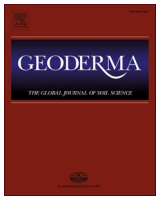
Geoderma 450 (2024) 117052

Contents lists available at ScienceDirect

Geoderma

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



Interpreting and evaluating digital soil mapping prediction uncertainty: A case study using texture from SoilGrids

Linda Lilburne [a,*], Anatol Helfenstein [b,c], Gerard B.M. Heuvelink [b,d], Andre Eger [a]

[a] Manaaki Whenua – Landcare Research, PO Box 69040, Lincoln 7640, New Zealand
[b] Soil Geography and Landscape Group, Wageningen University, PO Box 47, 6700 AA Wageningen, the Netherlands
[c] Soil, Water and Land Use Team, Wageningen Environmental Research, PO Box 47, 6700 AA Wageningen, the Netherlands
[d] ISRIC-World Soil Information, Wageningen, the Netherlands

ARTICLE INFO

Handling Editor: L. Morgan Cristine

Keywords:
Uncertainty
Soil texture
SoilGrids
Accuracy assessment

ABSTRACT

Soil information is critical for a wide range of land resource and environmental decisions. These decisions will be compromised when the soil information quality is unsatisfactory. Thus, users of soil information need to understand and consider the uncertainty of the available soil information and be able to judge whether it is fit for purpose. The uncertainty information provided with the SoilGrids 2.0 product was examined in a case study. We hypothesised that the soil property predictions for the Netherlands (NL) might be less uncertain than those of New Zealand (NZ) because there were more relevant training data for NL than for NZ. The study objectives were to: 1) understand whether the provided uncertainty information is correct for both countries; 2) explore spatial patterns and relationships in the prediction error and uncertainty information using quantitative tools and new graphical analyses; 3) analyse whether these patterns and relations can be explained; and 4) explore how the uncertainty information and insights derived from graphical analyses might assist an end user to determine whether a map is suitable for their purpose. The study focused on soil texture.

Independent datasets showed that the SoilGrids 2.0 uncertainty information was too optimistic for sand and too pessimistic for clay for both countries. The graphical analyses confirmed the initial assumption that NL predictions were more accurate than those for NZ, but they also indicated that some locations in NL have high uncertainty. The graphical analyses allowed only a limited identification of the four sources of uncertainty in digital soil maps, but were quite insightful in helping us to better understand the reliability of the information. A set of recommendations was developed for both producers and consumers of digital soil mapping (DSM) products. This includes the provision of a summary map of accuracy classes. We suggest that more research and educational effort is needed to ensure that digital soil maps are used appropriately.

1. Introduction

Soil information is critical for a wide range of land resource and environmental decisions, including progressing the United Nations Sustainable Development Goals (Keesstra et al., 2016). Decisions will be compromised when the soil information quality is unsatisfactory. Users of soil information need to understand and consider the uncertainty of the available soil information, and be able to judge whether it is fit for purpose.

Digital soil mapping (DSM)[1] has become a widely used approach to generate maps of soil properties at a range of scales, using machine learning and other statistical methods (McBratney et al., 2003; Minasny and McBratney, 2016). The general approach is to derive a model that predicts a soil property of interest based on relationships between the soil property and environmental covariates. The model is calibrated based on a training dataset of paired observations of the soil property and covariate data. Predictions are made based on the calibrated model and covariate values at prediction locations. In the case of geostatistical interpolation, the model also benefits from soil training data in the neighbourhood of the prediction location.

No map is perfect and this also holds for maps obtained with DSM. DSM maps have varying levels of predictive accuracy, depending on the following four sources of uncertainty:

* Corresponding author.
E-mail address: lilburnel@landcareresearch.co.nz (L. Lilburne).

[1] DSM – digital soil mapping.

https://doi.org/10.1016/j.geoderma.2024.117052
Received 20 July 2024; Received in revised form 22 September 2024; Accepted 26 September 2024
Available online 4 October 2024

1. There may be errors in the training data, including errors in the measured values and measurement locations.
2. The training dataset may be too small, spatially biased, or otherwise limited for optimal model calibration.
3. The available covariates may not fully explain variation in the soil property of interest.
4. The DSM model structure may not be able to fully represent the relationship between the covariates and the soil property.

These four sources capture all uncertainty, because if the covariates explain all soil spatial variation, if the model can capture all the information from the covariates, and if sufficient, error-free training data are available, then the model predictions must be error-free.

Since DSM maps are not perfect it is important to quantify their accuracy (Heuvelink, 2018). This is often achieved by withholding some of the data and using these as a test dataset, or by using a cross-validation procedure where the model is developed multiple times on different splits of the dataset and evaluated on the splits not used for model calibration. Sometimes a different, independent data set is used for evaluation, which has the advantage that these data can be truly unseen by the modeller and hence prevent data leakage (Kaufman et al., 2011). For example, data leakage occurs when modellers apply cross-validation on multiple models, until they get a well-performing model. In such a case they are effectively using the test data for calibration, which means that these data can no longer be used for evaluation. Ideally, the independent data are collected using probability sampling to ensure unbiased estimates of accuracy metrics and to compute confidence intervals of these estimates (Brus et al., 2011). However, in practice data are often collated from legacy (i.e., historical) datasets without a coherent sampling design.

Common metrics to quantify map accuracy of quantitative variables are the Mean Error (ME), Root Mean Squared Error (RMSE) and Model Efficiency Coefficient (MEC). Evaluation of map accuracy of categorical variables employs other metrics based on the confusion matrix. These are all summary metrics that assess the map accuracy and thus map uncertainty, across the entire study area. Wadoux et al. (2022) suggested an integrated approach of combining accuracy metrics through Taylor and solar diagrams.

Many DSM models allow for spatially explicit estimates of uncertainty in addition to summary metrics of accuracy derived from test data. For instance, kriging quantifies the prediction uncertainty at every prediction location through the kriging standard deviation (Webster and Oliver, 2007), while quantile regression forest (QRF, Meinshausen, 2006) does this by specifying the quantiles of the conditional predictive distribution. Prediction intervals (PI) at each prediction location can easily be computed from the quantiles, where the PI (its upper and lower limits) is defined in terms of a percentage probability. PIs can also be derived from the kriging standard deviation if a parametric distribution of the map error is assumed. For instance, under the normal distribution assumption the limits of the symmetric PI_{90}, which is an interval within which the true value is expected to fall 90 % of the time, is given by subtracting or adding 1.64 times the kriging standard deviation from or to the prediction. Another option to provide information on prediction uncertainty is using relative prediction interval ratios (PIR), being the ratio of a PI width to the median prediction.

The validity of the PI estimates can also be evaluated using independent test data, for example with the prediction interval coverage probability (PICP) (Shrestha and Solomatine, 2006). This metric is the proportion of the test data that fall within a specified PI. In the case of the PI_{90}, the PICP should be close to 90 %. PICP numbers greater than 90 mean that the PI_{90} is too pessimistic (the interval should be narrower), whereas numbers smaller than 90 indicate the PI_{90} is too optimistic – it should be wider. Reliability plots (also called accuracy plots) show the PICP for a large number of percentage probabilities (Goovaerts, 2001; Shrestha and Solomatine, 2006). To avoid the problem of one-sided bias, Schmidinger and Heuvelink (2023) suggested some additional metrics (quantile coverage probability, probability integral transform histograms, interval score and continuous ranked probability score).

Many studies have evaluated or compared DSM prediction layers. For example, Piikki et al. (2021) reviewed 188 peer-reviewed DSM papers, where 97 % of the studies included some type of map evaluation. About one-third of the studies estimated uncertainty, half of these as measures of spread, and a quarter using PIs. Of the 18 studies that provided PIs, 8 calculated the PICP. Other studies that have examined the reliability of spatial uncertainty information in DSM include Malone et al. (2011), Vaysse and Lagacherie (2017), Kasraei et al. (2021), Helfenstein et al. (2022), and Schmidinger and Heuvelink (2023). Szatmári and Pásztor (2019) compared the uncertainty estimates from four DSM methods, finding poor estimates of the uncertainty. While some attention has been given to the uncertainty information and its reliability, there remain many challenges in quantifying uncertainty of DSM products, and assessing and interpreting their reliability.

In this study, we evaluated the available uncertainty information for SoilGrids 2.0 (Poggio et al., 2021), a global DSM model that predicts basic soil properties at six standard depths at 250 m resolution. For the rest of the paper SoilGrids refers to SoilGrids 2.0 unless otherwise indicated. SoilGrids also quantifies the prediction uncertainty. We used two countries, the Netherlands (NL) and New Zealand (NZ), with different levels of training data and compared the SoilGrids prediction uncertainty of the two countries. We hypothesised that the soil property predictions for NL might be less uncertain than those of NZ because there were more relevant training data than for New Zealand.

The study objectives were to: 1) understand whether the provided uncertainty information is correct for both countries; 2) explore spatial patterns and relationships in the prediction error and uncertainty information using quantitative tools and new graphical analyses; 3) analyse whether these patterns and relations can be explained; and 4) explore how the uncertainty information and insights derived from graphical analyses might assist an end-user to determine whether the map is suitable for their purpose. These objectives were addressed by considering the spatial uncertainty information provided in the SoilGrids maps, by evaluating the DSM model performance using the SoilGrids training data, and by assessing the predictions and prediction uncertainty with independent datasets from NL and NZ. We focused on soil texture (clay, silt, sand content).

2. Materials and methods

2.1. Case study locations

The Netherlands with an area of 33,500 km^2 is located in the Rhine-Meuse-Scheldt delta in northwestern Europe. About one-third of the country lies below sea level, and is associated with loamy, clayey and peaty soils from Holocene marine and fluvial deposits. Dominated by sandy textures, the more elevated landforms that occupy the east and southeast of the country (maximum elevation 323 m) are of Pleistocene glacial and periglacial origin, often covered by aeolian or fluvial sands or minor amounts of loess. There is almost no occurrence of bedrock reaching the surface (de Bakker and Schelling, 1989; Edelmann, 1950; Hartemink, 2006). More than 90 % of all soils have groundwater within 140 cm of the surface in winter. Five soil orders are recognised in the Dutch soil classification system (Hartemink, 2006), of which the equivalents of Podzols/Spodosols and Luvisols/Alfisols (IUSS Working Group WRB, 2015; Soil Survey Staff, 1999) are the most pedogenically advanced soil types. The NL has a mild maritime climate with moderately warm summers, cool winters and typically high humidity (de Bakker and Schelling, 1989; Edelmann, 1950; Hartemink, 2006). The main land use types are agriculture (48 %), water (20 %), built-up areas (17 %), nature areas (8 %) and forests (7 %) (Hazeu et al., 2020). The human impact on soils in the Netherlands has been significant. In terms of the relative land surface area, peatlands have reduced from 50 % to 15 % (Vos et al., 2020; Erkens et al., 2016), 17 % is land that has been

historically reclaimed from water, 15 % has been converted to urban areas, and agricultural practices are widely considered among the most intensive in Europe (Debonne et al., 2022). A map of the main soil texture classes in the Netherlands is shown in Fig. 1.

New Zealand is a group of islands in the Pacific Ocean with a land area of 268,000 km^2. Located at a major plate boundary, it is both tectonically and volcanically active. Extending across 14° of latitude and reaching from sea level to > 3000 m altitude, the variability of the soil forming factors is large and soil diversity is high with all soil orders of Soil Taxonomy represented. Similar to the Netherlands, climate is generally temperate maritime, but variability in NZ is greater due to the 13° latitude of north-to-south extension of the main islands, and the effects of the axial mountain ranges, both with respect to temperature (i.e., lower temperatures with increasing altitude) and rainfall/humidity (i.e., orographic barriers against the prevailing westerly winds). Except for some geomorphically quiescent regions that experienced lower tectonic activity and less severe climate perturbations during the Quaternary and thus feature highly weathered soils, most soils started forming in the late Pleistocene and during the Holocene or are constantly rejuvenated by erosion. Unconsolidated, allochthonous deposits of Quaternary age are important parent materials and comprise alluvial, colluvial, aeolian, (peri)glacial, tephric, organic and marine deposits. Human land-use started in the 13th century with the arrival of the first Polynesian settlers but widespread agriculture only started in the late 1800 s. Before this, most of the country below tree line was covered by mixed conifer-broadleaf-hardwood-southern beech forests (Hewitt et al., 2021). Fig. 1 shows a map of the major soil texture classes in New Zealand.

While the natural environments and the degree of human impact in both countries are very different, the soils in both NL and NZ are mainly the result of soil formation processes under Late Pleistocene and Holocene environmental conditions of the mid-latitudes.

2.2. Data

2.2.1. SoilGrids texture maps

SoilGrids 2.0[2] is a system for global digital soil mapping that makes use of global soil profile information and covariate data to model the spatial distribution of soil properties across the globe (Poggio et al., 2021). It uses the quantile regression forest (QRF) method to predict various basic soil properties and the associated prediction uncertainty at six standard depths and at 250 m resolution. In this study, we obtained the mean and median predictions for sand, silt and clay at six standard depths for NL and NZ, as well as the 0.05 and 0.95 quantiles of the predictive distribution at each location and depth from the soilgrids.org platform. Values were divided by 10 to get percent values for each texture class. Zero values (for the sea and masked-out areas) were turned into Nulls after ensuring that these were not valid zero predictions.

The prediction uncertainty was quantified by the 90 % prediction interval width (PIW_{90}), defined as the difference between the estimates of the 0.95 and 0.05 quantiles:

$$PIW_{90} = q_{0.95} - q_{0.05} \quad (1)$$

where q_α is the α-quantile of the predictive distribution.

2.2.2. SoilGrids training data

Fig. 2 shows the locations of the SoilGrids training data for both countries. Table 1 presents summary statistics. The SoilGrids training data in NL are a merge of the WoSIS[3] (Batjes et al., 2020) and LUCAS[4] databases. For NZ the training data are those provided in WoSIS. The NL metadata indicated that the analytical instrument included pipette, sieve (sand), field hand estimate or unspecified. Silt was either 0.002–0.05 mm in size or unspecified. The NZ metadata indicated that the analytical method was pipette or sieve (sand), and silt was 0.002–0.05 mm in size. All point data were reprojected to the Goode Homolosine projection to match the SoilGrids maps.

Table 1 shows that NL has considerably more training points than does NZ. Indeed, the data for NZ are so few, that the SoilGrids predictions are relying almost entirely upon training data from locations elsewhere in the world, most likely those where the feature space (covariates) are similar to those in NZ. Note that while the models were evaluated by splitting the data using a tenfold cross-validation procedure, the models used to produce the SoilGrids layers were generated using all of the available data (Poggio et al., 2021).

2.2.3. Independent test data

The locations of the independent test datasets are shown in Fig. 2 and summary statistics of texture variables are given in Table 1.

For NL, we obtained independent data as part of the "Profielbeschrijving" (PFB) dataset, which is openly available from Helfenstein et al. (2024b). These were soil samples collected at locations arranged in a purposive sampling design, selected in the past to create the national 1:50,000 soil map of NL (de Vries et al. 2003), meaning that soil variability is covered reasonably well. Soil samples measured in the laboratory by the pipette method were collected by soil horizon between 1953 and 2012. Observations at PFB sites that were also in the WoSIS database (and therefore used as SoilGrids training data) were removed. Silt values were generated for soil samples where sand and clay values were provided as follows: silt = 100 – clay – sand.

The NZ independent dataset was extracted from the National Soil Data Repository (NSDR; Manaaki Whenua - Landcare Research (2020b)). This dataset is a collation of laboratory data collected over many decades for a range of projects (i.e. it has no coherent sampling design). Sampling depths are mostly related to observed soil horizons (rather than fixed depths). All data were analysed for soil texture using the pipette method. Soil samples with values for sand, silt and clay that summed to values smaller than 98 or larger than 102 were rejected, other data were scaled to sum to 100 if needed.

2.3. Data preprocessing

Both the training and independent test data have specified top and bottom depths that vary. In most cases these did not match the fixed standard depths of the SoilGrids predictions, so conversion of training and independent data to standard depths was needed before comparison. This was done as follows. We calculated the overlap of soil samples from a site with each standard depth interval. Only samples where the total coverage of each standard depth interval met the minimum overlap requirement (Table 2) were used. For example, a sample of 8–18 cm has 7 cm of overlap with the 5–15 cm interval, which is more than the minimum required overlap. Where the fixed interval comprised more than one physical sample, a weighted average was calculated in proportion to the amount of overlap. For example, with a second sample of say 20–30 cm, the weightings for the first and second samples, for the 15–30 cm fixed sample interval, would be 3/13 and 10/13, respectively.

The locations of the training and independent data for each fixed depth were overlaid on the relevant SoilGrids layers to obtain a predicted value for each measurement location and depth. The SoilGrids prediction at a sample location was taken as the grid cell value that the sample location fell within. Note that all training and independent data as well as predictions are at the point support, although some database observations are based on multiple aggregated samples covering 10 s of metres. Such observations would smooth out some of the short distance variation in soil.

[2] See https://soilgrids.org/.

[3] See https://www.isric.org/explore/wosis (downloaded November 2021).

[4] See https://esdac.jrc.ec.europa.eu/resource-type/soil-point-data (2009 & 2015 topsoil data).

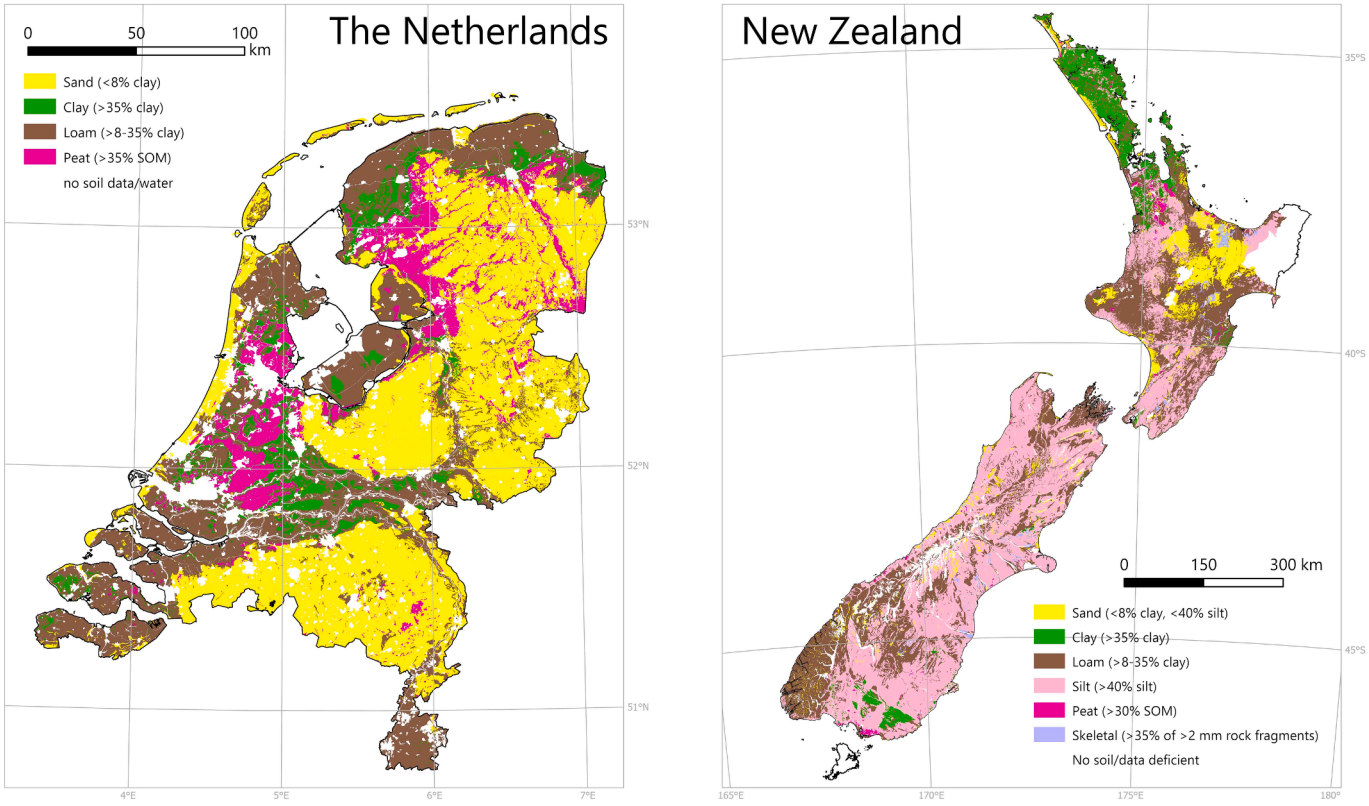


Fig. 1. Map of general soil texture classes in: left) the Netherlands (Wageningen UR-Alterra, 2006); right) New Zealand (derived from Manaaki Whenua - Landcare Research, 2020a). Note that the Dutch texture classes do not explicitly indicate soils dominated by silt but groups them with loam. The loam mapped in the southeast of the Netherlands is mainly derived from loess and is the equivalent to the silt as mapped in New Zealand.

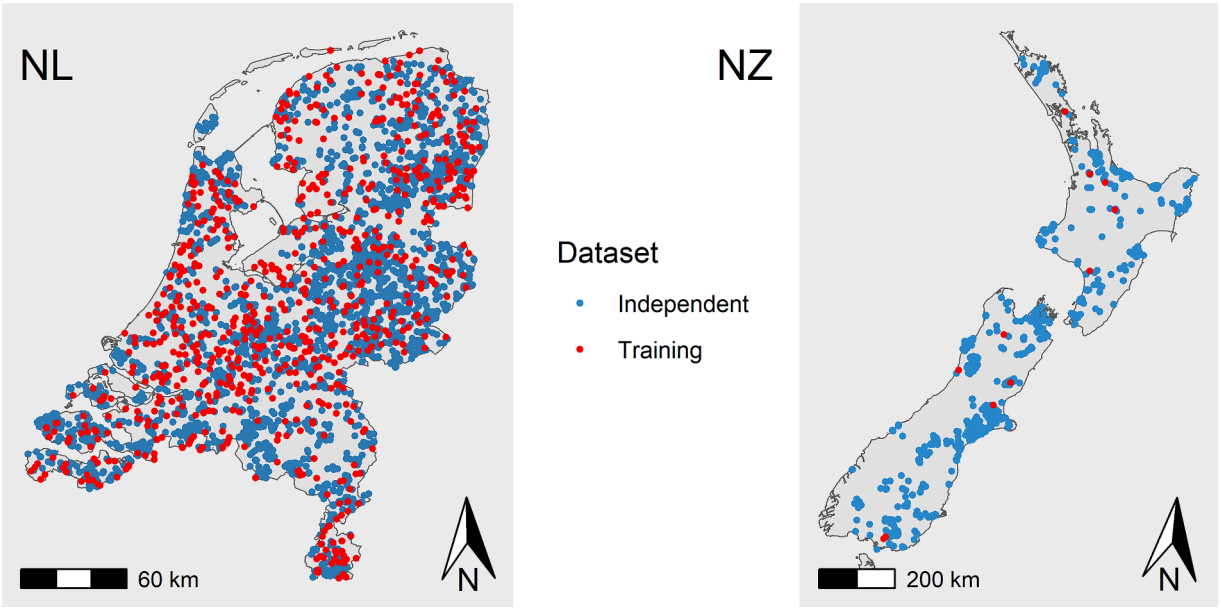


Fig. 2. Locations of the training and independent test data in the Netherlands (NL) and New Zealand (NZ).

2.4. Analyses

Commonly used accuracy metrics were generated by taking the predictions at each depth and location, then calculating the residuals by subtracting the observations from the predictions and estimating from them the ME (bias), the RMSE and the MEC:

$$ME = \frac{1}{n}\sum_{i=1}^{n}(pred_i - obs_i) \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(obs_i - pred_i)^2} \quad (3)$$

Table 1
Summary statistics of the training and independent test datasets. by particle size for the Netherland (NL) and New Zealand (NZ).

		Number of soil profiles	Number of soil samples	Min	Q1	Median	Mean	Q3	Max	SD
Clay [%]										
NL	*Training*	948	2,920	0	3.7	12.0	18.5	29.4	87.5	17.9
	Independent	3,116	11,580	0	3.1	6.2	14.1	20.0	90.3	16.1
NZ	*Training*	13	67	1	11.5	22.0	26.0	33.0	79.0	20.6
	Independent	1208	6,009	0	13.0	22.0	24.0	30.9	95.4	16.1
Silt [%]										
NL	*Training*	952	3125	0	10.0	19.1	26.5	44.0	94.7	20.3
	Independent	3,017	11,339	0	6.8	17.3	24.0	40.0	97.5	19.9
NZ	*Training*	13	61	1	32.0	52.0	47.7	64.0	82.0	22.0
	Independent	1166	5,761	0	36.3	50.0	47.3	61.4	95.0	19.1
Sand [%]										
NL	*Training*	857	2619	0	17.0	62.5	54.2	88.0	99.2	34.9
	Independent	3,019	11,391	0	33.1	75.6	62.0	89.7	100	32.4
NZ	*Training*	9	38	2	7.0	13.0	25.1	37.0	97.0	27.7
	Independent	1,188	5,861	0	9.4	20.7	29.5	43.2	100	25.9

Min = minimum, Q1 = 1st quartile, Q3 = 3rd quartile, Max = maximum, SD = Standard deviation.

Table 2
Minimum sample overlap required for each fixed-depth interval. The minimum overlap is approximately half of the interval width, except for the deepest interval where data are limited.

Fixed-depth interval (cm)	Minimum overlap (cm)
0–5	3
5–15	5
15–30	7.5
30–60	15
60–100	20
100–200	30

$$MEC = 1 - \frac{\sum_{i=1}^{n}(obs_i - pred_i)^2}{\sum_{i=1}^{n}(obs_i - \overline{obs})^2} \quad (4)$$

where n is the number of observations, obs_i is the observed/measured soil property of the i-th soil sample, $pred_i$ is the associated SoilGrids prediction, and $\overline{obs}$ is the mean of the observations. We also calculated the $PICP_{90}$, that is the PICP of the 90 % prediction interval, which was defined earlier in the Introduction.

3. Exploration of SoilGrids prediction uncertainty

This section reports on the uncertainty information provided by SoilGrids. Our evaluation of map accuracy and uncertainty estimates using SoilGrids training data will be presented in Section 4 and that using independent data in Section 5. We will focus on the 5–15 cm depth as other depths had very similar results. Most maps are of clay but some additional results for silt and sand are provided in the Supplemental Information.

The SoilGrids PIW_{90} for clay at 5–15 cm is mapped in Fig. 3. For NL the clay-rich Holocene parts in the north and west have a large prediction uncertainty, while the sandy Pleistocene parts in the east and southeast have much lower uncertainty. For NZ the PIW_{90} is much less spatially variable and is large everywhere. NL maps of silt PIW_{90} (Suppl. File Figure S.1) are similar to clay but the Dutch PIW_{90} maps of sand (Suppl. File Figure S.2) are more spatially variable. Suppl. File Fig. S.1-3 also show the mean predicted silt, sand and clay respectively. Fig. 4 compares the distribution of the PIW_{90} between the two countries and the three textures for the 5–15 cm depth. Only the NL clay map has areas of high certainty (e.g. $PIW_{90} < 20$ %).

The prediction intervals for other depths show the same pattern where the NZ prediction intervals are very wide, i.e. there is considerable uncertainty for all predictions in NZ, and the Dutch prediction intervals range from very narrow to very wide, i.e. some predictions in NL are very uncertain but others are relatively certain. The detail is shown in Suppl. File Table S.1. The very narrow prediction intervals (i.e. higher certainty) for clay largely coincide with Pleistocene and older landforms in the South and South-East of the Netherlands, where sandy soils dominate and clay content is low. The average PIW_{90} is smaller in NL than in NZ for clay and sand, and similar for silt. However, the maximum PIW_{90} is wider in NL than in NZ, and the minimum PIW_{90} in NL is considerably smaller than the NZ minimum.

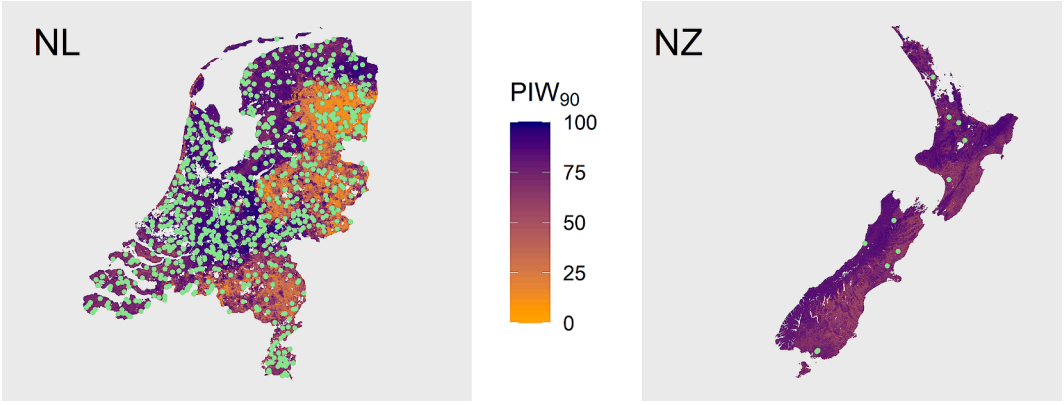


Fig. 3. Maps of SoilGrids PIW90 (90 % prediction interval width) by country for clay 5–15 cm, overlaid with training point locations (green) for the Netherlands (NL) and New Zealand (NZ). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

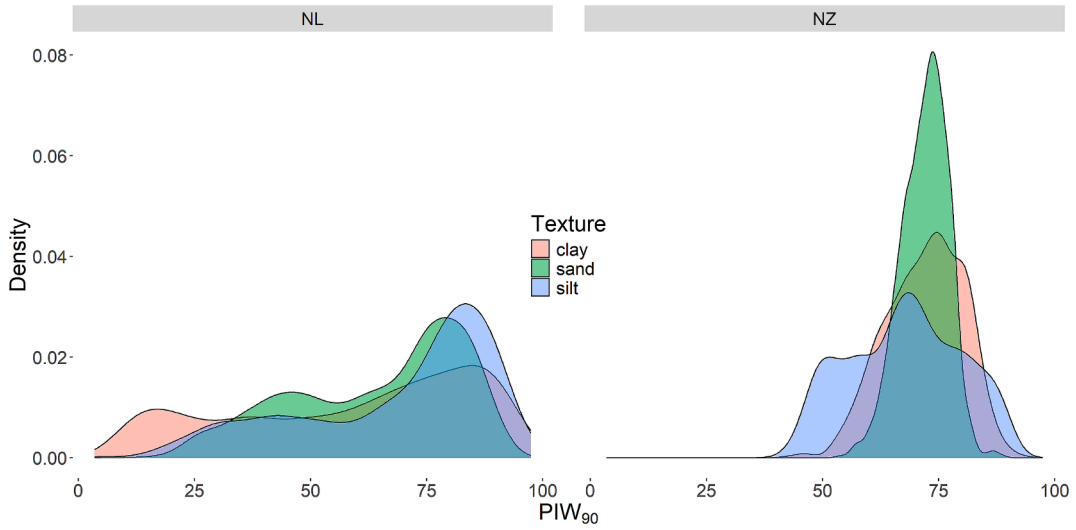


Fig. 4. Density plots of the SoilGrids PIW90 (90% prediction interval width) at 5–15 cm depth for the Netherlands (NL) and New Zealand (NZ).

4. Evaluation of SoilGrids using its training data

4.1. Relationship between the spatial density of the training data and the uncertainty

Fig. 3 shows that the PIW_{90} in NL does not appear to be higher where there are fewer training data nearby. The training points seem to be well distributed across the high (dark blue) to low (orange) uncertainty, and indeed have a higher density where there is higher uncertainty. There are insufficient training points in NZ to compare with the PIW_{90}. High uncertainty might be expected in areas where the training data are poorly represented, although we note that the QRF method is not per se spatial, and depends more on how many points in the training data are close to the prediction point in feature space. Note that SoilGrids is trained on a large global dataset. There might be training data outside NL and NZ that have similar covariate values as prediction locations in NL and NZ, thus reducing the PIW_{90} at these locations.

4.2. Global and local accuracy evaluation of uncertainty and predictions

The global prediction accuracy metrics from the SoilGrids training data (based on cross-validation) are reported in Poggio et al. (2021) and replicated in Table 3. Note that these are global metrics and that the metrics for specific countries, e.g. NL or NZ, may be better or worse. Therefore, local accuracy metrics were also computed according to Eqs. (2)–(4) by comparing training data to SoilGrids predictions in NL and NZ. Poggio et al. (2021) gave an overall global RMSE (all depths) of 13 %, 18 % and 13 % for clay, sand and silt, respectively. The respective global $PICP_{90}$ metrics for 5 – 15 cm are 96 %, 79 % and 96 %, which shows that the clay and silt PI_{90} are too wide and the PI_{90} for sand is too narrow.

Table 3 shows that based on the Dutch training data the PI_{90} for all textures are very pessimistic for NL because all $PICP_{90}$ values are much larger than 90. This is also the case for sand, where the global PIW_{90} was too optimistic. The NL predictions show good performance (MEC > 0.67) – considerably better than the global metrics. Clay and silt values are slightly overpredicted. Sand values are underpredicted and RMSE is higher for sand than for clay and silt. The metrics for the Dutch data are better than the global metrics. These improved accuracy metrics and pessimistic $PICP_{90}$ results are expected, as the SoilGrids predictions were directly informed by the training data and thus are not independent, so that the predictions should be closer to the observed values than for

Table 3
Published global accuracy metrics of SoilGrids texture maps based on cross-validation and local, country-level accuracy metrics based on the SoilGrids training data.

		Global		NL-WoSIS					NZ-WoSIS				
Texture	Depth	$PICP_{90}$	MEC	n	$PICP_{90}$	ME	RMSE	MEC	n	$PICP_{90}$	ME	RMSE	MEC
clay	0–5	0.96	0.45	578	99.0	3.79	8.36	0.67	13	92.3	0.45	16.16	−0.23
clay	5–15	0.96	0.42	797	99.2	2.87	7.56	0.76	13	100	0.53	15.09	−0.19
clay	15–30	0.96	0.42	443	99.5	1.87	7.98	0.78	12	100	1.37	19.34	−0.13
clay	30–60	0.95	0.41	398	99.5	1.39	8.51	0.78	13	92.3	4.42	22.44	−0.05
clay	60–100	0.95	0.40	363	98.1	2.14	9.70	0.69	9	100	6.76	26.16	−0.2
clay	100–200	0.96	0.40	84	98.8	3.46	8.54	0.76	3	100	16.18	18.77	−3.36
sand	0–5	0.82	0.59	509	95.9	−5.86	13.55	0.81	8	62.5	25.49	30.09	−3.22
sand	5–15	0.82	0.58	754	95.4	−5.72	13.47	0.82	8	62.5	24.27	29.06	−2.94
sand	15–30	0.80	0.57	443	96.6	−6.12	13.89	0.83	7	57.1	25.16	29.46	−3.25
sand	30–60	0.78	0.54	343	96.8	−4.21	13.38	0.86	8	50	19.29	31.29	−0.49
sand	60–100	0.78	0.50	299	93.3	−6.72	15.51	0.81	5	40	14.69	36.96	−0.12
sand	100–200	0.78	0.48	71	94.4	−8.70	15.51	0.80	3	33.3	−8.17	38.1	−0.11
silt	0–5	0.96	0.71	626	99.5	2.09	7.65	0.83	12	100	−22.64	27.71	−2.13
silt	5–15	0.95	0.64	837	99.5	2.02	7.78	0.84	12	100	−22.99	28.34	−2.11
silt	15–30	0.96	0.68	509	99.8	2.26	7.29	0.86	11	81.8	−22.58	29.98	−1.42
silt	30–60	0.96	0.62	461	99.1	2.55	8.28	0.84	12	83.3	−15.21	28.84	−0.36
silt	60–100	0.96	0.57	428	99.5	3.63	9.20	0.79	8	75.0	−17.15	31.24	−0.45
silt	100–200	0.96	0.54	72	100.0	4.84	9.90	0.78	3	66.7	−7.98	29.05	−0.14

independent data.

The training dataset for NZ indicates very poor performance for both the uncertainty and the predictions, with negative MEC values and $PICP_{90}$ values much larger or smaller than 90, but it should be noted that the number of points is very small.

4.3. Distribution of the training data and the predicted uncertainty by texture class

As well as exploring the spatial distribution of the PIW_{90} (Fig. 3), it can be useful to see if there is a pattern with respect to the soil texture classes. A ternary plot (Fig. 5) shows where a sample sits on the World Reference Base soil texture triangle (IUSS Working Group WRB, 2015), thus providing a quick way to visualise differences in soil texture feature space. Estimates are coloured according to the averaged PIW_{90} values for clay, silt and sand of each sample. This shows that in NL, the sand, sandy loam, and silty loam samples are more accurate (lower uncertainty) than for the other soil texture classes. The number of training data in NZ is too small to draw any conclusions.

Fig. 6 maps the range of soil texture values that are predicted at the training data locations in both countries overlaid with the measured soil texture values. The NL measurements reflect the distribution of the predicted soil texture classes. Only the high clay values are missing in the predictions. There is a clear mismatch between the distributions of the NZ measurements and the predicted soil texture classes.

5. Evaluation of SoilGrids using independent datasets

The locations of the independent data (sites) have already been shown in Fig. 2 and summary statistics presented in Table 1.

5.1. Evaluation of the uncertainty estimates and predictions

Table 4 gives the assessment of the proportion of independent observations that are within the prediction interval ($PICP_{90}$). In both countries, the clay PI_{90} values are much too wide (too pessimistic) and the sand PI_{90} values much too narrow (too optimistic), matching the global metrics. The silt PI_{90} is too wide in NL but too narrow in NZ. As can be seen in Fig. 4 the PI_{90} is in some cases very wide for silt and clay in NL.

Table 4 also provides the accuracy metrics for the independent dataset. The positive ME values show that in NL clay and silt values are overpredicted. Sand values are underpredicted, and their RMSE is quite high. An underprediction in one texture variable means that another texture variable (or two) must be overpredicted. MEC values in NL are low, varying between 0.20 and 0.47. In NZ the model performance is very poor (negative MEC values) for all three textures. Sand predictions are strongly overpredicted and silt values are strongly underpredicted. The bias and RMSE are smaller for clay in NZ. In some cases, ME is quite large compared to RMSE (e.g. sand and silt down to 60 cm in NZ), indicating that systematic prediction errors are larger than random prediction errors.

Fig. 7 shows that in NL, the predicted values for the independent dataset (blue) do not extend as far as the observed very high or very low sand content values. This reflects the smoothing effect of the modelling (Rossiter et al., 2022). In NZ there is a very poor match between the distribution of the predicted values (clay loams and sandy clays on the WRB soil texture triangle) and the observed values from the independent dataset.

5.2. Distribution of the independent data predictions and the predicted uncertainty by texture class

Fig. 8 shows the independent data predictions on the texture triangle along with the mean PIW_{90}, (i.e. the mean of the three PIW_{90} values for sand, silt, and clay). The predictions in NL range from sandy to loamy soils. As with the training data, the very high sand predictions have the least overall uncertainty (i.e. where the mean PIW_{90} of clay, silt and sand is low). The NZ predictions are all located at the intersection of sandy clay loams, clay loams and loams, and have large mean PIW_{90}.

Fig. 9 plots the predicted uncertainty against the prediction at each independent site location. In both NL and NZ, the low predictions of silt and clay, and both the high and low predictions of sand in NL are expected to be more reliable. For bounded soil properties like percent values, a decreasing uncertainty at either end of the range seems realistic, as the prediction interval is also bounded by 0. The red points are those where the observed value is outside the PI_{90}. The only clear pattern is that most of the NZ silt predictions that were outside the PI_{90} had smaller estimated PIW_{90}, i.e. they were (incorrectly) estimated to be more certain than other points.

Fig. 10 shows all the measured values along with their associated probabilistic values for the 0.05, 0.50 and 0.95 quantiles of sand and clay at 5–15 cm depth. Plots (a)–(c) are ordered by the predicted value, i.e. the mean. The NL sand plot (Fig. 10a) shows that the prediction

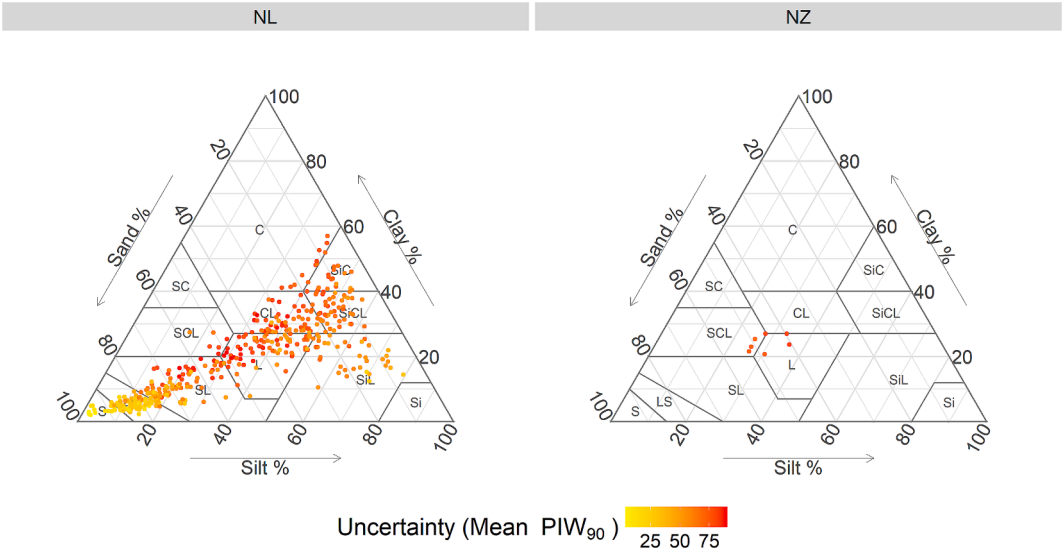


Fig. 5. The mean PIW90 (90 % prediction interval width) for the clay, silt and sand 5–15 cm depth predictions at the training data locations on a WRB soil texture diagram for the Netherlands (NL) and New Zealand (NZ).

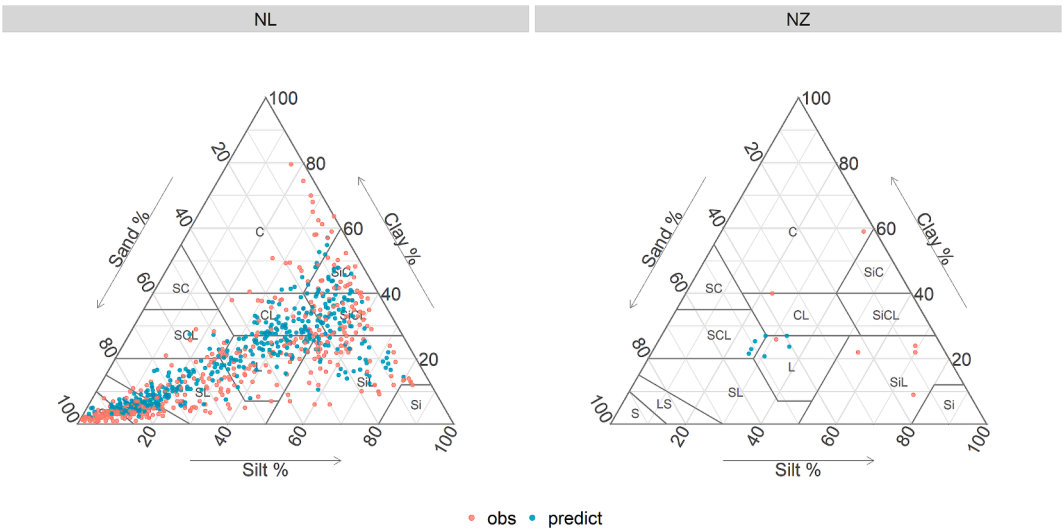


Fig. 6. Ternary plot showing the texture distribution of the measured (obs) and predicted (predict) soil texture values at 5–15 cm depth for the training data locations overlying the WRB texture classification for the Netherlands (NL) and New Zealand (NZ).

Table 4
Accuracy metrics from the independent datasets for New Zealand (NZ) and the Netherlands (NL).

		NL – PFB					NZ – NSDR				
Texture	Depth	n	$PICP_{90}$	ME	RMSE	MEC	n	$PICP_{90}$	ME	RMSE	MEC
clay	0–5	1,645	97.3	5.34	11.26	0.24	1,074	98.5	−1.1	13.3	−0.1
clay	5–15	2,695	97.2	4.01	11.29	0.38	1,021	98.4	0.6	13.5	−0.1
clay	15–30	2,094	96.9	4.23	12.32	0.33	981	98.4	2.6	15.5	−0.1
clay	30–60	1,796	95.8	4.40	13.21	0.30	832	97.2	7.8	18.7	−0.2
clay	60–100	1,615	95.0	5.69	13.41	0.09	571	95.4	8.7	21.0	−0.2
clay	100–200	233	95.7	4.98	12.4	0.18	155	97.4	8.6	20.9	−0.2
sand	0–5	1,562	88.1	−7.31	21.74	0.40	1,059	61.9	27.3	34.4	−2.6
sand	5–15	2,608	87.8	−6.15	21.37	0.47	1,017	66.0	24.9	32.8	−2.0
sand	15–30	2,095	86.7	−6.13	22.74	0.42	972	68.3	21.2	31.6	−1.2
sand	30–60	1,795	80.0	−8.40	24.90	0.40	817	69.6	10.4	29.4	−0.2
sand	60–100	1,529	77.4	−11.68	27.53	0.20	579	65.3	5.6	31.3	−0.1
sand	100–200	232	86.2	−8.72	24.42	0.39	156	70.5	5.4	31.4	−0.1
silt	0–5	1,562	96.7	2.08	14.03	0.40	1,038	78.9	−27.3	31.3	−4.1
silt	5–15	2,607	97.0	2.23	13.66	0.44	996	75.0	−26.7	31.0	−3.7
silt	15–30	2,097	96.2	1.95	14.32	0.39	951	78.0	−25.0	30.5	−2.7
silt	30–60	1,798	94.2	4.01	15.33	0.39	796	82.5	−19.6	28.1	−1.2
silt	60–100	1,535	94.5	6.13	16.86	0.24	558	83.9	−16.1	27.7	−0.7
silt	100–200	232	96.5	3.76	15.83	0.41	232	96.5	3.76	15.83	0.41

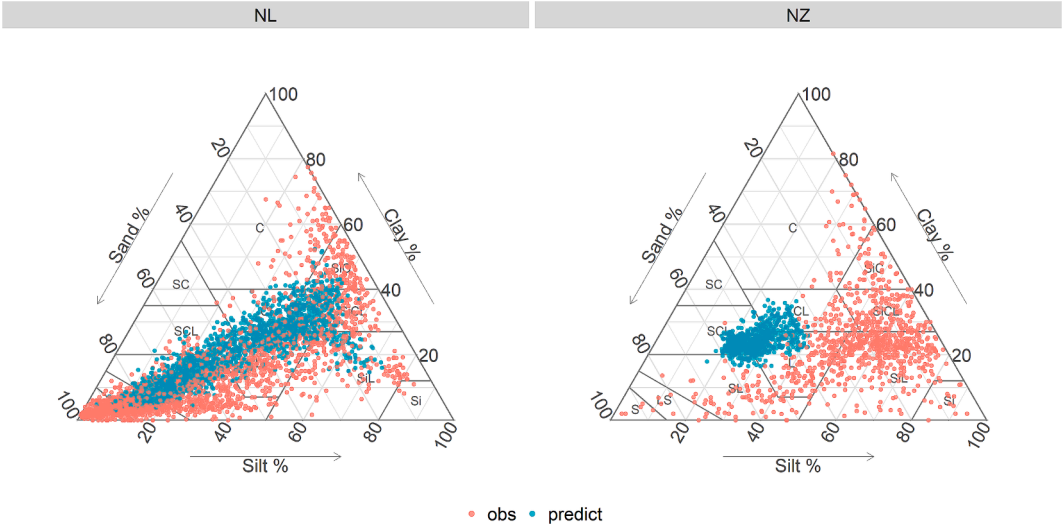


Fig. 7. Observed (obs) and predicted (predict) independent data plotted on texture triangle plot by country for the Netherlands (NL) and New Zealand (NZ).

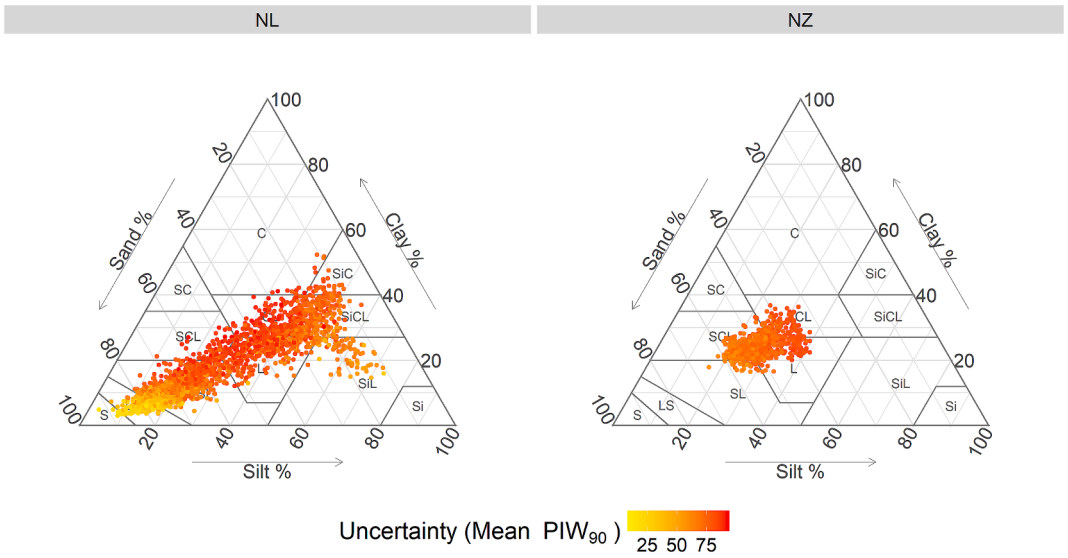


Fig. 8. Texture triangle plot showing the predicted values of the independent points with their averaged uncertainty (i.e. averaged 90% prediction interval width (PIW90) for sand, silt, clay) by country: the Netherlands (NL) and New Zealand (NZ).

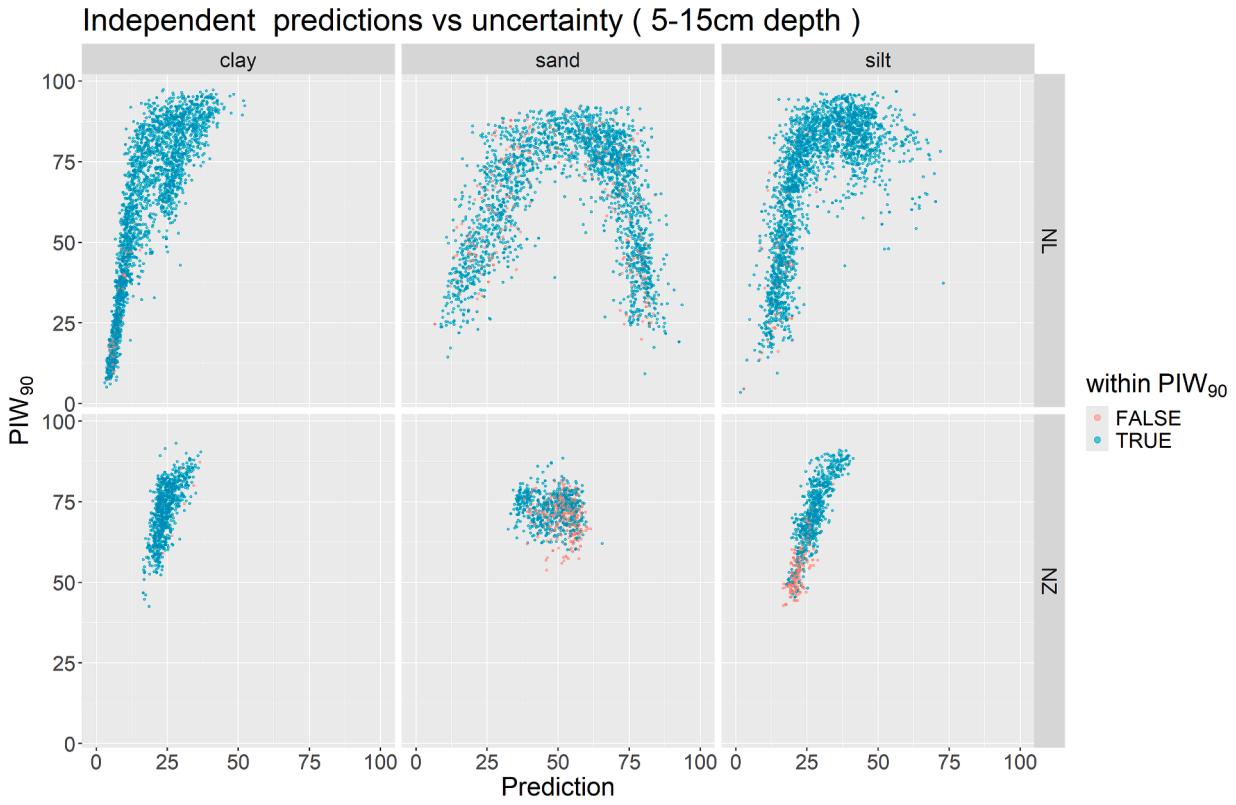


Fig. 9. Plot of 90 % prediction interval width (PIW90) for uncertainty against the prediction for 5–15 cm depth at each of the independent site locations. Red points indicate where the measured value is outside the 90 % prediction interval for the Netherlands (NL) and New Zealand (NZ). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

interval is smaller where the sand prediction value is either very low or very high. The NL clay plot (Fig. 10b) has narrow prediction intervals where clay is predicted to be low, becoming wider with increasing clay content. Even when the clay prediction is very high the lower 0.05 quantile stays low. Intriguingly, all the plots show that the median is generally lower than the mean prediction, except for sand > 50 %. The NZ clay plot (Fig. 10d) is ordered by observation (rather than prediction, which is shown in Fig. 10c), highlighting the lack of relationship between the predictions, prediction intervals, and observed values.

5.3. Relationship between residuals and uncertainty

Higher residuals might be expected in locations with higher uncertainty. Fig. 11 shows the residuals plotted against the PIW_{90} for the 30–60 cm depth layer. Red points are outside the PIW_{90}, blue are inside. The greater range of PIW_{90} size in NL shows that larger residuals are

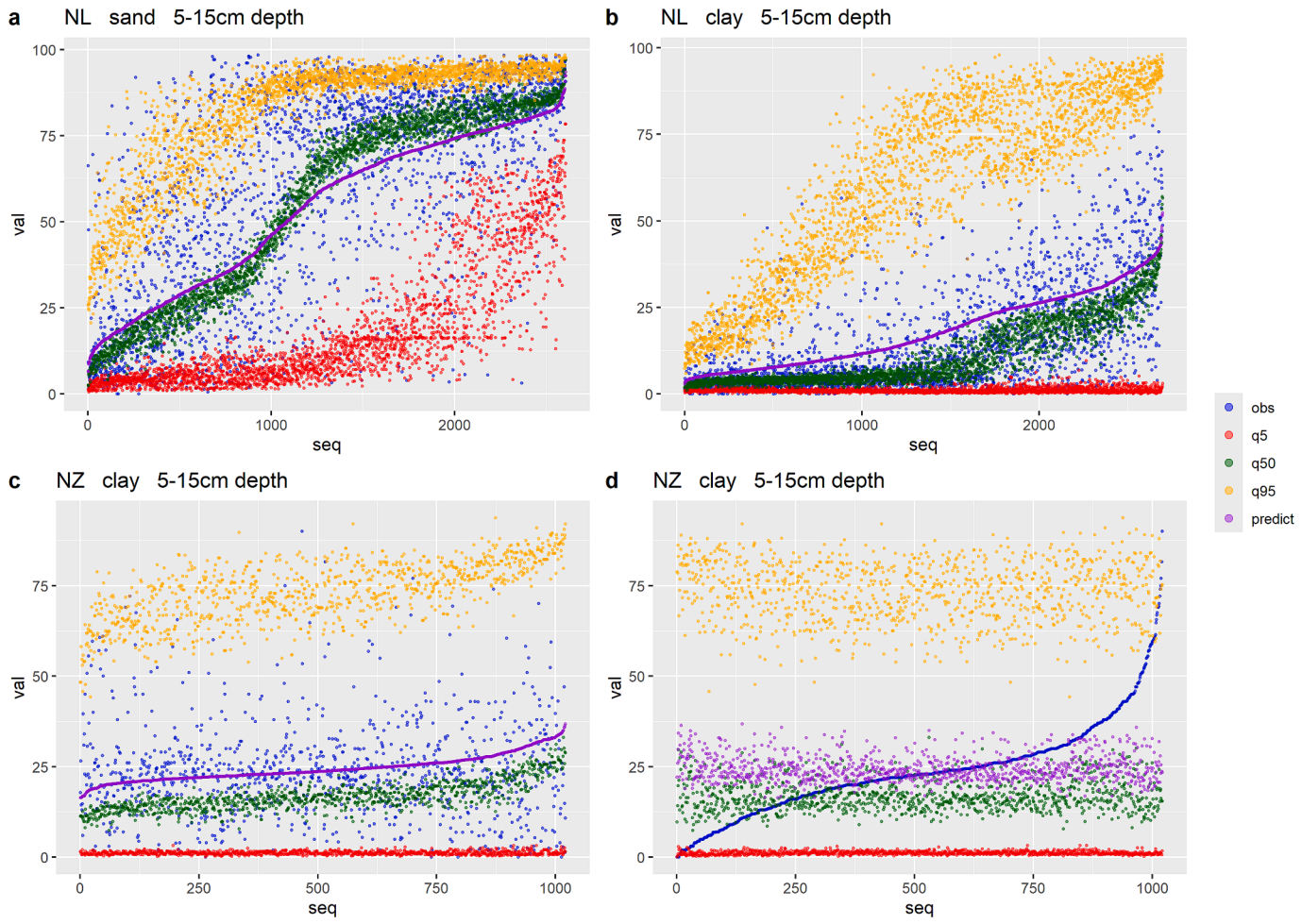


Fig. 10. Plots of measurements, predictions and quantiles of sand and clay 5–15 cm depth for the independent datasets in the Netherland (NL) and New Zealand (NZ): (a) NL sand ordered by increasing predicted value; (b) NL clay ordered by increasing predicted value; (c) NZ clay ordered by increasing predicted value; (d) NZ clay ordered by increasing observed value. obs = measurement, q5,q50,q95 are the 5th, 50th, and 95th quantile values, predict is the mean predicted value.

generally associated with larger PIW_{90}. Other depths show a similar pattern. Interestingly the NL points that are outside the PIW_{90} are associated with both low and high estimates of uncertainty, i.e. narrow and wide PIs. In NZ there is no relationship between residual and PIW_{90}, but the PIW_{90} values are all quite high.

6. Discussion

6.1. Findings from this study

Some novel visualisations were used to explore the SoilGrids predictions and their uncertainty in NL and NZ. Some visualisations were solely based on the provided SoilGrids information, while others relied on the training or independent datasets to gain further insight into the accuracy of the soil information. Evaluating the predictions against the training data of the respective countries allowed us to get a sense of the accuracy of the local predictions (while being aware of the effects of data leakage) and to look for patterns in the error residuals and uncertainty (e.g. some soil types or locations being more accurate). Evaluation against the independent data permitted a much more robust analysis of the errors and uncertainty of the SoilGrids predictions and prediction uncertainty estimates. The key findings from the analyses presented in Sections 3–5 are:

- There can be considerable variation in the level of uncertainty of texture predictions within a country. The PIW_{90} in NL had spatial structure and a wide distribution, indicating that some SoilGrids predictions in some areas are more certain than others, whereas the NZ predictions were consistently uncertain.
- The level of SoilGrids prediction uncertainty in NL did not appear to relate to the spatial sampling density of the training data (Fig. 3). This might be because the PIW_{90} computed using QRF depends on distances in feature space instead of distances in geographic space. It could also be that these are areas in feature space that have high variability in soil texture (i.e. observations show large variation in silt, sand and clay while covariate values are similar). Or it might be that there is significant variability in the covariates, such that the feature space is not well represented, despite the good geographical representation.
- The decreased level of SoilGrids uncertainty was found in the parts of NL with very low or very high sand content, or very low clay content. Fig. 11 indicates that the higher level of certainty was supported by the independent dataset in that errors were generally smaller where there was higher certainty.
- The level of uncertainty for texture of SoilGrids in NL and NZ is quite high (Figure 4).
- The clay and silt uncertainty is pessimistic, i.e. the PIW_{90} is often too large, as demonstrated by the overly high $PICP_{90}$ values in the independent dataset.
- The local accuracy metrics from the SoilGrids training data generally indicated similar characteristics as the independent data, including:
  - o NL sand was more poorly predicted than clay or silt.

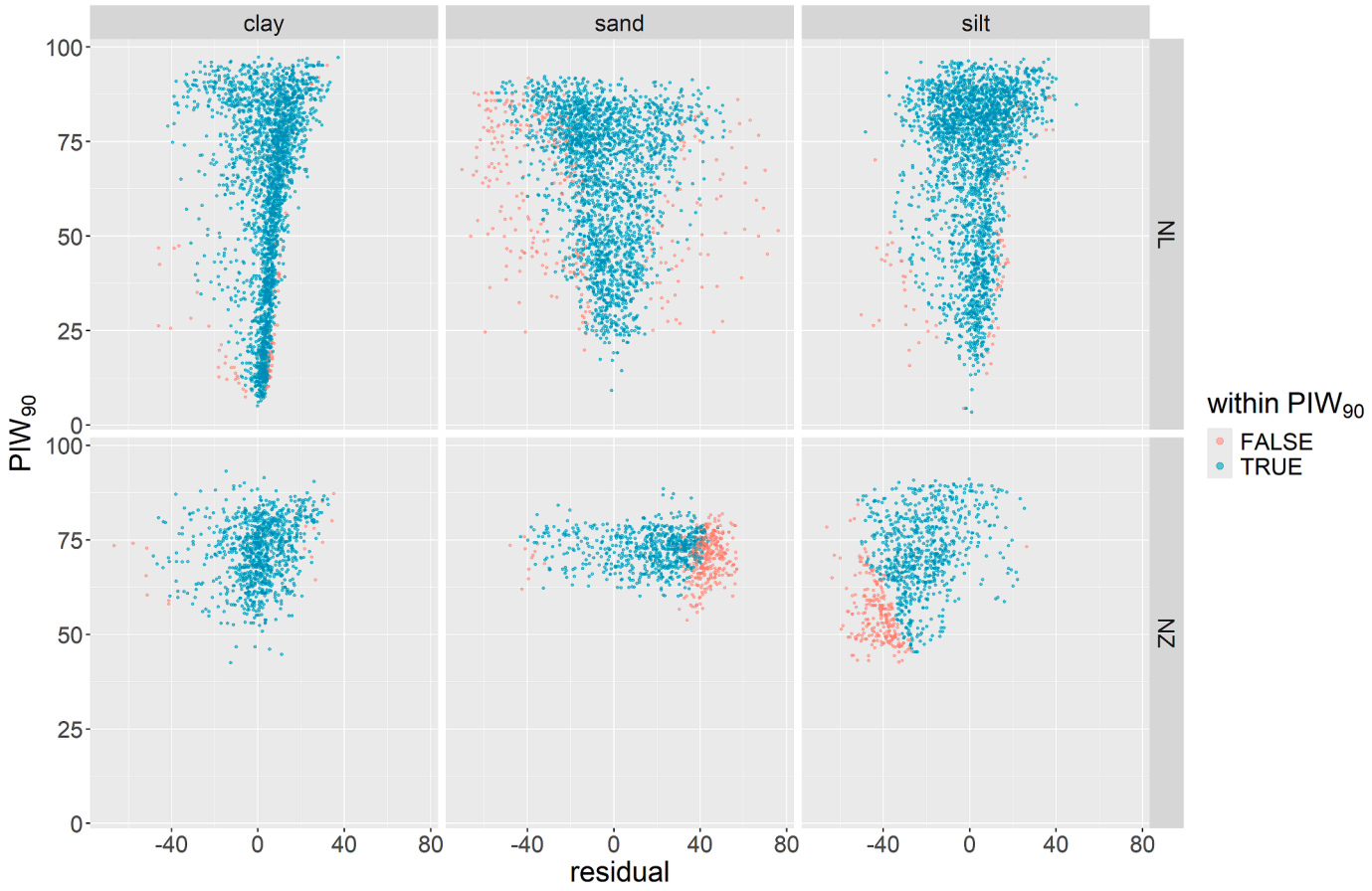


Fig. 11. Residuals vs 90 % prediction interval width (PIW90) by texture and country for 5–15 cm depth layer for the independent dataset for the Netherlands (NL) and New Zealand (NZ).

- o NL clay and silt were overpredicted whereas sand was underpredicted.
- o NL clay and silt uncertainty was pessimistic.
- o The accuracy metrics of the NL predictions displayed the same patterns across clay, silt, and sand between the training and independent datasets. However, as expected given that the training data were used to develop the model (data leakage), the level of accuracy in the independent dataset was not as good as the training data accuracy (for both NL and published global metrics).
- o NZ texture was very poorly predicted (worse than global metrics of SoilGrids stated in Poggio et al., 2021)
- o NZ sand was overpredicted, and silt was underpredicted (i.e. the opposite to NL).

- Some differences in the results between the training and independent datasets are:
  - o NL sand uncertainty was too optimistic in locations where we had independent data, unlike in the training dataset. Note that the NZ training dataset was too small to draw any conclusions, but the NZ independent dataset indicated that the SoilGrids silt as well as sand uncertainty was too optimistic.
  - o The more extreme soils (very high clay, sand or silt) were not well predicted in the independent dataset – but this is less evident in the training dataset.

6.2. Comparison with other studies

Various studies have developed national maps and compared them to the global SoilGrids products. Bahri et al. (2022) report poor performance in Tunisia for soil organic carbon 0–30 cm (ME = 0.79, RMSE = 2.52, $R^2 = 0.15$) for SoilGrids2.0 (Poggio et al., 2021), where many high SOC sites were underestimated. Conversely, Mulder et al. (2016) in an analysis of France found that the SoilGrids 1 km product (Hengl et al., 2014) overestimated SOC levels. They found that the training data located in France were limited and non-representative (biased towards high SOC values). Chen et al. (2019) studied soil pH in China and found that the SoilGrids 250 m product (Hengl et al., 2017) had a modest performance (RMSE = 1.02 pH units and Lin's correlation concordance coefficient (CCC) = 0.67) based on an independent dataset of 4,700 profiles across China. This was lower than their national map derived from a larger dataset (RMSE = 0.71 pH units and CCC = 0.84), but had a similar spatial pattern. Chen at al. comment that the SoilGrids 250 m map showed great potential where data were sparse.

Helfenstein et al. (2022) compared their national-scale maps of pH with SoilGrids2.0 predictions in NL. The global SoilGrids accuracy metrics (Poggio et al., 2021) were similar to the accuracy metrics of the national pH maps (Helfenstein et al., 2022). However, deriving NL-specific metrics from their independent dataset showed much poorer metrics for SoilGrids, e.g. the SoilGrids uncertainty was too optimistic at $PICP_{90} = 0.71$. They also evaluated the prediction accuracy and obtained RMSE = 1.23 pH units and MEC 0.34. This supports our results in that the local metrics for SoilGrids2.0 predictions and uncertainty for NL were worse than the global metrics. In contrast, Rossiter et al. (2022) compared alternative maps of pH with soilGrids2.0, rather than measurement points. They felt that the very large uncertainty in SoilGrids2.0 was unduly pessimistic.

Cramer et al. (2019) found that clay and pH varied markedly from the SoilGrids 250 m product (Hengl et al., 2017) in a study in the Greater Cape area of South Africa. They noted that training data from only 209 sites were available to SoilGrids, and that these were biased towards agriculturally productive soils. They suggested that another reason for

the improved local modelling could be that patterns with environmental predictors that are important globally may not be relevant to the local area.

Han et al. (2022) in a multi-scale study in Australia found that the local accuracy metrics of the SoilGrids 250 m clay and carbon products (Hengl et al., 2017) were worse than the published global metrics, and that national and state maps improved upon SoilGrids. Their RMSE for point samples of clay was 19.34 %, comparable with our NZ result but worse than our NL metric. Radočaj et al. (2023) in a study of texture from the SoilGrids2.0 (Poggio et al., 2021) product in Croatia found low MEC values based on an independent dataset (e.g. 0.267, 0.039, 0.039 for clay, silt, and sand, respectively [averaged over three depths]). The RMSE were 12.59 %, 24.6 %, 28.7 %, respectively. These results indicate slightly worse performance than our NL results but much better than the NZ results. They note that the local RMSE values for clay, silt and sand were similar to the global RMSE metrics.

6.3. Sources of uncertainty

Source 1: Training data errors.

The very poor accuracy results in the training data for NZ (Table 3) prompted an investigation into the cause of this. It turned out that most of the points were incorrectly located, possibly a mix of reprojection issues and transcription errors (some were incorrect by < 1 km, others by 50, 125 and 430 km). The correct information has now been supplied to the WoSIS team and will be incorporated in the next SoilGrids release. The training data metrics presented in this paper (Table 3) are based on the correct locations. These metrics are just as poor as the metrics derived using the incorrect WoSIS locations (not shown).

Many training data are taken from depth intervals that differ from the GSM depth intervals. A given measurement might also relate to soil below or above the target soil depth interval – or it may only partially cover the interval. An equal-area spline can be fitted to the available measurements in a soil profile to harmonise data to specified depth intervals (Bishop et al., 1999). However, this spline interpolation process is not error-free. Note that the SoilGrids process did not need to harmonise data over depth because it took a 3D modelling approach (Ma et al., 2021). But depth harmonisation is often needed in 2D digital soil mapping, for which equal-area spline and weighted averaging are often used. This error source of a mismatch between measured and predicted depth intervals also pertains to model evaluation using independent data. It would be worthwhile investigating the effect of depth harmonisation methods and the errors they bring about in future research.

While re-checking the metadata, it became apparent that there were a range of methods used across the full training dataset. Methods included pipette, hydrometer, and laser diffraction. In addition, different countries used different thresholds for the definition of silt and sand, as recently noted in Batjes and van Oostrum (2023). Table S.2, adapted from Nemes and Rawls (2006), shows the thresholds used by different countries. We note that the WoSIS metadata for the NZ training data indicated a threshold of 0.05, which is incorrect. While there are approximate methods for converting sand/silt fractions to a different threshold (GSM, 2015; Minasny and McBratney, 2001; Nemes et al., 1999), these have not been applied to the training data as the development, application and testing of such transfer functions is beyond the remit of WoSIS (ISRIC, 2024). Development of a generic threshold transformation process is challenging and will not lead to error-free conversions, thus contributing to this source of uncertainty.

The errors in the very small number of training points for NZ are unlikely in our view to affect the SoilGrids model as the NZ samples are a tiny fraction (e.g. 0.006 % for clay 5–15 cm) of the SoilGrids training data. However, the threshold definition issue may be more significant. In Australia, for example, a sample will be analysed as having a higher proportion of sand content and lower silt content than the same sample in another country. This is because a particle size of, for example 0.04 mm, is classified as 'sand' in Australia but as 'silt' in most other countries. This may be the cause of the poor performance of sand depending on the influence of the training samples from the non-0.05 mm threshold countries as these samples have not been converted to a common threshold of 0.05 mm.

The difference in threshold could result in sand predictions for NZ (and maybe NL) that are slightly higher than NZ observations, and silt predictions that are slightly lower than NZ observations – due to: 1) the influence of training data samples based on a smaller threshold; and 2) the independent observations being based on a different threshold than the SoilGrids predictions (where the threshold is defined as being 0.05 mm). A quick analysis of NZ data with texture fraction data at 0.05 and 0.06 mm indicates that the observed silt values derived using the 0.06 mm threshold, are on average higher by 3.7 % (and hence sand is lower by 3.7 %) with a standard deviation of 3 %. This is consistent with Tables 3 and 4 in terms of the direction of the bias, but the magnitude of the bias is considerably greater than –3.7 %.

In general, significant errors in the training data are likely to lead to a poorer relationship with the covariates and thus wider prediction intervals in the areas of feature space with training data errors.

Source 2: Limited training data.

The low number of data points for NZ and poor performance suggests a problem with a lack of training data. However, it is not necessarily a problem if there were sufficient training data from other countries with similar covariate values. Meyer and Pebesma (2020) introduced the area of applicability (AOA) which shows where model prediction error is expected to be high due to extrapolation in feature space. This is calculated by means of a dissimilarity index that describes how different the covariates at a predicted location are to the covariates of the training data. It would be very interesting to evaluate whether NZ has a high dissimilarity, and higher than NL. Unfortunately, the size of the training and covariate data used in SoilGrids means that this calculation is computationally very challenging (Laura Poggio, pers. comm. March 2022). Hateffard et al. (2024) tested four indicators of similarity including AOA and PIW_{90} in a random forest DSM study, finding that none had a strong correlation with the evaluation metrics.

Limited training data also includes the situation where the training data is biased. For example, if an area of the feature space is strongly dominated by observations from a region with higher values of sand than other regions with the same covariates, then the sand predictions in the other regions will be biased and uncertainty estimates will not reflect the additional uncertainty. An example of biased data is described under Source 1 (i.e. different regional sand/silt thresholds).

Note that the prediction intervals may not reflect limited and/or biased training data.

Source 3: Inadequate covariates.

In general, soil particle size distribution is mainly controlled by two factors (Schaetzl and Thompson, 2015). Firstly, physical and chemical weathering of rocks and minerals will produce increasingly finer grain sizes, depending on rock/mineral type, duration of weathering and environmental conditions. Secondly, and particularly relevant for soils where these weathering processes have not yet sufficiently progressed, soil texture can be inherited through the geomorphological origin of the soil parent material. Some sediment transport processes result in soil parent materials with strata of uniform particle size distributions, such as fluvial, aeolian or tephric deposits, where factors like river discharge, wind speed, type of volcanism or transport distance will determine the outcome. Other processes produce highly inhomogeneous, unsorted deposits, such as landslide or glacier movements. To predict soil particle sizes adequately, we argue that the covariates not only have to describe environmental factors like climate, lithology or vegetation, but also should convey information about the depositional age of the soil parent material and geomorphic environment in which the soil formed.

Unfortunately, none of the covariates used in the global model are direct indicators of this, limiting the applicability of the model to predict soil particle size. However, there are indirect links between the available covariates and geomorphic processes. For instance, in the spatial model

of clay distribution for NL using national data, elevation was found to be one of the most important variables (Helfenstein et al., 2024a). In the Netherlands, elevation differences coincide with the change between elevated Pleistocene (or older) and low-lying Holocene surfaces. Soils on the higher elevated landforms are dominated by relatively clay-poor Pleistocene sands. This spatially more homogeneous area is represented by an adequate number of observations enabling lower prediction uncertainties and higher accuracy of model predictions. The lower-lying surfaces are the result of dynamic fluvial and coastal processes (e.g. river avulsions, marine trans-/regressions) during the Holocene, producing highly variable spatial patterns of loamy and clayey soils (de Bakker and Schelling, 1989; Edelmann, 1950), coinciding with areas of higher model uncertainty and lower accuracy. While this makes elevation a useful indirect proxy for spatial distribution of soil texture and its prediction uncertainties in NL, a similar relationship may not be applicable elsewhere, particularly not in NZ, where active tectonics and volcanism throughout the Quaternary (Hewitt et al., 2021) have created very different geomorphological conditions compared to NL.

With respect to scale, it is not surprising that the global model using covariate data of $\geq$ 250 m resolution is more uncertain in regions with high spatial variability over short distances ($<$ 250 m), like the Holocene fluvial surfaces in NL, compared to the more uniformly sand-covered older surfaces. For example, in the riverine and peat areas in NL south of Amsterdam, heterogeneous mixtures of clay, silt, sand and even peat can be found within relatively short distances due to changing river flowing and flooding regimes in the past (Brouwer et al., 2023; Brouwer et al., 2018). This degree of variability is most likely not captured by global covariates. So even though the density of sampling points is high, their variability and the poor relationship with the available covariates means that predictions in these areas are highly uncertain. However, this is not an explanation for the uncertainties in NZ, where soil landscapes with a more uniform soil texture (e.g. regions with widespread loess and tephra cover) have similarly high uncertainties as more variable locations (Suppl. File Figure S.2).

Another limitation is the accuracy of the covariate layers. SoilGrids relies on global covariates, some of which have substantial uncertainty. The impact of this limitation could be assessed by quantifying these uncertainties and propagating them through the SoilGrids algorithm. While this analysis was beyond the scope of the current study, it represents a valuable direction for future research.

In general, inadequate covariates are likely to lead to a poorer relationship and thus wider prediction intervals.

Source 4: Model structure.

This error source relates to whether the choice of model type, model parameters and preprocessing (e.g. transformations, scale) can fully represent the relationship between the covariates and the soil property.

The random forest model is a popular choice for modelling of soil properties. In this study, model performance was poor for soil texture. Global accuracy MEC values for silt ranged from 0.40 to 0.71. The local metrics for NL training data were higher (0.67–0.86), probably due to data leakage. The NL independent dataset achieved a worse performance (MEC ranges from 0.24 to 0.44 for silt) than the global metrics, suggesting possible overfitting of the random forest model. Of course, this poor performance may be due to Source 3 rather than model structure.

Other model structures that could have been employed include regression kriging, deep learning (convolutional neural networks), use of contextual information (e.g. Behrens et al., 2018), or different scales (e.g. Samuel-Rosa et al., 2015). Comparison of results of these approaches to the same case study with our RF results might shed light on the significance of Source 4.

Prediction intervals are expected to be wider in the case of model structural errors as the relationship between the covariates and soil properties will likely be weaker.

6.4. Communication of uncertainty

This study makes it clear that the SoilGrids texture layers have limited accuracy, particularly for NZ, but even in NL there are areas with high uncertainty (Fig. 3). Lark et al. (2022) point out that users of spatial information may not benefit from the uncertainty information that is generally supplied, i.e. global model statistics and a prediction interval or variance.

The SoilGrids layers include prediction interval ratio (PIR) as an uncertainty metric. This is the ratio of PIW_{90} and the predicted median. On the soilgrids.org website, this value is associated with the labels Low to High – but the conversion of the ratio value to a label varies for each soil property. We note that this metric is not suitable for a bounded variable such as a proportion, as the metric is very sensitive to the median value, making it very hard to derive a consistent interpretation of the metric. For example, a PIR value of 1 is achieved with a PIW_{90} of 10 and a median of 10, as well as with a PIW_{90} of 90 and a median of 90. The first PIW_{90} seems very reasonable, while the second PIW_{90} is very bad.

Like Poggio et al. (2021), we think the use of uncertainty classes or categories would be beneficial in communicating uncertainty to stakeholders. This might better highlight to potential users that the texture predictions in the north and west of NL are less reliable than in the sandy Pleistocene areas. Helfenstein et al. (2022) provide the accuracy rating classes that were devised in a GlobalSoilMap (GSM) meeting. The thresholds for each class are reproduced in Table 5.

We have interpreted this as meaning that a pixel will be assigned class A if $\frac{PIW_{90}}{pred} < 0.4$, AA if $\frac{PIW_{90}}{pred} < 0.25$, and AAA if $\frac{PIW_{90}}{pred} < 0.15$ (Dominique Arrouays, pers. comm. 31 May 2024). The A rating is not achieved anywhere in NL or NZ for clay 5–15 cm depth. Note that defining accuracy thresholds in relative terms (ratio of prediction interval width and mean prediction) suffers the same problem as PIR in that it is very dependent on the prediction. There is a greater chance of achieving a good rating where the prediction is high. This can be seen in the sand map (5–15 cm) where only the high sand areas in NL achieve an A or AA rating (Fig. 12) while the PIW_{90} map for sand has similar values in the low and high sand areas, as shown in Fig. 9.

We suggest that the GSM accuracy classes be redefined to be based on the absolute magnitude of PIW_{90} for soil properties that are bounded, e.g. proportions. Just as the PIR metric is not suitable for texture, neither are accuracy classes based on a relative measure. Using PIW_{90} thresholds of 10, 25 and 50 (Table 5) results in the accuracy maps in Fig. 13 and Suppl. File Figures S.4 and S.5. Note that we have also chosen to use more descriptive labels. The selection of threshold values seemed useful for the NL and NZ case studies – but needs further evaluation.

The AOA concept could be a useful approach to highlight areas that might be less reliable due to them being dissimilar to the feature space of the training data. However, this requires a computational effort by the provider – and as discussed above, is not currently possible for SoilGrids (Laura Poggio, pers. comm. May 2022). AOA has also been shown to be weakly correlated with predictive accuracy in a study of four African countries (Hateffard et al., 2024).

One of the clearest statements about fitness for use of the SoilGrids maps is found in Poggio et al. (2021, pg 228):

Table 5
Accuracy thresholds for GlobalSoilMap (GSM) Tier 4 texture products (based on Helfenstein et al., 2022) and our proposed revision of the classification.

GlobalSoilMap		Revised	
None		Poor	$PIW_{90} > 50$ %
A	Mean ± 40 % (Mean)	Marginal	$PIW_{90} \leq 50$ %
AA	Mean ± 25 % (Mean)	Moderate	$PIW_{90} \leq 25$ %
AAA	Mean ± 15 % (Mean)	Good	$PIW_{90} \leq 10$ %

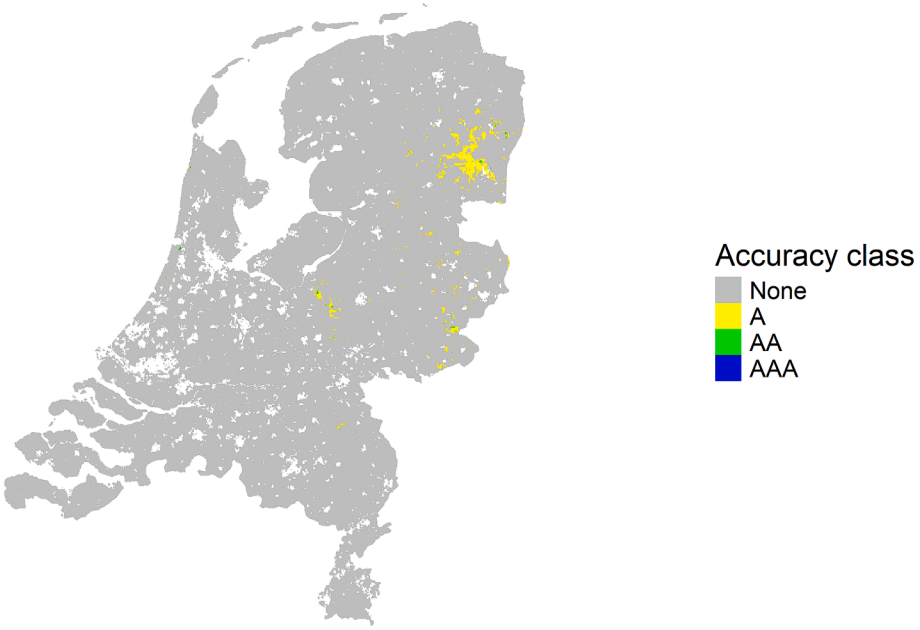


Fig. 12. GlobalSoilMap (GSM) accuracy classes for sand 5–15 cm depth in the Netherlands. We have shown the map of sand as the clay map has a quality rating of 'None'.

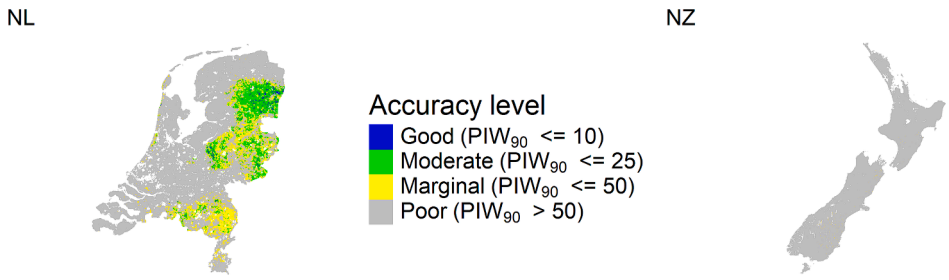


Fig. 13. Maps of accuracy for clay 5–15 cm depth in the Netherlands (NL) and New Zealand (NZ) based on 90 % prediction interval width (PIW_{90}) under our Revised accuracy classification.

> … *it should be realised that SoilGrids250m predictions are not meant for use at a detailed scale,* i.e. *at the subnational or local level* …

It seems to us that the SoilGrids products are potentially useful for global or large regional use, e.g. Europe or Asia, where the focus is on broad patterns, but may be inadequate for use in modelling at the national scale. This could be made clearer in the metadata and SoilGrids website. Arrouays et al. (2020) suggest that more information is needed, such as factsheets explaining uncertainty from a practical point of view and user instructions, along with tools to help visualise and understand uncertainty. They also suggest that minimum standards are required for DSM products to avoid poor information being used.

Assessing fitness for use for any specific purpose depends on a number of factors including the uncertainty. Some applications require higher accuracy than others. de Bruin et al. (2001) provide examples of how a user might assess the fitness for use given probabilistic measures of accuracy. Lark et al. (2022) suggest the use of a loss function in decision theory to quantify the cost of incorrect information.

6.5. Recommendations

6.5.1. For DSM producers

It has been accepted practice to provide metadata with data that are supplied for use by another party. The more comprehensive this is, the more users can be informed. We suggest that both technical information and interpretative information is supplied, as outlined below.

- Provide more metrics and information about the DSM analysis. Detail the covariates, provide a range of goodness-of-fit metrics and plots, graphs and maps of residuals, and covariate importance metrics. Provide results from an evaluation with an independent dataset (or cross-validation analysis). Be clear about the spatial support of the predictions.
- Put a narrative fitness-for-use comment in the metadata.
- Provide accuracy maps (e.g. Fig. 13), and layers of uncertainty information (e.g. PI for a range of quantiles – not just the 90 %). Provide the accuracy plot of PICP, and the interval score. Do not use PIR for bounded variables.
- Undertake and provide the results of an AOA analysis where this is technically possible (as recommended by Meyer and Pebesma 2020). However, we note additional research is still needed to confirm the value of AOA information.
- Provide the training data (if possible) and scripts to encourage reproducibility.
- Consider withholding maps (or parts of maps) that do not meet a defined minimum accuracy threshold, or proactively alerting users of the potential risks of using such maps.
- Ensure that training data are standardised – perhaps following GLOSOLAN standards (FAO, 2024). In the case of texture training data that are based on different sand/silt thresholds – these should be transformed to the same threshold (and analytical method) when

preparing the training datasets. Be clear in the metadata as to what this threshold definition is, as well as the analytical methods.
- Consider aggregating the information spatially, e.g. mean predictions for a larger spatial area, where this is appropriate as this can reduce the uncertainty (Wadoux and Heuvelink, 2023). However, it can be difficult to verify these predictions as typically there are no independent test data at large spatial support.

6.5.2. For end users of DSM products

Users of DSM data (or any data) are advised to consider the uncertainty information provided with a dataset of interest. The following steps will help give users a sense of the data and their reliability.

- Read the metadata which may be formal metadata as part of the spatial dataset, or an associated document or Readme file, or journal paper.
- Check the definitions and caveats. In the case of texture, ensure that sand and silt have the same definition as the use case.
- Consider the provided goodness of fit metrics. How significant is the evaluation RMSE and bias (ME)?
- Review the accuracy plot of PICP.
- Download the predictions and uncertainty information.
- Consider the map of predictions. Is the range of the predicted values consistent with other information? Or have predictions been oversmoothed to a cluster of central values (potentially indicating a poor model). Is the spatial pattern consistent with other information?
- Examine the PI (or variance) if provided. Derive the PIW. What is its range and distribution? Where (spatially or in feature space) is the PIW bigger (less reliable) or smaller (more reliable)? How does the PIW relate to the RMSE?
- If the area of interest is smaller than the dataset and the PI is large, then:
  - o consider how well is the area represented by the training data. Are the main soil formation processes in the smaller area of interest reflected in the map?
  - o Is AOA information supplied? Does this indicate areas of lower reliability – do these areas have wider PI?
  - o If the training data can be accessed, it can be useful to derive goodness of fit metrics for the area of interest if this is a subset of the dataset. But beware of data leakage. Are these 'local' metrics reasonable? Look at the pattern of the residuals (spatial and in feature space).
  - o If possible, use an independent dataset – are the goodness-of-fit metrics worse (indicating overfitting)? Look at the pattern of the residuals (spatial and in feature space).
- Bearing in mind what has been learned from the steps above, consider the sensitivity to errors given the purpose for using the data, and the consequences of getting it wrong. This can be achieved by using uncertainty propagation techniques and decision theory (e.g. Crosetto and Tarantola, 2001; Heuvelink et al., 1989; Lark et al., 2022).

7. Conclusion

This study examined the uncertainty in the soil texture information provided by SoilGrids for NL and NZ. Our learnings from working through the four objectives are as follows. The uncertainty information was too pessimistic for NL and indicated varied levels of uncertainty, whereas the information was too optimistic for NZ and indicated poor predictions. A range of maps and graphs were used to explore the uncertainty information, along with information from the training dataset and an independent dataset. This confirmed our initial assumption that NL would be better modelled than NZ. But we were surprised at the high uncertainty in parts of NL, and the very poor results in NZ. Contrary to our hypothesis, some of the NL predictions were more uncertain than the NZ predictions. The graphical analyses allowed only a limited identification of the four sources of uncertainty, but were quite insightful in helping us to better understand the accuracy and reliability of the information. As a result of these insights, a set of recommendations was made for both producers and consumers of DSM products. The research reinforced our view that more effort and training is needed to ensure DSM data are used appropriately.

CRediT authorship contribution statement

Linda Lilburne: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Conceptualization. **Anatol Helfenstein:** Writing – review & editing, Software, Data curation. **Gerard B.M. Heuvelink:** . **Andre Eger:** Writing – review & editing, Visualization.

Declaration of competing interest

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

Data availability

The authors do not have permission to share data.

Acknowledgements

We thank Nathan Odgers and Lauren O'Brien for their assistance with the NZ soil data and R code, and Helen O'Leary for editing. We are indebted to the ISRIC team: Laura Poggio, Niels Batjes and Maria Ruiperez-Gonzales for their assistance with SoilGrids. Two anonymous reviewers provided very useful comments. This work was supported by the Ministry of Business, Innovation and Employment (MBIE), New Zealand, Strategic Science Investment Fund.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geoderma.2024.117052.

References

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 Reg. 20, e00255. http://www.sciencedirect.com/science/article/pii/S2352009420300043.

Bahri, H., Raclot, D., Barbouchi, M., Lagacherie, P., Annabi, M., 2022. Mapping soil organic carbon stocks in Tunisian topsoils. Geoderma Reg. 30, e00561. https://www.sciencedirect.com/science/article/pii/S2352009422000815.

Batjes, N., van Oostrum, A., 2023. World Soil Information Service (WoSIS) -Procedures for standardizing soil analytical method descriptions., ISRIC - World Soil Information, Wageningen, https://www.isric.org/sites/default/files/WoSIS_procedures_for_standardizing_soil_analytical_method_descriptions_2023.pdf.

Batjes, N.H., Ribeiro, E., van Oostrum, A., 2020. Standardised soil profile data to support global mapping and modelling (WoSIS snapshot 2019). Earth Syst. Sci. Data 12 (1), 299–320. https://essd.copernicus.org/articles/12/299/2020/.

Behrens, T., Schmidt, K., MacMillan, R.A., Viscarra Rossel, R.A., 2018. Multiscale contextual spatial modelling with the Gaussian scale space. Geoderma 310, 128–137. https://www.sciencedirect.com/science/article/pii/S0016706117311175.

Brouwer, F., de Vries, F., Walvoort, D., 2018. Basisregistratie Ondergrond (BRO) actualisatie bodemkaart : Herkartering van de bodem in Flevoland. WOt technical report 143, WOT Natuur & Milieu, WUR Wageningen, https://library.wur.nl/WebQuery/wurpubs/549064.

Brouwer, F., Assinck, F., Harkema, T., Teuling, K., Walvoort, D., 2023. Actualisatie van de bodemkaart in degemeente Vijfheerenlanden: herkartering van de verbreiding van veen. WOt-rapport 151, WOT Natuur & Milieu Wageningen, Wageningen, https://research.wur.nl/en/publications/actualisatie-van-de-bodemkaart-in-degemeente-vijfheerenlanden-her.

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. https://bsssjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2389.2011.01364.x.

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. https://www.sciencedirect.com/science/article/pii/S0048969718345868.
Cramer, M.D., Wootton, L.M., van Mazijk, R., Verboom, G.A., 2019. New regionally modelled soil layers improve prediction of vegetation type relative to that based on global soil models. Divers. Distrib. 25 (11), 1736–1750. https://doi.org/10.1111/ddi.12973.
Crosetto, M., Tarantola, S., 2001. Uncertainty and sensitivity analysis: tools for GIS-based model implementation. Int. J. Geogr. Inf. Sci. 15 (5), 415–437.
de Bakker, H., Schelling, J., 1989. Systeem van bodemclassificatie voor Nederland: de hogere niveaus: With Engl. summary: A system of soil classification for the Netherlands, Second revised edition ed. Centrum voor Landbouwpublikaties en Landbouwdocumentatie, Wageningen, the Netherlands.
de Bruin, S., Bregt, A., Ven, M.v.d., 2001. Assessing fitness for use: the expected value of spatial data sets. Int. J. Geogr. Inf. Sci. 15 (5), 457–471. https://doi.org/10.1080/13658810110053116.
de Vries, W., Kros, J., Oenema, O., de Klein, J., 2003. Uncertainties in the fate of nitrogen II: A quantitative assessment of the uncertainties in major nitrogen fluxes in the Netherlands. Nutr. Cycl. Agroecosyst. 66 (1), 71–102.
Debonne, N., Bürgi, M., Diogo, V., Helfenstein, J., Herzog, F., Levers, C., Mohr, F., Swart, R., Verburg, P., 2022. The geography of megatrends affecting European agriculture. Glob. Environ. Chang. 75, 102551. https://www.sciencedirect.com/science/article/pii/S0959378022000899.
Edelmann, C.H., 1950. Soils of the Netherlands. North-Holland Publishing Company, Amsterdam.
Erkens, G., van der Meulen, M.J., Middelkoop, H., 2016. Double trouble: subsidence and CO2 respiration due to 1,000 years of Dutch coastal peatlands cultivation. Hydrgeol. J. 24, 551–568.
FAO, 2024. GLOSOLAN Standard Operating Procedures (SOPs), https://www.fao.org/global-soil-partnership/glosolan-old/soil-analysis/standard-operating-procedures/en/.
Goovaerts, P., 2001. Geostatistical modelling of uncertainty in soil science. Geoderma 103, 3–26.
GSM, 2015. Specifications Tiered GlobalSoilMap products, https://www.isric.org/sites/default/files/GlobalSoilMap_specifications_december_2015_2.pdf.
Han, S.Y., Filippi, P., Singh, K., Whelan, B.M., Bishop, T.F.A., 2022. Assessment of global, national and regional-level digital soil mapping products at different spatial supports. Eur. J. Soil Sci. 73 (5), e13300. https://bsssjournals.onlinelibrary.wiley.com/doi/abs/10.1111/ejss.13300.
Hartemink, A., 2006. Classification system: Netherlands. In: Lal, R. (Ed.), Encyclopedia of Soil Science. Taylor & Francis, New York, pp. 265–268.
Hateffard, F., Steinbuch, L., Heuvelink, G.B.M., 2024. Evaluating the extrapolation potential of random forest digital soil mapping. Geoderma 441, 116740. https://www.sciencedirect.com/science/article/pii/S0016706123004172.
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 (No. 3010). Wageningen Environmental Research, Wageningen.
Helfenstein, A., Mulder, V.L., Heuvelink, G.B.M., Okx, J.P., 2022. Tier 4 maps of soil pH at 25 m resolution for the Netherlands. Geoderma 410. https://edepot.wur.nl/561488.
Helfenstein, A., Mulder, V.L., Hack-ten Broeke, M.J.D., van Doorn, M., Teuling, K., Walvoort, D.J.J., Heuvelink, G.B.M., 2024a. BIS-4D: mapping soil properties and their uncertainties at 25 m resolution in the Netherlands. Earth Syst. Sci. Data 16, 2941–2970. https://doi.org/10.5194/essd-16-2941-2024.
Helfenstein, A., Teuling, K., Walvoort, D.J.J., Hack-ten Broeke, M.J.D., Mulder, V.L., van Doorn, M., Heuvelink, G.B.M., 2024b. Georeferenced point data of soil properties in the Netherlands. Wageningen University, Https:/. https://doi.org/10.4121/c90215b3-bdc6-4633-b721-4c4a0259d6dc.
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., Mendes de Jesus, J., Heuvelink, G.B.M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS One 12 (2), e0169748.
Heuvelink, G., 2018. Uncertainty and uncertainty propagation in soil mapping and modelling. In: McBratney, A., Minasy, B., Stockmann, U. (Eds.), Pedometrics. Progress in Soil Science. Springer, pp. 439–461.
Heuvelink, G.B.M., Burrough, P.A., Stein, A., 1989. Propagation of error in spatial modelling with GIS. Int. J. Geogr. Inf. Syst. 3, 303–322.
Hewitt, A., Lowe, D.J., Balks, M.R., 2021. The Soils of Aotearoa New Zealand. Springer International Publishing.
ISRIC, 2024. FAQ - WoSiS: Towards Soil Harmonisation, https://www.isric.org/explore/wosis/faq-wosis#Towards_soil_harmonisation.
IUSS Working Group WRB, 2015. World reference base for soil resources 2014, Update 2015 International soil classification system for naming soils and creating legends for soil maps. . World Soil Resources Reports No. 106, Rome, https://www.fao.org/3/i3794en/I3794en.pdf.
Kasraei, B., Heung, B., Saurette, D.D., Schmidt, M.G., Bulmer, C.E., Bethel, W., 2021. Quantile regression as a generic approach for estimating uncertainty of digital soil maps produced from machine-learning. Environ. Model. Softw. 144, 105139. https://www.sciencedirect.com/science/article/pii/S1364815221001821.
Kaufman, S., Rosset, S., Perlich, C., 2011. Leakage in data mining: formulation, detection, and avoidance, Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. Association for Computing Machinery, San Diego, California, USA, pp. 556–563.
Keesstra, S.D., Bouma, J., Wallinga, J., Tittonell, P., Smith, P., Cerdà, A., Montanarella, L., Quinton, J.N., Pachepsky, Y., van der Putten, W.H., Bardgett, R.D., Moolenaar, S., Mol, G., Jansen, B., Fresco, L.O., 2016. The significance of soils and soil science towards realization of the United Nations Sustainable Development Goals. Soil 2 (2), 111–128. https://www.soil-journal.net/2/111/2016/.
Lark, R.M., Chagumaira, C., Milne, A.E., 2022. Decisions, uncertainty and spatial information. Spatial Statistics 50, 100619. https://www.sciencedirect.com/science/article/pii/S2211675322000161.
Ma, Y., Minasny, B., McBratney, A., Poggio, L., Fajardo, M., 2021. Predicting soil properties in 3D: Should depth be a covariate? Geoderma 383, 114794. https://www.sciencedirect.com/science/article/pii/S0016706120325490.
Malone, B.P., McBratney, A.B., Minasny, B., 2011. Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. Geoderma 160 (3–4), 614–626. http://www.sciencedirect.com/science/article/pii/S0016706110003666.
Manaaki Whenua - Landcare Research, 2020. National Soils Data Repository (NSDR). Manaaki Whenua - Landcare Research. https://doi.org/10.26060/gcxk-e905.
Manaaki Whenua - Landcare Research, 2020a. FSL Particle Size Classification. In: Manaaki Whenua - Landcare Research (Ed.), https://lris.scinfo.org.nz/layer/48112-fsl-particle-size-classification/.
McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117 (1–2), 3–52. https://www.sciencedirect.com/science/article/pii/S0016706103002234?via%3Dihub.
Meinshausen, N., 2006. Quantile Regression Forests. J. Mach. Learn. Res. 7, 983–999.
Meyer, H., Pebesma, E., 2020. Predicting into unknown space? Estimating the area of applicability of spatial prediction models, http://arxiv.org/abs/2005.07939.
Minasny, B., McBratney, A.B., 2001. The Australian soil texture boomerang: a comparison of the Australian and USDA/FAO soil particle-size classification systems. Soil Res. 39 (6), 1443–1451. https://www.publish.csiro.au/paper/SR00065.
Minasny, B., McBratney, A.B., 2016. Digital soil mapping: A brief history and some lessons. Geoderma 264, 301–311. http://www.sciencedirect.com/science/article/pii/S0016706115300276.
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. https://www.sciencedirect.com/science/article/pii/S001670611530063X.
Nemes, A., Rawls, W.J., 2006. Evaluation of different representations of the particle-size distribution to predict soil water retention. Geoderma 132 (1), 47–58. https://www.sciencedirect.com/science/article/pii/S0016706105001291.
Nemes, A., Wösten, J.H.M., Lilly, A., Oude Voshaar, J.H., 1999. Evaluation of different procedures to interpolate particle-size distributions to achieve compatibility within soil databases. Geoderma 90 (3), 187–202. https://www.sciencedirect.com/science/article/pii/S0016706199000142.
Piikki, K., Wetterlind, J., Söderström, M., Stenberg, B., 2021. Perspectives on validation in digital soil mapping of continuous attributes—A review. Soil Use Manag. 37 (1), 7–21. https://onlinelibrary.wiley.com/doi/abs/10.1111/sum.12694.
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. https://soil.copernicus.org/articles/7/217/2021/.
Radočaj, D., Jurišić, M., Rapčan, I., Domazetović, F., Milošević, R., Plaščak, I., 2023. An Independent Validation of SoilGrids Accuracy for Soil Texture Components in Croatia. Land 12 (5), 1034. https://www.mdpi.com/2073-445X/12/5/1034.
Rossiter, D.G., Poggio, L., Beaudette, D., Libohova, Z., 2022. How well does digital soil mapping represent soil geography? An Investigation from the USA. SOIL 8 (2), 559–586. https://soil.copernicus.org/articles/8/559/2022/.
Samuel-Rosa, A., Heuvelink, G.B.M., Vasques, G.M., Anjos, L.H.C., 2015. Do more detailed environmental covariates deliver more accurate soil maps? Geoderma 243–244, 214–227. https://www.sciencedirect.com/science/article/pii/S001670611400456X.
Schaetzl, R.J., Thompson, M., 2015. Soils : genesis and geomorphology. Cambridge University Press, New York, USA.
Schmidinger, J., Heuvelink, G.B.M., 2023. Validation of uncertainty predictions in digital soil mapping. Geoderma 437, 116585. https://www.sciencedirect.com/science/article/pii/S0016706123002628.
Shrestha, D.L., Solomatine, D.P., 2006. Machine learning approaches for estimation of prediction interval for the model output. Neural Netw. 19 (2), 225–235. https://www.sciencedirect.com/science/article/pii/S0893608006000153.
Soil Survey Staff, 1999. Agriculture Handbook 436. 2nd ed. United States Department of Agriculture Natural Resources Conservation Service.
Szatmári, G., Pásztor, L., 2019. Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms. Geoderma 337, 1329–1340. https://www.sciencedirect.com/science/article/pii/S0016706117318499.
Vaysse, K., Lagacherie, P., 2017. Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma 291, 55–64. https://www.sciencedirect.com/science/article/pii/S001670611631059X.
Vos, P., Meulen, M., Weerts, H. v. d., Bazelmans, J., 2020. Atlas of the Holocene Netherlands, landscape and habitation since the last ice age. University Press, Amsterdam https://www.cultureelerfgoed.nl/onderwerpen/bronnen-en-kaarten/overzicht/paleografische-kaarten.
Wadoux, A.-M.-J.-C., Heuvelink, G.B.M., 2023. Uncertainty of spatial averages and totals of natural resource maps. Methods Ecol. Evol. 14 (5), 1320–1332. https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.14106.
Wadoux, A.M.J.C., Walvoort, D.J.J., Brus, D.J., 2022. An integrated approach for the evaluation of quantitative soil maps through Taylor and solar diagrams. Geoderma

405, 115332. https://www.sciencedirect.com/science/article/pii/S0016706121004122.
Wageningen UR-Alterra, 2006. Grondsoortenkaart 2006 - Simplified Soil Map of the Netherlands. Wageningen UR- Alterra, https://doi.org/10.17026/dans-xky-fsk5.
Webster, R., Oliver, M., 2007. Geostatistics for environmental scientists. Statistics in practice, 2nd edition ed. Wiley.