Geoderma 377 (2020) 114579

Contents lists available at ScienceDirect

Geoderma

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

Improvements to the Australian national soil thickness map using an integrated data mining approach

Brendan Malone[a,*], Ross Searle[b]

[a] CSIRO Agriculture and Food, Black Mountain, ACT, Australia
[b] CSIRO Agriculture and Food, St Lucia, QLD, Australia

ARTICLE INFO

Handling Editor: Budiman Minasny

Keywords:
Digital soil mapping
Soil thickness
National soil infrastructure
Censored data
Data mining
Integrated modelling

ABSTRACT

Soil thickness is not easily measured in situ, making it also a challenging variable to reliably map. This study improves on previous digital mapping of soil thickness across Australia using an approach suited to the continent's unique pedo-geomorphic history. Leveraging three large, in situ observation datasets and a wide range of spatial environmental variables, we developed three models depicting rock outcrops, intermediate and deep soils respectively. Our modelling approach addressed right-censored data, which is a common attribute of soil thickness data, and we applied an iterative, data re-sampling framework to quantify prediction uncertainties. We integrated the three models to create soil thickness maps and associated products of soil thickness exceedance probabilities. Using data excluded from model calibrations, we achieved an overall accuracy of 99% for the binary outcome rock outcrops model, and 85% for the binary outcome deep soils model. Modelling soil thickness of shallow to deep soils resulted in a concordance coefficient of 0.77. Of all the environmental variables considered in this study, those associated with climate data (including topo-climate) were consistently the most often used and important. We associate this finding with the direct and indirect effects of climate on biota and weathering of parental materials along with other factors driving spatial heterogeneity in soil thickness across Australia. While the products generated by this research are not without error, the overall pattern of soil thickness is consistent with previous observations from historical soil surveys across Australia and the results are demonstrably more skilful than previous digital soil mapping efforts.

1. Introduction

Soil thickness, as defined in Australia (National Committee on Soil and Terrain, 2009), is the length of distance from the soil surface to para-lithic or lithic contact (i.e. the A and B soil horizons). The term is often used synonymously with soil depth. Maps of soil thickness have a wide range of uses as inputs to land capability and crop/species suitability assessments, in models of biophysical dynamics such as carbon and water storage and balance, or infrastructure planning, to name just a few. However, accurate mapping of soil thickness is fraught with numerous technical difficulties. Besides a dearth of point observation data relative to other measurable soil variables such as pH or total soil carbon (Searle, 2015), the most significant technical difficulty is the problem of appropriately handling right-censored data (Chen et al., 2019). Here, right-censored data arises when the observed soil thickness does not correspond to the actual soil thickness. This happens for several reasons: 1) soil sampling often does not involve the need to determine depth to lithic contact; 2) in the absence of exposed soil pits or road cuttings, soil sampling to depth is costly and requires specialised equipment; 3) even with specialist equipment there is a maximum feasible sampling depth. These sampling constraints manifest in legacy data bases where for example in Australia we find 69% of soil site observations do not go beyond 1.5 m and 62% of soil site observations that do not have C or R type horizon descriptors. These descriptors are used in Australia to describe layers below the soil of consolidated or unconsolidated material, usually partially weathered, little affected by pedogenic processes, and either like or unlike the material from which the soil was presumably formed (National Committee on Soil and Terrain, 2009).

Compared to other parts of the world, Australian soils are relatively deep because of the age and weathering processes these landscapes have undergone (Young and Young, 2001). For example, some alluvial floodplain soils west of the Australian Great Dividing Range in the Riverina region (eastern Australia) have been measured at more than 20 m thick (Chen, 1997). While similar observations of soil thickness are known anecdotally in numerous other places, in different settings

* Corresponding author.
E-mail address: brendan.malone@csiro.au (B. Malone).

https://doi.org/10.1016/j.geoderma.2020.114579
Received 3 December 2019; Received in revised form 29 January 2020; Accepted 6 July 2020

and contexts, this knowledge is not well represented in digital soil mapping of soil thickness across Australia (Viscarra Rossel et al., 2014), which predicts a maximum thickness of 1.84 m. The same model predicts a minimum soil thickness of 0.1 m (mean of 0.82). In later work, Wilford et al. (2016) presented work on mapping the regolith thickness across Australia. While both the Viscarra Rossel et al. (2014) and Wilford et al. (2016) approaches were largely framed around machine learning predictive modelling approaches, the latter approach incorporated more data and of the type to better distinguish shallow and deep profiles, for example from rock outcrop observations and bore hole data from the National Groundwater Information System (NGIS) database, which have also been used in global soil thickness digital mapping models in Shangguan et al. (2017). However, what is difficult to disentangle from regolith thickness mapping is the relative thicknesses of the two constituent layers: the solum (i.e. the A and B horizons) and the saprolite. Without distinguishing these layers, it is difficult to draw meaningful comparisons between the Wilford et al. (2016) and the Viscarra Rossel et al. (2014) maps. Biological systems interact predominantly within the solum, which provides a foundation for agricultural productivity and ecosystem diversity. Given this importance of the solum, there is a clear need to firm up our understanding of its variation thickness across the Australian landscape, using the best available data, knowledge and tools.

The spatial prediction of soil thickness has received considerable attention because of its importance for quantifying various soil functions. Mechanistic based approaches have included those described by Minasny and McBratney (1999), McKenzie et al. (2003) and Pelletier and Rasmussen (2009). In these studies, mechanistic understanding of weathering and sediment transport processes, including interaction with topographic position is used to estimate soil thickness. The parametrisation of these models is informed by empirical observations, but this approach cannot be generalised across large spatial extents. Implied assumptions about how the model will operate apply to the setting in which it was calibrated, outside of which extrapolations can generate spurious predictions.

Empirical approaches include those from Odeh et al. (1991) and Moore et al. (1993) among others of more recent times and provide a way of dealing with heterogeneity. This is because explicit relationships between observations (target variables) and associated predictor variables are defined via some model structure that could range from being relatively simple (such as multiple linear regression) to rather complex (such as machine learning or neural network models). These models are good at extracting relationships from vast amounts of diverse predictor variables. This can be beneficial for explorative work when considering lesser known relationships and incorporating complex statistical relationships that mechanistic based models are unable to capture. Presuming the data with which the models are fitted are representative of the environment to be mapped, successful outcomes are to be expected if the relationships between target and predictor variables are particularly strong. Such models however are not immune to the model extensibility issue that mechanistic based models suffer from, but this could potentially be overcome by increasing the number and diversity of data in the model, as well as calibrating the model to the intended spatial geographical extent (assuming there is sufficient data within the newly defined extent of course).

In any case, while one could potentially entertain any number or type of predictor variables, it is helpful, particularly for spatial modelling of soil phenomena, to have an underlying soil-landscape concept with which to frame and constrain the empirical models. This is established through the SCORPAN function introduced in McBratney et al. (2003) which borrows from earlier established concepts (namely Jenny's CLORPT function) of soil-landscape relationships. That is, soil at a particular location is the result of various chemical and physical processes due to where it is in the landscape, affected by the prevailing climate (C), the biology such as vegetation and land management imprint (O), relief (R), underlying parent material (P) and its age (T). As well as these factors, McBratney et al. (2003) contend that we can also predict soil properties from other soil phenomena, and that there is a spatial context to soil formation, for example distance to a certain landmark or geographical feature. These additional factors are the S and N factors of the SCORPAN function. What sets SCORPAN apart is its purely empirical nature and that it can be easily melded to exploit capability in modelling approaches and digital environmental data availability. It is the SCORPAN function that provides the framework and basis for what is commonly called digital soil mapping (DSM). Spatial modelling of soil thickness is in the purview of DSM. This is useful because through using spatially exhaustive environmental variables such as from digital elevation models, remote sensing data, geological information and other environmental data, together with soil observations, one can derive a soil-landscape relationship understanding via this quantitative model framework. For soil thickness mapping this ultimately enables the ability to bring together as much environmental information as feasibly possible, and then using (especially in more recent times) quite complex model structures to get a comprehensive picture of its spatial variability. This is demonstrated in the earlier cited studies by Viscarra Rossel et al. (2015), Wilford et al. (2016) and Shangguan et al. (2017). Soon we may expect to see a coupling of mechanistic and empirical methods such as that described in Ma et al. (2019) as a way of exploiting and integrating the benefits of both approaches.

DSM-based studies that focus on addressing the right-censored data issue are relatively scarce. Forays into this space have been undertaken by Kempen et al. (2015) for predicting soil peat thickness in the Netherlands. They used simulation wherein, for every iteration, a certain amount of thickness was added to the right-censored data by drawing a value from a beta distribution with given parameters. The simulated data was combined with interval data (similarly subjected to a simulation sampling approach) and actual measured thickness data to generate a data series that was then combined to derive probabilistic estimates of peat thickness. Rather than drawing from a beta distribution, Lacoste et al. (2016) simply treated right-censored data by adding 30 cm to those affected profiles before implementing their spatial model framework. Formalised model structures for dealing with right-censored data fall predominately into the survival analysis branch of statistics. As indicated in Chen et al. (2019), survival analysis is widely used in medical and engineering research where one application of these models is for analysing the expected duration of time until one or more events happen, such as death in biological organisms or failure in mechanical systems. In their study, Chen et al. (2019) reviewed several models that can be used in survival analysis including the Random Survival Forest model (RSF, Ishwaran et al., 2008) which they used for probabilistic mapping of soil thickness across France. This proved a successful approach and echoed sentiments from Styc and Lagacherie (2016) about such models being useful; they are objective, specific and a necessary addition to the suite of modelling approaches used for DSM.

Our own forays with RSF were not hugely successful. This was mainly because we had difficulty constructing and sourcing a RSF computational workflow to fit a model with our to-be-described ~150,000 observations. Secondly, even negating the presence of right censored data (in order to fulfil an explorative exercise) with our data mix and just trying to model soil thickness using a machine learning model structure where we can relax many assumptions around things like distributions, we found such models incapable of generating a multimodal distribution of the type observed when one compiles a database of soil thickness observations. On balance after consideration of the sorts and types of data at our disposal, a mechanistic based model was not appropriate. Instead, the empirical approach was determined as most suitable, but something more tailored to the Australian soil context rather than using a standard modelling approach. The present study describes the process we ultimately undertook to adequately model the spatial distribution of deep soils, rock outcrops, skeletal soils and everything in-between. Our approach integrates observational data

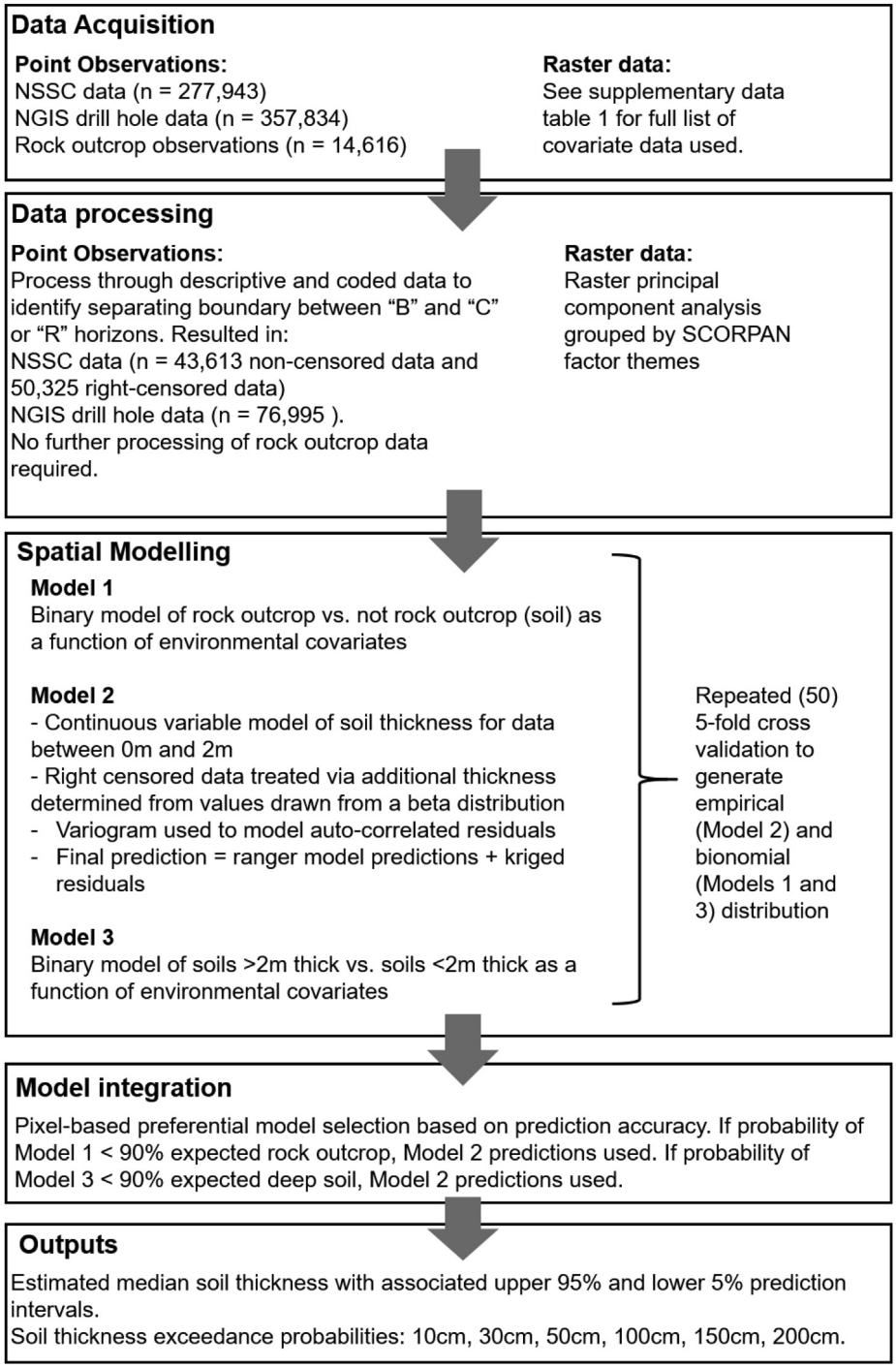

Fig. 1. Summarised workflow of methodological framework for digital mapping of soil thickness across Australia.

from different sources and implements a piecewise or multi-model approach where separate models estimate shallow, intermediate and deep soils independently. These models are integrated via a basic gating routine to generate continental probabilistic maps of soil thickness. We propose that this approach is well suited for the Australian context. Here we describe the number of processing steps undertaken to prepare the data for use and the spatial modelling framework applied.

2. Materials and methods

This study encompasses the continent of Australia, where there is a need to consistently map the spatial variation in soil thickness. We used publicly available datasets and applied integrated modelling and bespoke computational workflows to generate a final product (as shown in Fig. 1).

2.1. Datasets

2.1.1. Observational data

We harmonised three point observation datasets in this study (Fig. 2).

1. The Australian National Soil Site Collation (NSSC, Searle, 2015). The database has information for 277,943 soil profile observations,

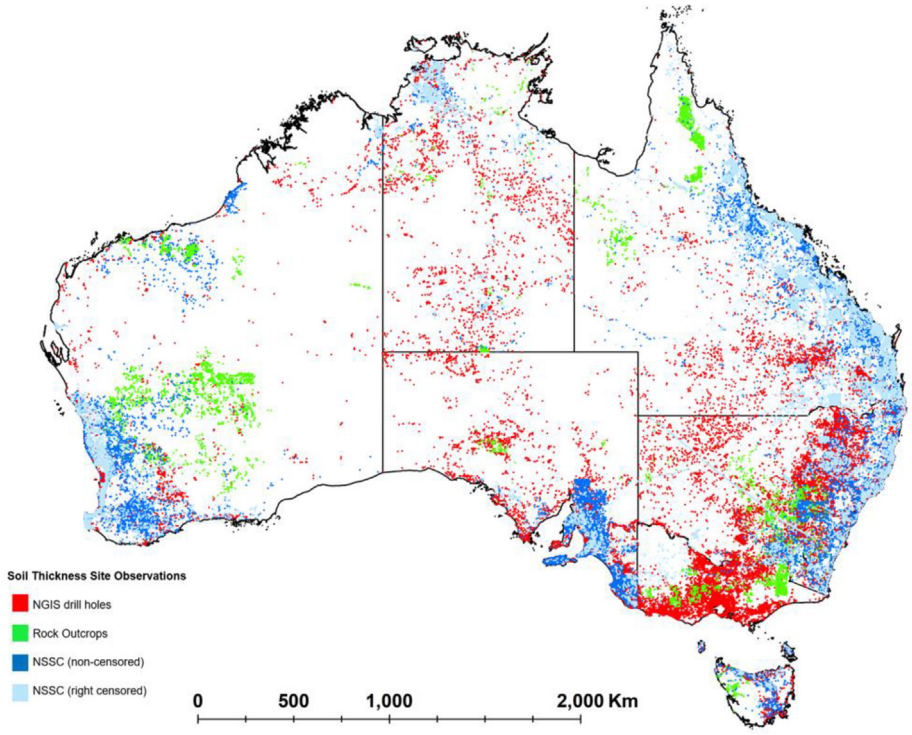

Fig. 2. Distribution of soil thickness point observation colour coded by the source of the data.

describing 1,019,823 horizons. These data are distributed across Australia, with the compilation being the result of collaboration between State and National agencies, and universities.

2. National Groundwater Information System (NGIS) database of bore hole data. This spatial database holds nationally consistent information about bores that were drilled as part of the Bore Construction Licensing Framework (http://www.bom.gov.au/water/groundwater/ngis/). The database contains 357,834 drill hole locations with associated lithology, bore construction and hydrostratigraphic records.
3. The Rock Properties database provided by Geoscience Australia give the locations of sampled rock outcrops across Australia (http://www.ga.gov.au/scientific-topics/disciplines/geophysics/rock-properties). Filtering this dataset on the sample types of "outcrop sample" resulted in 14,616 rock outcrop locations within areas where relief > 300 m.

2.1.2. Environmental covariate data

Most covariate data are the same as used by Viscarra Rossel et al. (2015) in generating the Soil Landscape of Australia (SLGA) digital soil information infrastructure. Each variable is described in Table 1 of the supplementary information. For consistency with end user requirements, we substituted the climatic data used in Viscarra Rossel et al. (2015) with a new suite of surfaces and associated solar energy and topo-climatic information. Climate surfaces for the present study were based on the ANUCLIM 6.1 (Xu and Hutchinson, 2011) 30-year average climate surfaces for Australia (1976–2005), with elevational lapse rate correction applied over the 3 arc second (~90 m) processed and hydrologically corrected STRM digital elevation model (DEM, Gallant et al. (2012)). Radiative correction derived from the same DEM was applied to radiation and maximum temperature before calculation of evaporation, using the CSIRO TerraFormer software (Harwood et al. 2014). Summary statistics for each variable were calculated including variables described in Harwood et al (2014) and Williams et al. (2012). Solar energy and topo-climatic adjusted variables were derived using the same DEM and computed with the SRAD model (Wilson and Gallant, 2000). In total 47 'climatic' surfaces were compiled for this study. Covariates related to relief were derived from the same DEM. In total, 16 primary and secondary terrain variables were available. Covariates related to organisms were derived from remote sensing platforms including the AVHRR, MODIS and Landsat satellite platforms and include vegetation indices with themes such as persistent vegetation, time series fractional vegetation cover (and their statistical moments) and FPAR (fraction of photosynthetically active radiation). A total of 21 'Organism' layers were compiled for this study. Parent material and associated soil covariates included coverages of the gamma radiometric data regions of interest (thorium, potassium and uranium), a material weathering index (Wilford, 2012), and maps of soil secondary clays (kaolinite, illite, smectite) from Viscarra Rossel (2011). In total 14 layers of the parent material and soil theme were compiled. If not already done so, each of the 98 environmental variables were reprojected from their native resolutions to WGS84 (EPSG:4326) projection with 3 arc second grid cell resolution. The bilinear interpolation method was used for this reprojection task. Each covariate data layer had the extent:

Table 1
Summary of data available and used for spatial modelling of soil thickness in Australia. NSSC (national soil site collation), NGIS (national groundwater information system), GA rocks (Geoscience Australia rock properties database).

Dataset	Number of available site data (before filtering steps)	Number of site data suitable for spatial modelling
NSSC	277,943	43,613 (non-censored); 50,325 (right-censored)
NGIS	357,834	76,995
GA rocks	14,616	14,616

112.99958°E–153.99958°E; 10.0004°S–44.00042°S.

2.2. Processing of datasets

2.2.1. Observational data

A significant amount of data processing was needed to harmonise and extract observed soil thickness values from both the NSSC and NGIS databases. The common goal for both datasets was to identify the depth of the bottom boundary of the soil before it transitions into consolidated or semi-consolidated lithic material or rock. For the NSSC database this identification process was done by analysing the horizon codes that were assigned to layers within each recorded soil profile. Horizon notation for all profiles followed the designation system as detailed in National Committee on Soil and Terrain (2009). The data challenge was to record the depth within a soil profile in the first instance or occurrence of a 'C' or 'R' horizon which directly underlaid either an A, B, O or P horizon. Our analysis also specified that the tops of transitional horizons such as CB, AC, BR and the many other related variants be classified as non-soil material, and in most cases, the depth at which these horizons occurred was deemed the soil thickness. Our analysis revealed a total of 3485 different horizon name designations, 822 of which contained C or R descriptors. The basic workflow of the analysis performed on the NSSC data was as follows:

For each profile;

1. Does soil profile have horizon descriptors?
   a. Yes: move to step 2
   b. No: Exclude profile from further analysis and move to next soil profile
2. Does soil profile contain a 'C' or 'R' horizon (or in other words, does the soil profile contain one of the 822 horizon designations that contain a 'C' or 'R')?
   a. Yes: Record the top layer depth in the first instance that such a horizon occurs. This is the recorded soil thickness.
   b. No: Record the maximum depth of the profile and indicate that this profile is right-censored.

This analysis revealed 43,613 credible soil profiles with recorded lithic or paralithic contact and a further 172,704 credible soil profiles that were right-censored. 61,626 soil profiles were excluded that did not have horizon descriptors. A further data reduction procedure was applied to the right-censored data to exclude soil profiles where the maximum recorded depth was 1.35 m. While there are few exceptions, we selected this depth from prior experience operating soil coring instruments which predominantly have a maximum substrate penetration no greater than this depth. By removing right-censored data that are less than 1.35 m, first it makes the modelling method (described below) more plausible and importantly, removes a large number of topsoil-only soil profile observations. This data reduction procedure resulted in having 50,325 right-censored soil profile observations for spatial modelling.

Processing of the NGIS database followed a similar sequence to that described in Wilford et al. (2016). That is, the boundary between regolith and fresh bedrock was designated using a query routine applied to more-or-less free-form text of bore hole layer descriptions which searched for the boundary between soil material and consolidated or semi-consolidated materials. For every bore log at each layer described, the database has variables for specifying major and minor lithologies of the drilled material. Then there is a free form text description of this material. Lithological information is often provided without associated text description and vice versa. Where no such information was provided, the data point was subsequently removed from further analysis. Single word, multi-word, and phrase-word libraries or lookup tables were manually compiled to distinguish soil material from lithic and para-lithic material. For lithologies that indicated the drilled material was soil material, some text examples included: "CLAY", "MARL", "TPSL" for clay, marl and topsoil respectively. The lithology categories also had an element of free-form language which we cross-referenced with text descriptions to confidently determine their meaning. Some example lithologies indicative of rock material included: "SHLE", "ROCK", and "LMST" for shale, rock, and limestone respectively. Some indicative words and text phrases included: "sands", "unconsolidated clay", "overburden", and "loam"; and for rock: "calcrete", "sandy siltstone", and "brecciated". The decision to allocate a layered material within a bore as either soil material or consolidated material was based on either the lithology code or text description when only one piece of information was available. Where both pieces of information were available each phrase-word library was used. This text processing to distinguish the boundary between soil material and consolidated material was an exacting procedure where any sign of ambiguity or non-match with the compiled word libraries meant a decision could not be made, and therefore the bore log in question was removed from further analysis. Ambiguity in this situation included words or phrases for soil materials mixed with those of consolidated material. The success of this text analysis procedure was highly dependent upon richly populated word libraries to capture the predominant words and phrases in order to reduce ambiguity and clearly distinguish the two categories of soil material and semi- or fully consolidated material. This analysis was necessarily iterative. Of the 357,834 bore log records in the NGIS database we extracted 76, 995 credible records that met the rigorous search criteria.

A summary of the data composition before and after filtering steps is given in Table 1.

2.2.2. Environmental covariate data

In order to get a more balanced or even contribution of SCORPAN variables in the spatial modelling, while also aiming for a simpler model configuration with minimal redundant covariates, we opted to implement a principal component analysis (PCA) of the available continuous covariate data. In our study all but one of the environmental covariates was continuous which was the geomorphons terrain classification layer which was a categorical variable of landform characterisations. The PCA was performed for each grouping of covariates for each factor of the SCORPAN function. In our case this meant performing PCA for the 47 climatic variables in one workflow, 21 organism variables in another, and 16 relief variables in another. We combined the 14 soil and parent material layers for the last PCA workflow. For each PCA, 500,000 randomly allocated point locations were distributed within the spatial data extent of the rasters. We performed a raster data extraction at each point location which resulted in a 500,000 × N (number of variables) matrix, which were subjected to PCA after the data were centred and scaled. We selected the number of PCs that cumulatively summarised at a minimum 95% of the data variation. Once done we mapped the PCs using the PCA equation and the stacked raster layers. These workflows resulted in 12, 4, 9 and 10 PCA layers for the climate, organism, relief and parent material + soil SCORPAN variables.

2.3. Spatial modelling

Upon compiling all the available soil thickness observation datasets to observe the distribution of the data, we noted it to be very non-symmetrical with a long tail, which was also quite lumpy or multi-modal (Fig. 3). There were many observations with zero or close to zero, which predominantly would have come from the rock outcrop observations, and many samples between 0 and 2 m. The long tail of the distribution spanned from about 2 m up to over 35 m (note we cut off the tail of the distribution shown in Fig. 3 to 10 m for visualisation purposes). Some initial investigative modelling work involved using a machine learning model (such as Random Forest) on the data both untransformed and transformed –square-root and natural log transform of the soil thickness data were trialled. But this yielded unsatisfactory

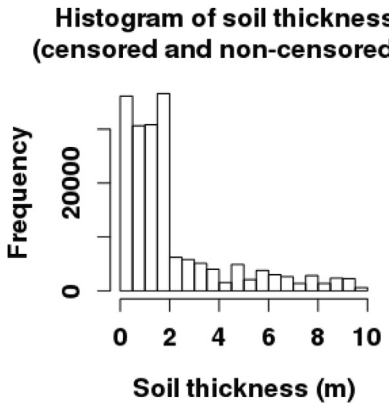

Fig. 3. Histogram of soil thicknesses of combined NSSC and NGIS datasets. Contains both censored and uncensored data. The tail of the distribution was cut off at 10 m for visualisation purposes.

results where we found the influence of the weighty long tail of the distribution resulted in overprediction when assessing the model with a test dataset. We therefore opted to use a multi-model approach which entailed three separate models for:

Model 1. Predicting the occurrence of rock outcrops.

Model 2. Predicting the thickness of soils within the 0–2 m range

Model 3. Predicting the occurrence of deep soils (soils greater than 2 m thick)

Prior to any modelling, all available points were intersected with each of the 36 PCA covariate layers to retrieve their values at those point locations. For all three models, 5-fold cross validation (repeated 50 times) was implemented in order to generate empirical distribution functions for each prediction that was made. For Models 1 and 3, the prediction is expressed as a probability of occurrence in both cases, whereas for Model 2, the predictions were expressed as the median value of the repeated 5-fold model predictions, and upper and lower bounds of the prediction distribution which corresponded to the 5th and 95th percentiles of the predicted values. We also provide exceedance probability estimates from this information for specified depths: 10 cm, 30 cm, 50 cm, 100 cm, 150 cm, and 200 cm. The exceedance probability estimate is just the estimated probability that the soil thickness at a given location exceeds the specified soil thickness threshold. Note that also for Model 2 we also modelled the spatial variation of model results with variograms, where more information about this is discussed further on.

All three models utilised the random forest (RF) data modelling algorithm (Breiman, 2001), in particular the 'Ranger' implementation of it (Wright and Ziegler, 2017) which is faster and particularly suited for high dimensional data. This model can be used for both cases of floating-point number prediction (regression) and categorical value prediction. The task for these models is to find relationships and patterns within the environmental data that optimises the prediction accuracy of a given target variable.

Models 1 and 3 used the categorical model variant of the Ranger RF which was preceded by distinguishing; for Model 1, the observations that were deemed as rock outcrops from soils. And for Model 3, distinguishing soils that were less than 2 m thick (and not rock outcrops) from soils greater than 2 m thick. Ultimately both Models 1 and 3 were binary categorical models. For Model 1 the balance of observation was 14,616 and 169,581 for rock outcrops and soils respectively. For Model 3, the balance of observations was 125,918 and 58,279 for less than 2 m and > 2 m soils respectively. 50 repeats of 5-fold CV (cross-validation) iterations of the Ranger RF model were run for each Model variant. Prior to running the models, we optimised the RF hyperparameter 'mtry' (number of variables to possibly split at in each node of the random forest model) using a purpose built cross-validation scheme that is facilitated in the caret (Kuhn et al., 2019) R package. The number of trees to grow was not optimised and set to 500 as this was computationally efficient for our computer system.

Model 2 used the regression form of the random forest model. After removing from the total data set the observations that were regarded as rock outcrops and soil greater than 2 m, there were 111,302 observations available. Of these, 67,698 had explicitly defined soil thickness values. The remaining 43,604 were right-censored data and were treated as follows. For each repeated 5-fold iteration, prior to splitting the data in calibration and validation datasets, values from a beta function were drawn at random of length 43,604. This value (between 0 and 1) was multiplied by the censored value soil thickness and then added to this same value, creating a simulated pseudo-soil thickness. In their work, Kempen et al. (2015) implemented a similar procedure when working with peat thickness data and used beta function shape parameters of 2 (a) and 5 (b) based on expert judgement. In our work we came upon the values of 2 and 5.5 for these same two parameters by experimentation. This experiment (results not shows here) entailed searching for sites that fulfilled the criteria of having observations of non-censored and right censored data relatively close to each other. Sequentially different values of a and b were trialled, and the ideal combination were the values resulting in the means of the actual soil thicknesses and pseudo-soil thicknesses being nearest to parity. We found numerous combinations of these values could achieve this requirement yet values near or close to 2 and 5.5 consistently gave the ideal outcome and so were selected for this study.

Once the simulated data were combined with actual soil thickness data, the values were square-root transformed to approximate a normal distribution. Ranger RF modelling proceeded after optimising the Hyperparameter settings as described above for the categorical modelling. Like the categorical modelling, 50 repeated 5-fold CV iterations were computed. Model 2 also entailed assessing the spatial autocorrelation of model residuals, with the intention that the spatial pattern of these errors could add further predictive skill in predicting soil thickness on top of what was captured in the environmental covariate data.

Rather than compute variograms upon each derived set of model residuals from each iteration, a far simpler approach in terms of computation was used. A small sub-hypothesis guiding the decision making was that predictive skill would not improve significantly by spatial modelling of the autocorrelated ranger RF residuals. There were two reasons why this was believed to be the case. First, the RF algorithm from experience tends to overfit or does not generalise too well. The algorithm will search for patterns in the data that maximises the accuracy. From the practical perspective, when this occurs, there is little to no spatial autocorrelation pattern amongst the residuals to work with, making the variography task somewhat redundant. Secondly, which is also informed strongly by the first is the fact that we doubted the validity of global variograms fitted at the scale at which our data existed. Much more would be gained by local fitting of variograms, but this would require significant computational resources and major modification of the overall workflow that was determined to be unfeasible for the present study.

Therefore, the simpler workflow which we describe now was implemented. This entailed, from the total combined dataset, selecting the observations that were non-censored. These data were then (with their covariates) passed into each fitted ranger RF model where model residuals were then calculated. This created for each observation a vector of residuals from which we estimated the median. We then automatically fitted a variogram to this data using the automap R package (Hiemstra et al., 2009). We then used this fitted variogram to predict a model residual layer onto the same grid used for all the covariate layers. In terms of generating maps of soil thickness, each model iteration was applied using each of the covariate layers as predictor variables, followed by adding the random forest prediction to the residual layer. We then took the median and 5th and 95th percentiles of the empirical distribution which were mapped accordingly.

For all model approaches, we report goodness of fit statistics in terms of the validation data (data withheld from each model iteration) on the basis of the average and standard deviation of the overall accuracy and Kappa statistic (for Models 1 and 3) and the root mean square error and concordance coefficient (Model 2).

All three model approaches were integrated via a simple 'if-then' pixel-based procedure. At each pixel, if Model 1 indicated the presence of rock outcrops 45 times or more out of 50 (90% of resampling iterations), the estimated soil thickness was estimated as rock outcrop, or effectively 0 cm. Similarly, for Model 3 which was the model based on prediction of deep soils (soils > 2 m deep). In no situations did we encounter both Models 1 and 3 predict in the positive on 90% or more occasions simultaneously. If Model 1 or 3 did not predict in the positive in 90% of iterations, the prediction outputs of Model 2 were used.

After model integration, we derived a set of soil thickness exceedance probability mapping outputs. These were derived simply by assessing the empirical probabilities (at each pixel) and then tallying the number of occasions the estimated soil depth exceeded given threshold depths of 10 cm, 50 cm, 100 cm, and 150 cm. This tallied number was divided by 50 to give an exceedance probability for each threshold depth.

3. Results

The quality of the integrated modelling of soil thickness across Australia is given in terms of each of the contributing models. The results are summarised in Table 2 and are based entirely on data that were excluded from model fitting. The median concordance coefficient for Model 2 (model for predicting soil thickness on a continuous scale from 0 to 200 cm) that did not consider further spatial modelling of the residuals was 0.768. Our analysis found there to be spatial structure in the residuals which was described with a Matern variogram (Stein's parameterisation) that had parameters: Nugget = 0.007, Sill = 0.01, distance = 245,605 m, and Kappa (smoothness parameter) = 0.2. Kriging with this variogram resulted in a map which is displayed as supplementary material. However, we decided not to add the RF model predictions with the map of kriged residuals because there was not any improvement in the goodness of fit statistics (the concordance coefficient reduced to 0.705 and the RMSE increased to 0.273 (square-root units) compared to 0.257 for just using the modelled predictions). The hex-plot (Fig. 4) shows the comparative modelled predictions and associated observations that binned for different depth increments. Note that the observations include the non-censored and censored data, where the observation for the censored data was simulated (from the beta distribution). Fig. 4 aggregates data from each model iteration that were excluded from the model fitting and the colours distinguish total counts in each 'bin'. Most of the grouping of the data are close to the 1:1 observed vs predicted line which is a desired outcome. Notwithstanding, there appears to be some overprediction evident in soils between 0 and 1 m and some underprediction in soils 1–2 m thick.

Model 1 which was the binary model to distinguish rock outcrops from soil cover with undefined thickness was found to have an overall accuracy of 99% (50th percentile). Model perturbation with iterative 5-fold random data splitting did not cause any fluctuations in goodness of fit for Model 1. This was the case also for Model 3, although with less overall accuracy of 85%. Kappa coefficients for both Models 1 and 3 confirmed their predictive skill with values of 0.879 and 0.641 respectively.

Table 2
Goodness of fit statistics of the 3 individual models based on data excluded from model fitting (validation data). Values represent the median of 50 model iterations and values in square brackets represent the 5th and 95th percentiles.

Model	RMSE	Concordance
Model 2 (Ranger model only)	0.257 [0.255–0.258]	0.768 [0.764–0.772]
Model 2 (Ranger model + kriged residuals)	0.273 [0.271–0.274]	0.705 [0.701–0.710]
	Overall accuracy (%)	Kappa coefficient
Model 1 (rock outcrops vs. soil)	99 [99–99]	0.879 [0.872–0.884]
Model 3 (soil > 2 m vs soil < 2 m	85 [85–86]	0.641 [0.636–0.648]

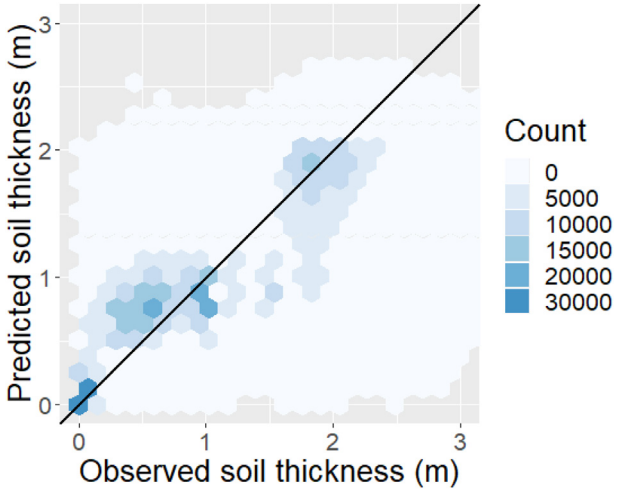

Fig. 4. Hex plot of observed soil thicknesses vs. associated modelled soil thickness. These data are based on data excluded from model fitting from each model iteration and include both non-censored and right-censored data. The colour scale represents the frequency of each observed-predicted pair. Black line corresponds to a 1:1 line.

Model integration and post-processing resulted in the derived spatial products shown in Fig. 5 as the 5th, 50th and 95th percentiles of the empirical distributions at each pixel. There is broad similarity in spatial pattern when comparing with the regolith depth map produced by Wilford et al. (2016). While not directly correlated, the pattern of deep regolith has some correspondence with deep soils. Compared with the Viscarra Rossel et al. (2014) soil thickness map, our study displays a more granular spatial heterogeneity and delineates deep soils (> 2m). For example, riverine and alluvial and lacustrine plain areas have relatively deep soils which corresponds well with legacy soil surveys (Chen, 1997).

For all three models, principal component variables of the climatic data theme featured strongest. In fact, for each model type, each of the 12 climatic data themed PCAs were in the top 20 most important model variables. Based on variable importance measures that help summarise random forest models (Louppe et al., 2013), we aggregated these across each model iteration to get an overall sense of which variables contributed most strongly to each model. For Model 1, the other variables which featured in the top 20 included a single Organism themed PCA variable that was in fact the single most important variable overall. The strong relationship between vegetation/landuse and the presence of rock outcrop or very skeletal soils is indicative of landscapes subject to minimal disturbances relative to those under intensive agricultural management (the latter for which can be relatively easy to distinguish with multispectral remote sensing satellite imagery). Relief and Parent Material themed PCA variables had 4 and 3 variables each in the top 20 most important for Model 1.

For Model 2 in addition to the 12 climatic data themed PCAs, Relief, Parent Material and Organism themes each had 4, 3, and 1 PCA variables in the top 20 most important. The Organism PCA of Model 2 was well down the list this time and the Relief PCA variables featured more prominently at the top of the list and included the geormorphons variable, which help describe particular terrain morphology in categorical terms. This outcome (Relief being relatively important predictor variable of soil thickness) coincides with earlier studies such as Patton

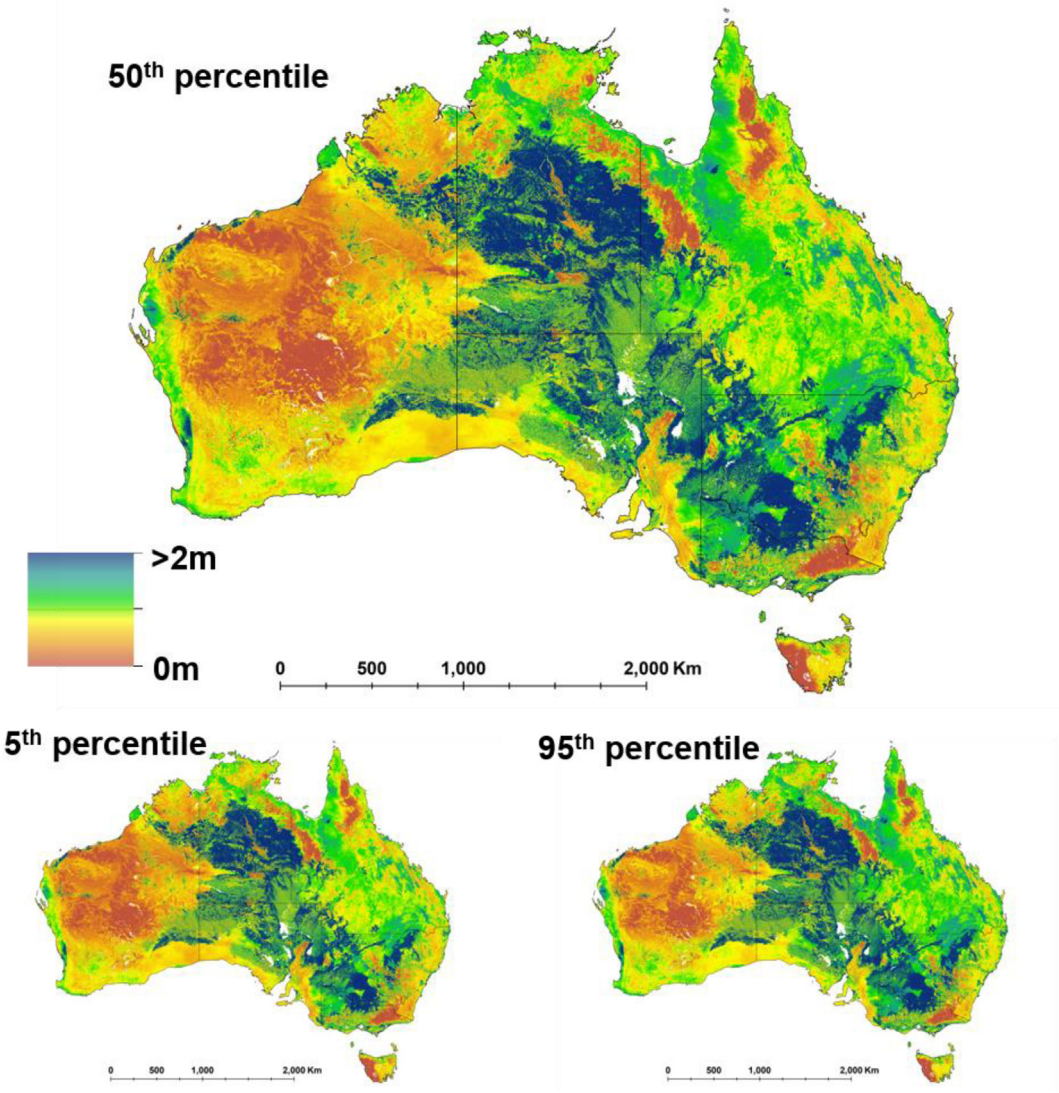

Fig. 5. Integrated maps of soil thickness showing predictions for the median (50th percentile), 5th and 95th prediction percentiles.

et al. (2018) and McKenzie et al. (2003) where geomorphologic factors are relied upon for skilled estimates of soil thickness across investigated catchments.

Model 3 had 4 Relief themed PCA variables in the 20 most important variables in addition to the geomorphons variable which was the single most important variable. This would indicate that across Australia the predominantly flatter landscapes are associated with deeper soils than those in more rugged and steep landscapes. Parent Material and organism themed PCA variables had 2 and 1 variables respectively in the 20 most important variables.

We note that climatic variables were included in all three Models. The climatic variables are statistical moments of long term (30 year) climatic data and meteorological derivatives of rainfall and temperature, incorporating topographic adjustment to account for the insolation effects of solar radiation exposure where relevant. The general spatial patterns of climatic data variability are intended to represent relatively deep time – stable climates over several thousand years since the last ice age. There is no assertion that such information are causative factors underpinning spatial heterogeneity of soil thickness across Australia. However, it is apparent that climatic regions of Australia would be a strong driver of soil formation processes interacting with landform, geology and biota. Climate regulates biota (e.g., the distribution of biomes) and the weathering of parental materials. The inclusion of both inter-annual and seasonal climate together with topo-climatic effects has resulted in relatively skilful predictive models of soil thickness. To generalise, the Australian land surface is very flat and geologically very old and has been subject to intensive and prolonged weathering (Young and Young, 2001). Notwithstanding the human-induced effects on present soil condition – although soil thickness maybe less affected by this in most context – Australian soils are likely to be less affected by soil forming factors such as modern geologic activities and geomorphic-induced transportation and translocation processes, at least at the continental extent of digital soil mapping.

The exceedance probability maps (Fig. 6) distinguish Australia's shallow and rocky soils and the relatively deep soils. Fig. 6a and b highlight large areas of shallow soils and rock surfaces in areas where one would expect to see them. For example, along Tasmania's west coast (Pemberton, 1989) and the alpine region in the south east of the island (Costin, 1954). In northwest Queensland, the relatively shallow soils appear in expected places such as the Isa Highlands which are coincident with the outcrop area of the folded and metamorphosed rocks and igneous rocks of the Cloncurry complex (Perry et al., 1964). In Queensland's Cape York region there are consistent correlations between shallow and rocky soils with soil landscapes described by Wilson and Philip (1999) and Biggs and Philip (1995). Over in Western Australia, the Pilbara and Goldfields regions show large areas of shallow skeletal soils which are consistent with descriptions in Tille (2006). In general, Fig. 6 shows large areas of soils deeper than 1 m

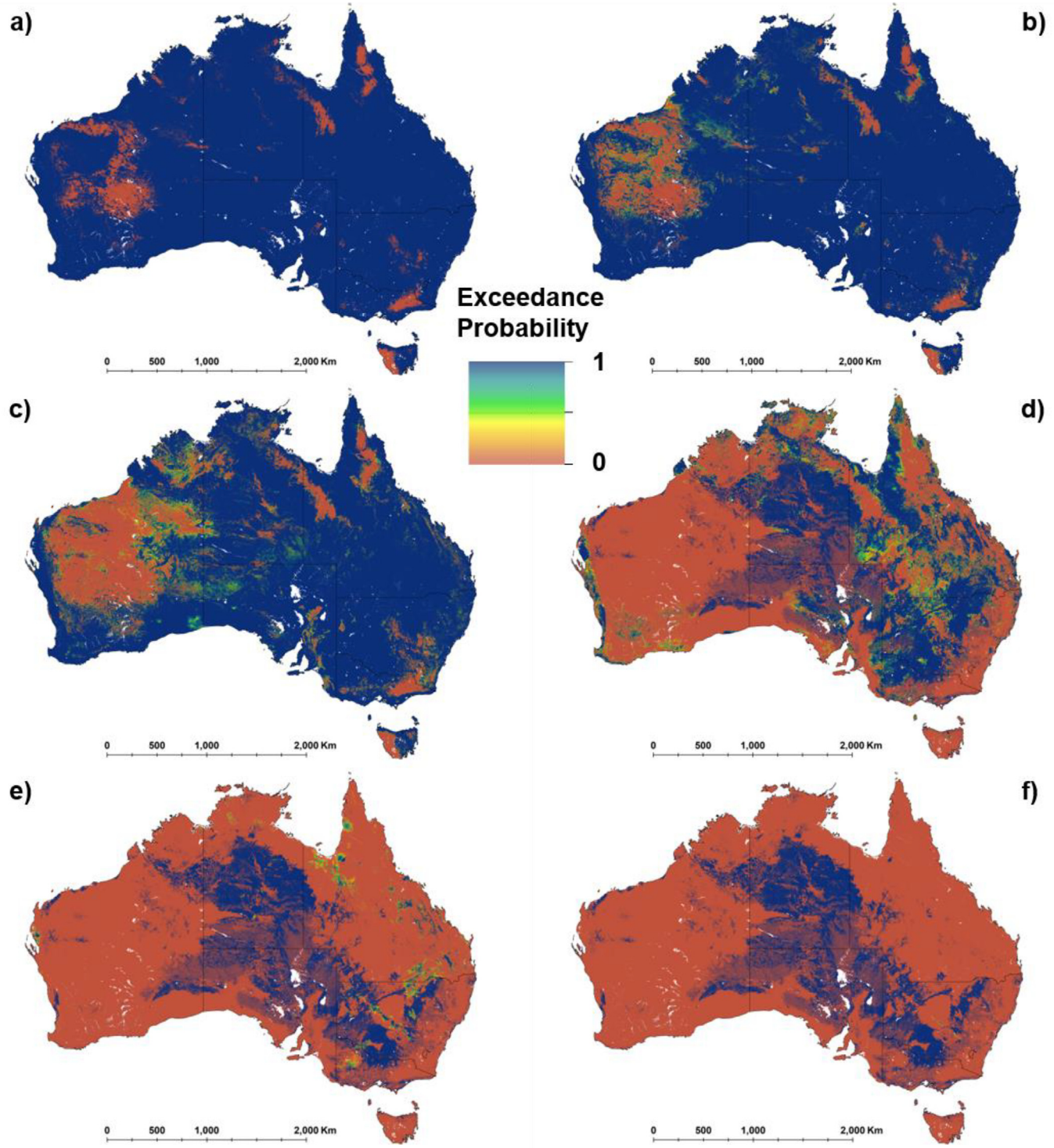

Fig. 6. Soil thickness exceedance probability maps for specified threshold depths. a) 10 cm, b) 30 cm, c) 50 cm, d) 100 cm, e) 150 cm, and f) 200 cm.

across much of Australia's interior, including agriculturally important areas in the weathered alluvial floodplains, and arid deep sands and dune systems.

4. General discussion

The soil thickness mapping products developed in this study are a significant improvement on prior attempts at the national extent. These maps are by no means error free, as our results have established, and will need to be updated and improved overtime as new techniques, data and modelling approaches evolve. It is very much in the purview of ongoing digital soil mapping activities at least at the national extent amongst other goals to set up an infrastructure to facilitate the updating and repeatability of the mapping products (Searle et al., 2019). Here, this is exemplified by making the workflow and importantly the underpinning code base available to the public in a version control repository (https://github.com/AusSoilsDSM/SLGA/tree/master/SLGA/Development/soilThickness). Note this repository does not contain the data that was used in the study but is nonetheless publicly available. Soon, work could involve a further integration with work done to date

on regolith mapping (Wilford et al., 2016). This could possibly allow for more of the drill hole data to be considered in the modelling, but perhaps more judiciously to account for issues with using drill hole data such that the data logging will be more useful in some areas than in others. Both regolith and soil thickness mapping could be improved where products would suit both hydrologists (for aquifer modelling work) and have the necessary granularity for agriculturists concerned with planning irrigation events.

We see immediate applications of this updated soil thickness mapping for whole soil estimates of soil carbon stocks to support improved inventories of land-based carbon flows for greenhouse gases assessments and reporting. Similarly, soil thickness models are also useful as an input with other variables for predicting terrestrial ecosystem dynamics and species/ecosystem distribution patterns (Williams et al., 2012). There is also a growing need for monitoring and forecasting of soil water storage and capacity in the agriculturally important areas of Australia. For example, for improved water usage efficiency in agricultural zones, and decision making around planting, and whether enough water is in the system to carry a crop through to harvest or to cut losses and hay off the available biomass. In general, a better handle on soil moisture status and dynamics will largely improve resilience across Australia particularly in drought affect times, and this needs to be underpinned by useful and indicative soil information such as soil thickness.

5. Conclusions

Acknowledging recent and historical approaches for mapping soil thickness around the world this study sought to develop an approach customised to Australian conditions and available data sources with which to derive inferences. This was achieved by separate modelling of rock outcrops, intermediate and deep soils that were then integrated into one output with associated quantified uncertainties. Each of the spatial models were calibrated using a suite of environmental covariates for which climatic themed variables consistently came up as the most important predictor variables. We deduce this because such variables have direct and indirect effects via regulating biota and the weathering of parental materials which ultimately drives spatial heterogeneity of soil thickness across Australian landscapes. This updated soil thickness mapping for Australia is an improvement on previous efforts and will provide better information to inform end-users in applications such as estimating soil and carbon stocks and soil water balance modelling and monitoring.

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.

Acknowledgements

The authors acknowledge the Terrestrial Ecosystem Research Network (TERN), an Australian Government NCRIS-enabled research infrastructure project, for facilitating and supporting this research. The authors also acknowledge CSIRO colleagues Kristen Williams, Anthony Ringrose-Voase and Sanji Pallegedara Dewage for their thoughtful and helpful insights on earlier versions of this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geoderma.2020.114579.

References

Biggs, A.J.W., Philip, S.R., 1995. Soils of Cape York Peninsula. Queensland Department of Primary Industries, Mereeba, Queensland.
Breiman, L., 2001. Random forests. Machine Learn. 45 (1), 5–32.
Chen, S., Mulder, V.L., Martin, M.P., Walter, C., Lacoste, M., Richer-de-Forges, A.C., Saby, N.P.A., Loiseau, T., Hu, B., Arrouays, D., 2019. Probability mapping of soil thickness by random survival forest at a national scale. Geoderma 344, 184–194.
Chen, X.Y., 1997. Quaternary sedimentation, parna, landforms, and soil landscapes of the Wagga Wagga 1: 100,000 map sheet, south-eastern Australia. Soil Res. 35 (3), 643–668.
Costin, A.B., 1954. A Study of the Ecosystems of the Monaro Region of New South Wales: With Special Reference to Soil Conservation. Government Printer, Sydney, Australia.
Gallant, J., Read, A., Dowling, T., 2012. Building the national one-second digital elevation model for Australia, Water. Information Research and Development Alliance: Science Symposium Proceedings.
Harwood, T., Ferrier, S., Harman, I., Ota, N., Perry, J., Williams, K., 2014. Gridded continental climate variables for Australia. CSIRO Land and Water, Canberra.
Hiemstra, P.H., Pebesma, E.J., Twenhöfel, C.J.W., Heuvelink, G.B.M., 2009. Real-time automatic interpolation of ambient gamma dose rates from the Dutch radioactivity monitoring network. Comput. Geosci. 35 (8), 1711–1721.
Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S., 2008. Random survival forests. Ann Appl. Stat. 2 (3), 841–860.
Kempen, B., Brus, D.J., de Vries, F., 2015. Operationalizing digital soil mapping for nationwide updating of the 1:50,000 soil map of the Netherlands. Geoderma 241–242, 313–329.
Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., Hunt, T., 2019. caret: Classification and Regression Training. R package version 6.0-84. CRAN, https://CRAN.R-project.org/package = caret.
Lacoste, M., Mulder, V.L., Richer-de-Forges, A.C., Martin, M.P., Arrouays, D., 2016. Evaluating large-extent spatial modeling approaches: a case study for soil depth for France. Geoderma Regional 7 (2), 137–152.
Louppe, G., Wehenkel, L., Sutera, A., Geurts, P., 2013. Understanding variable importances in forests of randomized trees, Proceedings of the 26th International Conference on Neural Information Processing Systems – Volume 1. Curran Associates Inc., Lake Tahoe, Nevada, pp. 431–439.
Ma, Y., Minasny, B., Welivitiya, W.D.D.P., Malone, B.P., Willgoose, G.R., McBratney, A.B., 2019. The feasibility of predicting the spatial pattern of soil particle-size distribution using a pedogenesis model. Geoderma 341, 195–205.
McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117 (1), 3–52.
McKenzie, N., Gallant, J., Gregory, L., 2003. Estimating Water Storage Capacities in Soil at Catchment Scales. Cooperative Research Centre for Catchment Hydrology.
Minasny, B., McBratney, A.B., 1999. A rudimentary mechanistic model for soil production and landscape development. Geoderma 90 (1), 3–21.
Moore, I.D., Gessler, P.E., Nielsen, G.A., Peterson, G.A., 1993. Soil attribute prediction using terrain analysis. Soil Sci. Soc. Am. J. 57 (2), 443–452.
Odeh, I.O.A., Chittleborough, D.J., McBratney, A.B., 1991. Elucidation of soil-landform interrelationships by canonical ordination analysis. Geoderma 49 (1), 1–32.
Patton, N.R., Lohse, K.A., Godsey, S.E., Crosby, B.T., Seyfried, M.S., 2018. Predicting soil thickness on soil mantled hillslopes. Nat. Commun. 9 (1), 3329.
Pelletier, J.D., Rasmussen, C., 2009. Geomorphically based predictive mapping of soil thickness in upland watersheds. Water Resour. Res. 45 (9).
Pemberton, M., 1989. Land Systems of Tasmania Region 7: South west. Department of Agriculture, Tasmania, Hobart, Tasmania.
Perry, J.A., Sleeman, J.R., Twidale, C.R., Prichard, C.E., Slatyer, R.O., Lazarides, M., Collins, F.H., 1964. General Report on lands of the Leichhardt-Gilbert Area, Queensland, 1953-54. Commonwealth Scientific and industrial Research Organisation, Melbourne. Australia.
Searle, R., 2015. The Australian site data collation to support the GlobalSoilMap. In: Arrouays, D., McBratney, A.B., Hempel, J., Richer-de-Forges, A.C., McBratney, A.B. (Eds.), GlobalSoilMap: Basis of the Global Spatial Soil Information System. CRC Press, London, UK, pp. 127–133.
Searle, R., Grundy, M.J., McBratney, A.B., Gregory, L., Wilson, P., Malone, B.P., Stenson, M., 2019. Phased development of digital soil infrastructure for Australia. University of Chile, Santiago, Joint Workshop for Digital Soil Mapping and GlobalSoilMap.
Shangguan, W., Hengl, T., Mendes de Jesus, J., Yuan, H., Dai, Y., 2017. Mapping the global depth to bedrock for land surface modeling. J. Adv. Model. Earth Syst. 9 (1), 65–88.
Styc, Q., Lagacherie, P., 2016. Predicting soil depth using a survival analysis model, 7th Global Workshop on Digital Soil Mapping, Aarhus, Denmark.
National Committee on Soil and Terrain, 2009. Australian Soil and Land Survey Field Handbook, third ed. CSIRO Publishing, Melbourne, Australia.
Tille, P.J., 2006. Soil-Landscapes of Western Australia's Rangelands and Arid Interior. West Australian Department of Agriculture and Food, Perth, WA.
Viscarra Rossel, R., Chen, C., Grundy, M., Searle, R., Clifford, D., Odgers, N., Holmes, K., Griffin, T., Liddicoat, C., Kidd, D., 2014. Soil and landscape grid national soil attribute maps – soil depth (3" resolution) – Release CSIRO. Data Collect. 1, v3. https://doi.org/10.4225/08/546F540FE10AA.
Viscarra Rossel, R.A., 2011. Fine-resolution multiscale mapping of clay minerals in Australian soils measured with near infrared spectra. J. Geophys. Res. Earth Surf. 116 (F4).
Viscarra Rossel, R.A., Chen, C., Grundy, M.J., Searle, R., Clifford, D., Campbell, P.H., 2015. The Australian three-dimensional soil grid: Australia's contribution to the

GlobalSoilMap project. Soil Res. 53 (8), 845–864.
Wilford, J., 2012. A weathering intensity index for the Australian continent using airborne gamma-ray spectrometry and digital terrain analysis. Geoderma 183–184, 124–142.
Wilford, J.R., Searle, R., Thomas, M., Pagendam, D., Grundy, M.J., 2016. A regolith depth map of the Australian continent. Geoderma 266, 1–13.
Williams, K.J., Belbin, L., Austin, M.P., Stein, J.L., Ferrier, S., 2012. Which environmental variables should I use in my biodiversity model? Int. J. Geograph. Inf. Sci. 26 (11), 2009–2047.
Wilson, J.P., Gallant, J.C., 2000. Secondary topographic attributes. In: Wilson, J.P., Gallant, J.C. (Eds.), Terrain Analysis: Principles and Applications. John Wiley & Sons, New York.
Wilson, P., Philip, S.R., 1999. An Assessment of Agricultural Potential of Soils in the Gulf Region, North Queensland. Queensland Department of Natural Resources, Merreba, Queensland.
Wright, M.N., Ziegler, A., 2017. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. 2017 77(1), 17.
Xu, T., Hutchinson, M.F., 2011. ANUCLIM Version 6.1 User Guide. Fenner School of Environment and Society, The Australian National University.
Young, A., Young, R., 2001. Soils in the Australian Landscape. Oxford University Press, South Melbourne, Victoria.