEC	mmol$_c$/kg	Mean ± 40% (Mean)	Mean ± 30% (Mean)	Mean ± 15% (Mean)
Depth 60 −100				
Depth to rock	cm	Mean ± 50% (Mean)	Mean ± 30% (Mean)	Mean ± 15% (Mean)
Plant Exploitable (Effective) Depth	cm	Mean ± 50% (Mean)	Mean ± 30% (Mean)	Mean ± 15% (Mean)
SOC	g/kg	Mean ± 40% (Mean)	Mean ± 25% (Mean)	Mean ± 15% (Mean)
pH x 10		**(±) 15**	**(±) 10**	**(±) 5**
Clay	g/kg	Mean ± 40% (Mean)	Mean ± 25% (Mean)	Mean ± 15% (Mean)
Silt	g/kg	Mean ± 40% (Mean)	Mean ± 25% (Mean)	Mean ± 15% (Mean)
Sand	g/kg	Mean ± 40% (Mean)	Mean ± 25% (Mean)	Mean ± 15% (Mean)
Coarse fragments	m^3/m^3	Mean ± 50% (Mean)	Mean ± 40% (Mean)	Mean ± 20% (Mean)
ECEC	mmol$_c$/kg	Mean ± 40% (Mean)	Mean ± 30% (Mean)	Mean ± 15% (Mean)
Depth 100 −200				
Depth to rock	cm	Mean ± 50% (Mean)	Mean ± 30% (Mean)	Mean ± 15% (Mean)
Plant Exploitable (Effective) Depth	cm	Mean ± 50% (Mean)	Mean ± 30% (Mean)	Mean ± 15% (Mean)
SOC	g/kg	Mean ± 40% (Mean)	Mean ± 25% (Mean)	Mean ± 15% (Mean)
pH x 10		**(±) 15**	**(±) 10**	**(±) 5**
Clay	g/kg	Mean ± 40% (Mean)	Mean ± 25% (Mean)	Mean ± 15% (Mean)
Silt	g/kg	Mean ± 40% (Mean)	Mean ± 25% (Mean)	Mean ± 15% (Mean)
Sand	g/kg	Mean ± 40% (Mean)	Mean ± 25% (Mean)	Mean ± 15% (Mean)
Coarse fragments	m^3/m^3	Mean ± 50% (Mean)	Mean ± 40% (Mean)	Mean ± 20% (Mean)
ECEC	mmol$_c$/kg	Mean ± 40% (Mean)	Mean ± 30% (Mean)	Mean ± 15% (Mean)

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://doi.org/10.1016/j.geoderma.2021.115659.

References

Adhikari, K., Hartemink, A.E., Minasny, B., Kheir, R.B., Greve, M.B., Greve, M.H., 2014. Digital Mapping of Soil Organic Carbon Contents and Stocks in Denmark. PLOS ONE 9 (8), e105519.

Adhikari, K., Kheir, R.B., Greve, M.B., Greve, M.H., Malone, B.P., Minasny, B., McBratney, A.B., 2014. Mapping Soil pH and Bulk Density at Multiple Soil Depths in Denmark. CRC Press Taylor & Francis Group, Boca Raton, p. 6.

AHN, 2021. Actueel Hoogtebestand Nederland (AHN). https://www.ahn.nl/.

Akpa, S.I.C., Odeh, I.O.A., Bishop, T.F.A., Hartemink, A.E., 2014. Digital Mapping of Soil Particle-Size Fractions for Nigeria. Soil Sci. Soc. Am. J. 78 (6), 1953–1966.

Alterra, 2004. Historisch Grondgebruik Nederland (HGN). https://www.wur.nl/nl/show/Kaarten-Historisch-Grondgebruik-Nederland-HGN.htm.

Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B.M., Hong, S. Y., Lagacherie, P., Lelyk, G., McBratney, A.B., McKenzie, N.J., Mendonca-Santos, M. d. L., Minasny, B., Montanarella, L., Odeh, I.O.A., Sanchez, P.A., Thompson, J.A., Zhang, G.-L., Jan. 2014. Chapter Three - GlobalSoilMap: Toward a Fine-Resolution Global Grid of Soil Properties. In: Sparks, D.L. (Ed.), Advances in Agronomy. Vol. 125. Academic Press, pp. 93–134.

Arrouays, D., Leenaars, J.G.B., Richer-de-Forges, A.C., Adhikari, K., Ballabio, C., Greve, M., Grundy, M., Guerrero, E., Hempel, J., Hengl, T., Heuvelink, G.B.M., Batjes, N., Carvalho, E., Hartemink, A., Hewitt, A., Hong, S.-Y., Krasilnikov, P., Lagacherie, P., Lelyk, G., Libohova, Z., Lilly, A., McBratney, A., McKenzie, N., Vasquez, G.M., Mulder, V.L., Minasny, B., Montanarella, L., Odeh, I., Padarian, J., Poggio, L., Roudier, P., Saby, N., Savin, I., Searle, R., Solbovoy, V., Thompson, J., Smith, S., Sulaeman, Y., Vintila, R., Rossel, R.V., Wilson, P., Zhang, G.-L., Swerts, M., Oorts, K., Karklins, A., Feng, L., Ibelles Navarro, A.R., Levin, A., Laktionova, T., Dell'Acqua, M., Suvannang, N., Ruam, W., Prasad, J., Patil, N., Husnjak, S., Pásztor, L., Okx, J., Hallett, S., Keay, C., Farewell, T., Lilja, H., Juilleret, J., Marx, S., Takata, Y., Kazuyuki, Y., Mansuy, N., Panagos, P., Van Liedekerke, M., Skalsky, R., Sobocka, J., Kobza, J., Eftekhari, K., Alavipanah, S.K., Moussadek, R., Badraoui, M., Da Silva, M., Paterson, G., Gonçalves, M. d. C., Theocharopoulos, S., Yemefack, M., Tedou, S., Vrscaj, B., Grob, U., Kozák, J., Boruvka, L., Dobos, E., Taboada, M., Moretti, L., Rodriguez, D., Dec. 2017. Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. GeoResJ 14, 1–19.

Arrouays, D., McBratney, A., Bouma, J., Libohova, Z., Richer-de-Forges, A.C., Morgan, C. L.S., Roudier, P., Poggio, L., Mulder, V.L., 2020. Impressions of digital soil maps: The good, the not so good, and making them ever better. Geoderma Regional, e00255.

Arrouays, D., McBratney, A., Minasny, B., Hempel, J., Heuvelink, G.B.M., MacMillan, R. A., Hartemink, A., Lagacherie, P., McKenzie, N., 2015. The GlobalSoilMap project specifications. In: Proceedings of the 1st GlobalSoilMap Conference, pp. 9–12.

Arrouays, D., McKenzie, N., Hempel, J., de Forges, A.R., McBratney, A. (Eds.), 2014. GlobalSoilMap: Basis of the Global Spatial Soil Information System. CRC Press Taylor & Francis Group, Boca Raton.

Arrouays, D., McKenzie, N., Hempel, J., de Forges, A.R., McBratney, A., 2014. Preface. In: GlobalSoilMap: Basis of the Global Spatial Soil Information System. CRC Press Taylor & Francis Group, Boca Raton, p. p. xiii..

Baake, K., Apr. 2018. Quantifying Uncertainty of Random Forest Predictions: A Digital Soil Mapping Case Study. Thesis Report GIRS-2017-14, Wageningen University, Wageningen, the Netherlands.

Bakker, J., van Dessel, B., van Zadelhoff, F., 1989. Natuurwaardenkaart 1988: natuurgebieden, bossen en natte gronden in Nederland. No. 266862. s-Gravenhage SDU.

Batjes, N.H., 2016. Harmonized soil property values for broad-scale modelling (WISE30sec) with estimates of global soil carbon stocks. Geoderma 269, 61–68.

BIJ12, 2019. Informatiemodel Natuur (IMNa). https://www.bij12.nl/onderwerpen/natuur-en-landschap/digitale-keten-natuur-ketensamenwerking/informatiemodel-natuur-imna/.

Boehmke, B., Greenwell, B., 2020. Hands-On Machine Learning with R. Taylor & Francis.

Breiman, L., 2001. Random Forests. Mach. Learn. 45 (1), 5–32.

Brenning, A., 2005. Spatial prediction models for landslide hazards: Review, comparison and evaluation. Natural Hazards Earth Syst. Sci. 5 (6), 853–862.

Brenning, A., 2012. Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. In: 2012 IEEE International Geoscience and Remote Sensing Symposium, pp. 5372–5375.

BRT, 2020. Basisregistratie Topografie (BRT): Catalogus en Productspecificaties. Kadaster Versie 1.2.0.3, BRT.

BRT, 2021. TOPNL. https://www.kadaster.nl/zakelijk/producten/geo-informatie/topnl.

Brus, D.J., 2014. Statistical sampling approaches for soil monitoring. Eur. J. Soil Sci. 65 (6), 779–791.

Brus, D.J., 2019. Sampling for digital soil mapping: A tutorial supported by R scripts. Geoderma 338, 464–480.

Brus, D.J., Hengl, T., Heuvelink, G.B.M., Kempen, B., Mulder, V.L., Olmedo, G., Poggio, L., Ribeiro, E., Omuto, C., 2017. Carbon Mapping: GSOC Map Cookbook Manual. Food and Agriculture Organization of the United Nations, Rome.

Brus, D.J., Heuvelink, G.B.M., 2007. Towards a Soil Information System with quantified accuracy: Three approaches for stochastic simulation of soil maps. In: Statutory Research Tasks Unit for Nature and the Environment 58 Alterra, Wageningen.

Brus, D.J., Kempen, B., Heuvelink, G.B.M., 2011. Sampling for validation of digital soil maps. Eur. J. Soil Sci. 62 (3), 394–407.

Brus, D.J., Vašát, R., Heuvelink, G.B.M., Knotters, M., de Vries, F., Walvoort, D.J.J., 2009. Towards a Soil Information System with quantified accuracy. A prototype for mapping continuous soil properties. In: Statutory Research Tasks Unit for Nature and the Environment 197 Alterra, Wageningen.

Buringh, P., Stuer, G.G.L., Vink, P., 1962. Some techniques and methods of soil survey in the Netherlands. Neth. J. Agric. Sci. 10 (2), 17.

CBS, 2015. Bestand Bodemgebruik (BBG): 1993, 1996, 2000, 2003, 2006, 2008, 2010, 2012, 2015. Centraal Bureau voor de Statistiek (CBS). https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/korte-onderzoeksbeschrijvingen/bodemgebruik.

Chatfield, C., 1995. Model Uncertainty, Data Mining and Statistical Inference. J. R. Stat. Soc. Series A (Statistics in Society) 158 (3), 419–466.

Chen, S., Liang, Z., Webster, R., Zhang, G., Zhou, Y., Teng, H., Hu, B., Arrouays, D., Shi, Z., 2019. A high-resolution map of soil pH in China made by hybrid modelling of sparse soil data and environmental covariates and its implications for pollution. Sci. Total Environ. 655, 273–283.

Clement, J., 2001. GIS Vierde Bosstatistiek: Gebruikersdocumentatie, Documentatie van bestanden. Tech. rep., Research Instituut voor de Groene Ruimte, Alterra, Wageningen.

Cochran, W.G., 1977. Sampling Techniques, 3rd Edition. John Wiley & Sons, New York.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., Böhner, J., 2015. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. Geoscientific Model Development 8 (7), 1991–2007.

de Gruijter, J.J., Brus, D., Bierkens, M., Knotters, M., 2006. Sampling for Natural Resource Monitoring. Springer, The Netherlands.

de Vries, F., Al, E.J., 1992. De groeiplaatsgeschiktheid voor bosdoeltypen in beeld met ALBOS. Tech. Rep. 234, DLO-Staring Centrum.

Dharumarajan, S., Hegde, R., Janani, N., Singh, S.K., 2019. The need for digital soil mapping in India. Geoderma Regional 16, e00204.

Dharumarajan, S., Kalaiselvi, B., Suputhra, A., Lalitha, M., Hegde, R., Singh, S.K., Lagacherie, P., 2020. Digital soil mapping of key GlobalSoilMap properties in Northern Karnataka Plateau. Geoderma Regional 20, e00250.

Dokuchaev, V., 1899. Report to the Transcaucasian Statistical Committee on Land Evaluation in General and Especially for the Transcaucasia. Horizontal and Vertical Soil Zones. (In Russian.). Off. Press Civ, Affairs Commander-in-Chief Cacasus, Tiflis, Russia.

Domburg, P., de Gruijter, J.J., van Beek, P., 1997. Designing efficient soil survey schemes with a knowledge-based system using dynamic programming. Geoderma 75 (3), 183–201.

EEA, 2007. CLC2006 technical guidelines. EEA Technical Report 17. European Environment Agency (EEA), Copenhagen.

EEA, 2018. CORINE Land Cover — Copernicus Land Monitoring Service: 1986, 2000, 2006, 2012, 2018. European Environment Agency (EEA). https://land.copernicus.eu/pan-european/corine-land-cover.

EZK, 2013. Fysisch Geografische Regio's 2013; Ministerie van Economische Zaken en Klimaat (EZK; Ministry of Economic Affairs and Climate). https://nationaalgeoregister.nl/geonetwork/srv/dut/catalog.search#/metadata/c8b5668f-c354-42f3-aafc-d15ae54cf170.

EZK, 2019. Basisregistratie Gewaspercelen (BRP): 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019. Ministerie van Economische Zaken en Klimaat (EZK; Ministry of Economic Affairs and Climate), Agrarische Areaal Nederland. https://www.pdok.nl/introductie/-/article/basisregistratie-gewaspercelen-brp-.

FAO, Dec. 2017. Global Soil Organic Carbon (GSOC) Map. http://www.fao.org/global-soil-partnership/pillars-action/4-information-and-data-new/global-soil-organic-carbon-gsoc-map/en.

FAO, 2018. Soil Organic Carbon Mapping Cookbook, 2nd Edition. FAO, Rome, Italy.

Felix, R., 1995. Bodemkartering voor 1943—het geologisch perspectief. In: Buurman, P., Sevink, J. (Eds.), Van Bodemkaart Tot Informatiesysteem. Wageningen, Wageningen Pers, pp. 1–17.

Filippi, P., Jones, E.J., Bishop, T.F.A., 2020. Catchment-scale 3D mapping of depth to soil sodicity constraints through combining public and on-farm soil databases – A potential tool for on-farm management. Geoderma 374, 114396.

Filippi, P., Jones, E.J., Ginns, B.J., Whelan, B.M., Roth, G.W., Bishop, T.F.A., 2019. Mapping the Depth-to-Soil pH Constraint, and the Relationship with Cotton and Grain Yield at the Within-Field Scale. Agronomy 9 (5), 251.

Finke, P.A., de Gruijter, J.J., Visschers, R., 2001. Status 2001 Landelijke Steekproef Kaarteenheden en toepassingen; Gestructureerde bemonstering en karakterisering Nederlandse bodems. In: Alterra, Research Instituut voor de Groene Ruimte. Wageningen.

Gallant, J.C., Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resour. Res. 39 (12).

GDAL/OGR contributors, 2020. GDAL/OGR geospatial data abstraction software library. Open Source Geospatial Foundation (OSGeo).

Gomes, L.C., Faria, R.M., de Souza, E., Veloso, G.V., Schaefer, C.E.G.R., Filho, E.I.F., 2019. Modelling and mapping soil organic carbon stocks in Brazil. Geoderma 340, 337–350.

Gregoire, T.G., Valentine, H.T., 2007. Sampling Strategies for Natural Resources and the Environment. CRC Press, Boca Raton, USA.

Group, G.S.D.T., Dec. 2000. Global Gridded Surfaces of Selected Soil Characteristics (IGBP-DIS). ORNL DAAC, Oak Ridge, Tennessee, USA.

Gupta, S., Hengl, T., Lehmann, P., Bonetti, S., Or, D., 2020. SoilKsatDB: Global soil saturated hydraulic conductivity measurements for geoscience applications. Earth Syst. Sci. Data Discuss. 1–26.

Hartemink, A.E., Hempel, J., Lagacherie, P., McBratney, A., McKenzie, N., MacMillan, R. A., Minasny, B., Montanarella, L., de Mendonça Santos, M.L., Sanchez, P., Walsh, M., Zhang, G.-L., 2010. GlobalSoilMap.net – A New Digital Soil Map of the World. In: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), Digital Soil Mapping: Bridging Research, Environmental Application. and Operation. Progress in Soil Science. Springer, Netherlands, Dordrecht, pp. 423–428.

Hartemink, A.E., McBratney, A., 2008. A soil science renaissance. Geoderma 148 (2), 123–129.

Hartemink, A.E., Sonneveld, M.P.W., 2013. Soil maps of The Netherlands. Geoderma 204–205, 1–9.

Hazeu, G.W., de Wit, A.J.W., 2004. CORINE land cover database of the Netherlands: Monitoring land cover changes between 1986 and 2000. EARSeL eProceedings 3 (3), 382–387.

Hazeu, G.W., Vittek, M., Schuiling, R., Bulens, J.D., Storm, M.H., Roerink, G.J., Meijninger, W.M.L., 2020. LGN2018: een nieuwe weergave van het grondgebruik in Nederland. Tech. Rep. 3010, Wageningen Environmental Research, Wageningen.

Helfenstein, A., Mulder, V.L., Heuvelink, G.B.M., Okx, J.P., Sep. 2021. Tier 4 maps of soil pH at 25 m resolution for the Netherlands. 4TU.ResearchData. Dataset.

Hempel, J., Libohova, Z., Thompson, J., Odgers, N., Smith, C., Lelyk, G., Geraldo, G., 2014. GlobalSoilMap North American Node progress. In: Arrouays, D., McKenzie, N., Hempel, J., de Forges, A., McBratney, A. (Eds.), GlobalSoilMap. CRC Press, pp. 41–45.

Hengl, T., de Jesus, J.M., Heuvelink, G.B.M., Gonzalez, M.R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLOS ONE 12 (2), e0169748.

Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B.M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Gonzalez, M.R., 2014. SoilGrids1km — Global Soil Information Based on Automated Mapping. PLOS ONE 9 (8), e105992.

Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., Mendes de Jesus, J., Tamene, L., Tondoh, J.E., 2015. Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. PLOS ONE 10 (6), e0125814.

Hengl, T., MacMillan, R.A., 2019. Predictive Soil Mapping with R. OpenGeoHub foundation. Wageningen, the Netherlands.

Hengl, T., Miller, M.A.E., Križan, J., Shepherd, K.D., Sila, A., Kilibarda, M., Antonijević, O., Glušica, L., Dobermann, A., Haefele, S.M., McGrath, S.P., Acquah, G. E., Collinson, J., Parente, L., Sheykhmousa, M., Saito, K., Johnson, J.-M., Chamberlin, J., Silatsa, F.B.T., Yemefack, M., Wendt, J., MacMillan, R.A., Wheeler, I., Crouch, J., 2021. African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. Sci. Rep. 11 (1), 6130.

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B.M., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6, e5518.

Heuvelink, G.B.M., 2014. Uncertainty Quantification of GlobalSoilMap Products. CRC Press 335–340.

Heuvelink, G.B.M., 2018. Uncertainty and Uncertainty Propagation in Soil Mapping and Modelling. In: McBratney, A.B., Minasny, B., Stockmann, U. (Eds.), Pedometrics. Springer International Publishing, Cham, Progress in Soil Science, pp. 439–461.

Hijmans, R.J., 2020. Geographic Data Analysis and Modeling: Package 'raster'. CRAN.

Hijmans, R.J., 2021. Spatial Data Analysis: Package 'terra'. CRAN.

IenM, TNO, Jun. 2017. Basisregistratie Ondergrond (BRO) Catalogus: Booronderzoek, Bodemkundige boormonsterbeschrijving. Tech. Rep. Versie 1.0, Ministrie van Infrastructuur en Milieu (IenM), Geologische Dienst Nederland (TNO), Den Haag.

Janitza, S., Binder, H., Boulesteix, A.-L., 2016. Pitfalls of hypothesis tests and model selection on bootstrap samples: Causes and consequences in biometrical applications. Biometr. J. 58 (3), 447–473.

Jenny, H., 1941. Factors of Soil Formation: A System of Quantitative Pedology. McGraw-Hill, New York.

Keesstra, S.D., Munkholm, L., Cornu, S., Visser, S., Faber, J., Kuikman, P., Thorsoe, M., de Haan, J., Vervuurt, W., Verhagen, J., Neumann, M., Fantappie, M., van Egmond, F., Bispo, A., Wall, D., Berggreen, L., Barron, J., Gascuel, C., Granjou, C., Gerasina, R., Chenu, C., Mar. 2021. Deliverable 2.4: Roadmap for the European Joint Programme SOIL. WP2 D2.4, European Joint Project COFUND.

Kempen, B., Brus, D.J., Heuvelink, G.B.M., 2012. Soil type mapping using the generalised linear geostatistical model: A case study in a Dutch cultivated peatland. Geoderma 189–190, 540–553.

Kempen, B., Brus, D.J., Heuvelink, G.B.M., Stoorvogel, J.J., 2009. Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. Geoderma 151 (3), 311–326.

Kempen, B., Brus, D.J., Stoorvogel, J.J., 2011. Three-dimensional mapping of soil organic matter content using soil type–specific depth functions. Geoderma 162 (1), 107–123.

Kempen, B., Heuvelink, G.B.M., Brus, D., Walvoort, D., 2014. Towards GlobalSoilMap.net products for The Netherlands. In: GlobalSoilMap: Basis of the Global Spatial Soil Information System – Proceedings of the 1st GlobalSoilMap Conference, pp. 85–90.

Keskin, H., Grunwald, S., 2018. Regression kriging as a workhorse in the digital soil mapper's toolbox. Geoderma 326, 22–41.

Keskin, H., Grunwald, S., Harris, W.G., 2019. Digital mapping of soil carbon fractions with machine learning. Geoderma 339, 40–58.

KNMI, May 2020. Koninklijk Nederlands Meteorologisch Instituut (KNMI) Dataplatform. https://www.knmidata.nl/.

Koomen, A., Maas, G., 2004. Geomorfologische Kaart Nederland (GKN); Achtergronddocument bij het landsdekkende digitale bestand. Altera-Rapport 1039, Alterra, Wageningen.

Kramer, H., Clement, J., 2015. Basiskaart Natuur 2013: Een landsdekkend basisbestand voor de terrestrische natuur in Nederland. WOt-technical report 41, Wettelijke Onderzoekstaken Natuur & Milieu. Wageningen.

KRW, 2004. Kaderrichtlijn Water (KRW) Grote grondwaterlichamen 2004.

Kuhn, M., Mar. 2019. The caret Package. https://topepo.github.io/caret/.

Kuhn, M., 2020. Classification and Regression Training: Package 'caret'. CRAN.

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

Lacoste, M., Minasny, B., McBratney, A., Michot, D., Viaud, V., Walter, C., 2014. High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape. Geoderma 213, 296–311.

Lagacherie, P., Arrouays, D., Bourennane, H., Gomez, C., Martin, M., Saby, N.P.A., 2019. How far can the uncertainty on a Digital Soil Map be known?: A numerical experiment using pseudo values of clay content obtained from Vis-SWIR hyperspectral imagery. Geoderma 337, 1320–1328.

Lagacherie, P., Arrouays, D., Bourennane, H., Gomez, C., Nkuba-Kasanda, L., 2020. Analysing the impact of soil spatial sampling on the performances of Digital Soil Mapping models and their evaluation: A numerical experiment on Quantile Random Forest using clay contents obtained from Vis-NIR-SWIR hyperspectral imagery. Geoderma 375, 114503.

Lamigueiro, O.P., Hijmans, R.J., 2021. Visualization Methods for Raster Data: Package 'rasterVis'. CRAN.

Lark, R.M., Ander, E.L., Cave, M.R., Knights, K.V., Glennon, M.M., Scanlon, R.P., 2014. Mapping trace element deficiency by cokriging from regional geochemical soil data: A case study on cobalt for grazing sheep in Ireland. Geoderma 226–227, 64–78.

Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J., Bretagnolle, V., 2014. Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. Glob. Ecol. Biogeogr. 23 (7), 811–820.

Liang, Z., Chen, S., Yang, Y., Zhou, Y., Shi, Z., 2019. High-resolution three-dimensional mapping of soil organic carbon in China: Effects of SoilGrids products on national modeling. Sci. Total Environ. 685, 480–489.

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

Liu, F., Zhang, G.-L., Song, X., Li, D., Zhao, Y., Yang, J., Wu, H., Yang, F., 2020. High-resolution and three-dimensional mapping of soil texture of China. Geoderma 361, 114061.

Ma, Y., Minasny, B., McBratney, A., Poggio, L., Fajardo, M., 2021. Predicting soil properties in 3D: Should depth be a covariate? Geoderma 383, 114794.

Maas, G., van der Meij, M., Delft, S., Heidema, A., 2019. Toelichting bij de legenda Geomorfologische kaart van Nederland 1:50 000 (2019). Wageningen Environmental Research, Wageningen.

McBratney, A., Mendonça Santos, M., Minasny, B., 2003. On digital soil mapping. Geoderma 117 (1–2), 3–52.

Meinshausen, N., 2006. Quantile Regression Forests. J. Mach. Learn. Res. 7, 17.

Meyer, H., Feb. 2021. 'caret' Applications for Spatial-Temporal Models: Package 'CAST'. CRAN.

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., Nauss, T., 2018. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environ. Modell. Software 101, 1–9.

Miller, R.O., Kissel, D.E., 2010. Comparison of Soil pH Methods on Soils of North America. Soil Sci. Soc. Am. J. 74 (1), 310–316.

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

Mulder, V.L., Lacoste, M., Richer-de-Forges, A.C., Martin, M.P., Arrouays, D., 2016. National versus global modelling the 3D distribution of soil organic carbon in mainland France. Geoderma 263, 16–34.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — A discussion of principles. J. Hydrol. 10 (3), 282–290.

Nussbaum, M., Walthert, L., Fraefel, M., Greiner, L., Papritz, A., 2017. Mapping of soil properties at high resolution in Switzerland using boosted geoadditive models. SOIL 3 (4), 191–210.

Papadopoulos, G., Edwards, P., Murray, A., 2001. Confidence estimation methods for neural networks: A practical comparison. IEEE Trans. Neural Networks 12 (6), 1278–1287.

Piikki, K., Wetterlind, J., Söderström, M., Stenberg, B., 2021. Perspectives on validation in digital soil mapping of continuous attributes—A review. Soil Use Manage. 37 (1), 7–21.

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., Pélissier, R., 2020. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nat. Commun. 11 (1), 4540.

Poggio, L., de Sousa, L.M., Batjes, N.H., Heuvelink, G.B.M., Kempen, B., Ribeiro, E., Rossiter, D., 2021. SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. SOIL 7 (1), 217–240.

Poggio, L., Gimona, A., 2017. 3D mapping of soil texture in Scotland. Geoderma Regional 9, 5–16.

Pohjankukka, J., Pahikkala, T., Nevalainen, P., Heikkonen, J., 2017. Estimating the prediction performance of spatial models via spatial k-fold cross validation. Int. J. Geogr. Inform. Sci. 31 (10), 2001–2019.

QGIS Development Team, 2021. QGIS: A Free and Open Source Geographic Information System.

R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Ramcharan, A., Hengl, T., Nauman, T., Brungard, C., Waltman, S., Wills, S., Thompson, J., 2018. Soil Property and Class Maps of the Conterminous United States at 100-Meter Spatial Resolution. Soil Sci. Soc. Am. J. 82 (1), 186–201.

RIVM, 2020. Grootschalige Concentratie- en Depositiekaarten Nederland (GCN, GDN), Rijksinstituut voor Volksgezondheid en Milieu (RIVM). https://www.rivm.nl/gcn-gdn-kaarten.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40 (8), 913–929.

Robinson, N.J., Benke, K.K., Norng, S., Kitching, M., Crawford, D.M., 2017. Improving the information content in soil pH maps: A case study. Eur. J. Soil Sci. 68 (5), 592–604.

Sanders, M.E., Prins, A.H., 2001. Provinciaal natuurbeleid: kwaliteitsdoelen voor de Ecologische Hoofdstructuur.

Schratz, P., 2019. Handling of Spatial Data. https://mlr.mlr-org.com/articles/tutorial/handling_of_spatial_data.html.

Scull, P., Franklin, J., Chadwick, O.A., McArthur, D., 2003. Predictive soil mapping: A review. Progr. Phys. Geogr.: Earth Environ. 27 (2), 171–197.

Stoorvogel, J.J., Bakkenes, M., Temme, A.J.A.M., Batjes, N.H., ten Brink, B.J.E., 2017. S-World: A Global Soil Map for Environmental Modelling. Land Degradation Devel. 28 (1), 22–33.

Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8 (1), 25.

Subburayalu, S.K., Slater, B.K., 2013. Soil Series Mapping By Knowledge Discovery from an Ohio County Soil Map. Soil Sci. Soc. Am. J. 77 (4), 1254–1268.

Szatmári, G., Pásztor, L., 2019. Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms. Geoderma 337, 1329–1340.

Szatmári, G., Pásztor, L., Heuvelink, G.B.M., 2021. Estimating soil organic carbon stock change at multiple scales using machine learning and multivariate geostatistics. Geoderma 403, 115356.

Thunnissen, H. a. M., van Middelaar, H.J., 1995. The CORINE Land Cover database of the Netherlands; final report of the CORINE Land Cover project in the Netherlands. Tech. Rep. 78, SC-DLO, Wageningen.

TNO, G.D.N., 2020. BROloket: Ondergrondgegevens. https://www.broloket.nl/ondergrondgegevens.

Van den Berg, F., Tiktak, A., Hoogland, T., Poot, A., Boesten, J., van der Linden, A.M.A., Pol, J.W., 2017. An improved soil organic matter map for GeoPEARL_NL: Model description of version 4.4.4 and consequence for the Dutch decision tree on leaching to groundwater. Tech. rep.. Wageningen Environmental Research (Alterra), Wageningen.

Van der Meulen, M., Doornenbal, J., Gunnink, J., Stafleu, J., Schokker, J., Vernes, R., van Geer, F., van Gessel, S., van Heteren, S., van Leeuwen, R., Bakker, M., Bogaard, P., Busschers, F., Griffioen, J., Gruijters, S., Kiden, P., Schroot, B., Simmelink, H., van Berkel, W., van der Krogt, R., Westerhoff, W., van Daalen, T., 2013. 3D geology in a 2D country: Perspectives for geological surveying in the Netherlands. Netherlands J. Geosci. – Geologie en Mijnbouw 92 (4), 217–241.

van der Westhuizen, S., Heuvelink, G.B.M., Hofmeyr, D., Apr. 2021. Measurement error-filtered machine learning in digital soil mapping. In: EGU21-9704. Copernicus Meetings, online, p. 1.

Van Ebbenhorst Tengbergen, T., Feb. 2021. Critical evaluation and improvement of cross-validation strategies for accuracy assessment of digital soil maps. Thesis Report GIRS-2021-13, Wageningen University, Wageningen.

Van Leeuwen, C.C.E., Mulder, V.L., Batjes, N.H., Heuvelink, G.B.M., 2021. Statistical modelling of measurement error in wet chemistry soil data. Eur. J. Soil Sci. 1–17.

Vaysse, K., Lagacherie, P., 2017. Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma 291, 55–64.

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

Visschers, R., Finke, P.A., de Gruijter, J.J., 2007. A soil sampling program for the Netherlands. Geoderma 139 (1), 60–72.

Vos, P., 2015. Origin of the Dutch Coastal Landscape: Long-Term Landscape Evolution of the Netherlands during the Holocene, Described and Visualized in National, Regional and Local Palaeogeographical Map Series. Barkhuis, Groningen.

Vos, P., v. d. Meulen, M., Weerts, H., Bazelmans, J., 2020. Atlas of the Holocene Netherlands, landscape and habitation since the last ice age. University Press, Amsterdam.

Wadoux, A.M.J.C., Heuvelink, G.B.M., de Bruin, S., Brus, D.J., 2021. Spatial cross-validation is not the right way to evaluate map accuracy. Ecol. Model. 457, 109692.

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

Wadoux, A.M.J.C., Minasny, B., McBratney, A.B., 2020. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth Sci. Rev. 210, 103359.

Wallig, M., Microsoft, Weston, S., Oct. 2020. Provides Foreach Looping Construct: Package 'foreach'. CRAN.

Wallig, M., Corporation, Microsoft, Weston, S., 2020. Foreach Parallel Adaptor for the 'parallel' Package: Package 'doParallel'. CRAN.

Walvoort, D., Hoogland, T., Jul. 2017. Metadata for the Dutch contribution to the Global Soil Organic Carbon (GSOC) map.

Webster, R., Oliver, M.A., 2007. Geostatistics for Environmental Scientists, 2nd Edition. John Wiley & Sons Ltd, Chichester.

Weil, R., Brady, N., 2017. The Nature and Properties of Soils. fifteenth Edition. Pearson Education.

WENR, 2020. Landelijk Grondgebruik Nederland (LGN). https://www.wur.nl/nl/Onderzoek-Resultaten/Onderzoeksinstituten/Environmental-Research/Faciliteiten-tools/Kaarten-en-GIS-bestanden/Landelijk-Grondgebruik-Nederland.htm.

Wilks, D., 2011. Chapter 8: Forecast Verification. In: Statistical Methods in the Atmospheric Sciences, 3rd Edition. Vol. 100. Elsevier, pp. 301–394.

Wood, J., 1996. The geomorphological characterisation of Digital Elevation Models. Ph. D. thesis, University of Leicester.

Wood, J., Jan. 2009. Chapter 14 Geomorphometry in LandSerf. In: Hengl, T., Reuter, H.I. (Eds.), Developments in Soil Science. Vol. 33 of Geomorphometry. Elsevier, pp. 333–349.

Wright, M.N., Ziegler, A., 2017. Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J. Stat. Softw. 77 (1).

Zhang, Y., Ji, W., Saurette, D.D., Easher, T.H., Li, H., Shi, Z., Adamchuk, V.I., Biswas, A., 2020. Three-dimensional digital soil mapping of multiple soil properties at a field-scale using regression kriging. Geoderma 366, 114253.