Influence of local data and local calibration on SoilGrids predictions

Merel Jager

May 2020
Influence of local data and local modelling on SoilGrids predictions

Merel Jager

Registration number 950710387120

Supervisors:
S. de Bruin (Laboratory of Geo-Information Science and Remote Sensing)
G.B.M. Heuvelink (Soil Geography and Landscape / ISRIC – World Soil Information)
B. Kempen (ISRIC – World Soil Information)

A thesis submitted in partial fulfilment of the degree of Master of Science
at Wageningen University and Research Centre,
The Netherlands.

Date: 05/2020
Wageningen, The Netherlands

Thesis code number: GRS-80436
Thesis Report: GIRS-2020-21
Wageningen University and Research Centre
Laboratory of Geo-Information Science and Remote Sensing
Abstract
Soils are the major component of the terrestrial ecosystem and the largest organic carbon pool on earth and therefore soils have to be managed properly. To facilitate proper management, digital soil maps of good quality are needed. This study aims to help ISRIC understand the impact of local data and local modelling to improve their digital soil maps.

This study aims to compare globally and locally calibrated SoilGrids (SG) models and assess the influence of adding local data to a globally calibrated SoilGrids model. This was assessed in three steps, using Soil Organic Carbon (SOC) and soil pH as soil properties to predict. The first step was calibrating an SG model using soil observations from a study area in India (Andhra Pradesh) and selecting local covariates. Predictions with this local SG model were compared to those of an SG model calibrated with soil observations from the entire globe and globally selected covariates, using a test dataset containing soil observations within Andhra Pradesh. The top ten most informative predictors differed 75% to 80% between the models. The predictions of local SG model were much closer to the observed values than those obtained by the global model with an RMSE of 0.286 for the local SOC model, 1.720 for the global SOC model, 0.467 for the local pH model, and 0.837 for the global pH model.

The second step compared the global SG model with an SG model calibrated using all global soil observations and all local observations. The covariates used for both models were identical, and predictions were made for three areas, i.e., Andhra Pradesh - India, Nampula – Mozambique (similar to Andhra Pradesh in feature space) and the Netherlands (very dissimilar to Andhra Pradesh in feature space). It was anticipated that for Nampula differences between the two models would be relatively large whereas for the Netherlands they would be small. Results showed that this was not the case. The areas dissimilar to Andhra Pradesh in feature space showed more deviation in the difference maps of the SG global and SG local predictions than the areas close in feature space. A reason for this deviation can be the difference in variable importance between both models, where predictions for the Netherlands were more influenced by in variable importance of the SG model calibrated using all global soil observations and all local observations, than Nampula was. To find out what the exact influence of covariate importance is, future research is needed. The out-of-bag (OOB) (see appendix I, OOB) model statistics of both models, showed that adding local soil observations to an SG model resulted in a better RMSE, ME and R² for predictions around the entire globe. To find out if the model predictions accuracy also improved for Nampula and the Netherlands, future research is needed.

The third and final step of this research was to gain insight into the influence of the size of the dataset added to a globally calibrated SG model. This was done by adding the local data in seven successive steps to the global soil observations and subsequent calibration of the global SG model, each time using the same covariates. The models were used to make soil pH predictions for Andhra Pradesh and showed that adding more local soil observations to the model increased prediction accuracy to a power-law (the more training data, the more accurate the model) and then reached a plateau from whereon the accuracy of the predictions is just slightly changing, as expected.

This study shows that adding local data observations to the SoilGrids DSM changes the variable importance of the model. This resulted in different model results in areas close in feature space, but also areas distant in feature space. The OOB model statistics showed that adding local data to a global model positively influenced the model accuracy for predictions all over the globe. However, to make even more accurate predictions with a SoilGrids model, a combination between locally selected covariates and global soil observations can be used to make predictions for a local area.
Acknowledgements

I would like to express my sincere gratitude to my supervisors Gerard Heuvelink and Sytze de Bruin for providing guidance throughout this research and for having the patience and time to read this document multiple times to provide feedback. In addition, a thank you to Bas Kempen who provided extra knowledge and information when needed and for helping me through the (dumb) mistakes I sometimes made when programming in R.

I also would like to thank all the people and supervisors of thesis ring 2 for their feedback and ideas regarding this thesis.

Last but not least, I would like to thank my family and friends for their constant source of inspiration and encouragements to keep going and finish this thesis in time.
# Table of contents

1. Introduction .................................................................................................................. 6
   1.1 Research questions ................................................................................................. 7

2. Methodology .................................................................................................................. 8
   2.1 The SoilGrids Framework ....................................................................................... 8
   2.1.1 SoilGrids rebuild .............................................................................................. 10
   2.2 Comparison of a global and local SoilGrids model ............................................... 10
   2.3 Influence of new soil observations ....................................................................... 13
   2.4 Local data density .................................................................................................. 14
   2.5 Case study ............................................................................................................... 14
   2.6 Data and materials ................................................................................................. 17

3. Results ............................................................................................................................ 21
   3.1 Comparison of a global and local SoilGrids model ............................................... 21
   3.1.1 Soil Organic Carbon content ............................................................................ 21
   3.1.2 Soil pH .............................................................................................................. 3
   3.2 Influence of new soil observations ....................................................................... 7
   3.2.1 model comparison ............................................................................................ 7
   3.2.2 Andhra Pradesh ............................................................................................... 9
   3.2.3 Nampula .......................................................................................................... 11
   3.2.4 Netherlands ...................................................................................................... 13
   3.3 Local data density .................................................................................................. 15

4. Discussion ....................................................................................................................... 17
   4.1 Comparison of a global and local SoilGrids model ............................................... 17
   4.2 Influence of new soil observations ....................................................................... 18
   4.3 Local data density .................................................................................................. 20
   4.4 limitations of this study ......................................................................................... 20

5. Conclusions and recommendations ............................................................................ 22
   5.1 Comparison of a global and local SoilGrids model ............................................... 22
   5.2 Influence of new soil observations ....................................................................... 22
   5.3 Effect of local data density .................................................................................... 22
   5.4 Recommendations ................................................................................................. 22

6. References ..................................................................................................................... 24

Appendix ............................................................................................................................. 28
   I Theoretical background ............................................................................................ 28
   II Covariates explanation ........................................................................................... 31
   III Homosoil results ................................................................................................. 36
Glossary:

Calibrating a model  
Model calibration is the process of adjustment of the model parameters. This includes that the model learns patterns, structures and parameters directly from the training data, and all settings that cannot be learned directly from the data are set manually. After a model is calibrated, it can be used to make predictions (Molnar, 2019). A synonym for calibration can also be: training a model.

Covariates  
A covariate is a variable that is correlated with the soil property of interest. According to this definition, any variable that is measurable, considered to have a statistical relationship with the dependent variable and that is available at all observation and prediction points would qualify as a potential covariate. (Fan, 2019).

Depth interval layers  
The six layers used by Soil Grids to represent soil properties at different depth intervals (0-5 cm, 5-15cm, 15-30cm, 30-60 cm, 60-100cm and 100-200 cm depth interval layers) (Hengl et al., 2017).

Global model  
A model is a global model when: the input soil profiles contain data of the entire globe, the covariate stack covers the entire globe and the model is calibrated using both global soil profiles and covariates.

Horizons  
A soil horizon is a layer in the earth with unique soil properties relative to parent and child layers (Zhang & Hartemink, 2019).

Local model  
A model is a local model when: the input soil profiles contain data of a country or region the covariate stack covers that same country or region and the model is calibrated using both country or region soil profiles and covariate stack.

Machine Learning  
(supervised) Machine learning are algorithms which can be trained to recognise patterns based on past event or experience with respect to some class of tasks. ML Algorithms are able to predict future events for any new input after sufficient training (Frankenfield, 2018).

Prediction depth  
This depth is the centre point of the depth interval layer you want to predict for (e.g. a prediction depth of 10 cm is used to predict for the 5-15 cm depth layer).

Sample depth  
The actual depth at which a soil observation sample was taken. This is the depth used as a covariate during model calibration. It stands apart from the depth interval layers and it should not be confused with prediction depth.

Soil observations:  
Soil observations are soil samples taken in the field and analysed in a lab to give a measure of the soil p. For this research, soil observations for Soil pH and Soil Organic Carbon are used.

Soil properties  
All soils contain mineral particles, organic matter, water and air. The combinations of these determine the soil’s properties – its texture, structure, porosity, chemistry and colour. Soil properties can therefore be e.g. soil organic carbon, soil pH, proportion of clay particles, total nitrogen or the proportion of sand particles.
List of figures

Figure 1. SoilGrids statistical framework (Hengl et al., 2017). ................................................................. 8
Figure 2. Distribution of major soil types in Andhra Pradesh – India (Maschinen et al., 2012). .......................... 15
Figure 3. Distribution of major soil types in the Netherlands (Maschinen et al., 2012)............................... 16
Figure 4. Distribution of major soil types in Nampula (Maschinen et al., 2012). ........................................ 17
Figure 5. Locations of soil observations provided with the 'WoSIS September 2019 snapshot' (Batjes, Ribeiro, & Van Oostrum, 2020). ........................................................................................................ 18
Figure 6. Variable importance of the SGGlobal and SGLocal SOC models ranked by SGGlobal variable importance. .................................................................................................................................................. 21
Figure 7. Soil pH prediction maps and difference map (top) of SGLocal (bottom left) and SGGlobal (bottom right) for Andhra Pradesh - India ........................................................................................................ 1
Figure 8. Scatterplots of the global (left) and local (right) SOC model. ............................................................ 2
Figure 9. Variable importance of the global and local soil pH models. .......................................................... 3
Figure 10. Soil pH prediction maps and difference map (top) of SGLocal (left) and SGGlobal (right) for Andhra Pradesh. ................................................................................................................................. 5
Figure 11. Scatterplots of the global (left) and local (right) soil pH model. ..................................................... 6
Figure 12. Variable importance of the SGGlobal and SGPlus SOC models based on the variable importance ranking of SGGlobal. ........................................................................................................ 7
Figure 13. Variable importance of the SGGlobal and SGPlus soil pH models. .................................................. 8
Figure 14. Difference maps of SOC (%) for three interval depths. .................................................................... 9
Figure 15. Difference maps of soil pH for three interval depths, Andhra Pradesh - India. .............................. 10
Figure 16. Difference maps of SOC (%) for three interval depths, Nampula – Mozambique. ......................... 11
Figure 17. Difference maps of soil pH for three interval depths, Nampula – Mozambique. ......................... 12
Figure 18. Difference maps of soil SOC (%) for three interval depths in the Netherlands. .............................. 13
Figure 19. Difference maps between SGPlus and SGGlobal of soil pH for three interval depths – the Netherlands. ............................................................................................................................................... 14
Figure 20. Learning curve of the predicted mean of the SGPlus model with respect to the mean observed soil observations, for soil pH in Andhra Pradesh – India, with the size of the dataset used for calibrating the model. .................................................................................................................. 15
Figure 21. Difference maps with ascending new data samples used for calibrating the model for soil pH in Andhra Pradesh – India. ........................................................................................................ 16
Figure 22. Prediction value Occurrence for SGLocal and SGGlobal for SOC .................................................. 18
Figure 23. Estimated SOC (%) for 0 - 30 cm depth layer in the Netherlands according to two different sources. Left: Dutch soil map, land use map and LSK data (Lesschen et al., 2012) and right HWSD based on soil types (figure 3) from Hiederer & Köchy, 2012.. The legend represents the SOC % for both maps. ........................................................................................................................................ 19
Figure 24. Estimated SOC (%) for 0 - 30 depth interval layer in the Netherlands according to the SGPlus model (left) and the SGGlobal model (right). ........................................................................................................ 19
Figure 25. Random Forest decision tree example to predict numbers and their colours (Yiu, 2019) .... 28
List of tables

- Table 1. SoilGrids model settings to compare a global and a local model. .................................................. 11
- Table 2. SoilGrids model settings to determine the influence of new soil observations. ......................... 13
- Table 3. Number of added local sample data used for assessing the effect of local data density. .......... 14
- Table 4. SoilGrids model settings to determine the influence of local data density. ................................. 14
- Table 5. Head of the global soil observations data. ......................................................................................... 18
- Table 6. Head of the local soil observations data. ......................................................................................... 19
- Table 7. Used R libraries to create the results for the research................................................................. 20
- Table 8. Descriptive statistics of the SOC models vs Observed values in Andhra Pradesh in % SOC..... 22
- Table 9. Descriptive statistics of the global and local SOC models with SOC in %................................. 2
- Table 10. Descriptive statistics of the Soil pH models vs Observed values in Andhra Pradesh. ............ 4
- Table 11. Descriptive statistics of the local and Global soil pH models...................................................... 6
- Table 12. OOB statistics for SG\textsubscript{Global} and SG\textsubscript{Plus} for SOC (%)........................................... 7
- Table 13. OOB statistics for SG\textsubscript{Global} and SG\textsubscript{Plus} for soil pH...................................................... 8
- Table 14. Statistics table of SOC (%) in Andhra Pradesh – India................................................................. 9
- Table 15. Statistics table of soil pH for Andhra Pradesh – India................................................................. 10
- Table 16. Statistical table of SOC (%) in Nampula – Mozambique............................................................. 11
- Table 17. Statistical table of soil pH in Nampula – Mozambique.............................................................. 12
- Table 18. Statistical table of SOC (%) in the Netherlands.......................................................................... 13
- Table 19. Statistical table of soil pH in the Netherlands.............................................................................. 14
- Table 20. Descriptive statistics of SG\textsubscript{Plus}\textsubscript{n} models, compared to the SG\textsubscript{Global} predictions and observed values all predicted for Andhra Pradesh for the soil property soil pH. ................................................. 15
1. Introduction

Many global problems such as climate change, freshwater scarcity, loss of agricultural land through erosion, biodiversity decline and feeding 10 billion people by 2050 present immense challenges to humanity (Bouma, Montanarella, & Evanylo, 2019). Soil has a vital part to play in all these challenges (Sanchez et al., 2009). Soil degradation has, for example, detrimental consequences for the already limited land and water resources available for agricultural productivity (Koch et al., 2013). To be able to protect our carbon-rich and biodiverse rainforests, wetlands, and grasslands from changing into croplands and still be able to feed 10 billion people in the near future, we will have to increase current crop yields on existing farmland through sustainable intensification. In many areas, it is possible to sustainably intensify crop production through proper soil management and the use of fertilisers (Kopittke, Menzies, Wang, McKenna, & Lombi, 2019). Soil information is important to determine the amount of fertilisers needed to increase in yield. Another reason why soil is essential is that soil constitutes the earth’s largest terrestrial carbon (C) pool (Jobbágy & Jackson, 2000). Our challenge is to keep C in the soil through proper soil management and preferably increase soil C as a climate change mitigation measure. When soil is not rightly maintained (e.g. deforestation), it will release C into the atmosphere resulting in an aggravating climate change measure (Davidson & Janssens, 2006). Because the soil has a vital part to play in addressing global problems, the demand for soil information is increasing.

The growing demand for soil information in combination with the increase in available detailed satellite data has led to the development of Digital Soil Mapping (DSM) (McBratney, Mendonça Santos, & Minasny, 2003). DSM produces soil maps using soil mapping based on supervised machine learning (ML) algorithms or geostatistical methods like kriging to predict soil properties from soil point observations and environmental covariate layers that are typically derived from satellite imagery and digital elevation models. These covariate layers related to the five major soil-forming factors that were identified by Pendleton & Jenny, 1945 and formalised into a model for soil development. These factors are climate (cl), organisms (o), relief (r), parent material (p) and time (t), or ‘CLORPT’ and are widely used as input for DSM.

In many parts of the world, DSM has shifted from an academic pursuit to operational initiatives on both local (regional or national) (Hengl et al., 2015; Kempen, Brus, & de Vries, 2015) and global scale (Arrouays, Lagacherie, & Hartemink, 2017). SoilGrids is a DSM framework developed by ISRIC that provides global (world covering) scale prediction maps for a standard set of numeric soil properties (Bulk density of the fine earth fraction, Cation exchange capacity of the soil, Volumetric fraction of coarse fragments, Proportion of clay particles, Total nitrogen, Soil pH, Proportion of sand particles, Proportion of silt particles, Soil organic carbon content, Organic carbon density and Organic carbon stocks) at six standard depths intervals (0-5 cm, 5-15 cm 15-30 cm 30-60 cm 60-100 cm and 100-200 cm). Those predictions are based on a collection of worldwide soil point observations and remote sensing-based soil covariates (Hengl et al., 2017)(see 2.1, the SoilGrids framework). These maps are used in various initiatives, for instance, to fill in the gaps for the Global Soil Organic Carbon Map (FAO & ITPS, 2018), as a data source for assessing land degradation trends (UNCCD & The Global Mechanism, 2016) and as data input for ecological niche modelling of plant species (Velazco, Galvão, Villalobos, & De Marco, 2017).

Because of its global nature, SoilGrids might not represent local patterns of the spatial distribution of soil well or give accurate local predictions, for instance, for a specific country or region. This could be caused by the lack of soil point observations for that specific county/region or predictive relationships that are locally different from globally. To address this, it might be more accurate to calibrate a DSM.
model locally. Kempen et al., 2019 uses a locally fitted ML model to predict Soil Organic Carbon content (SOC) for Tanzania instead of a globally fitted ML model. This model very likely captures local predictive relationships much better than a globally calibrated model would, and will thus result in much more accurate predictions (Hand & Vinciotti, 2003). A local model selects covariates that are locally relevant. This selection might differ from a covariate set that is selected by a global model. The latter might not be optimal for a more local application. In this way, the model will be better tuned to local conditions. A negative side of this is that an extensive calibration dataset is needed to ensure there is enough calibration data available to calibrate a DSM model locally.

To provide more accurate local SoilGrids maps, one could consider serving the SoilGrids framework local data only (local soil samples and local covariates) and thereby make it a local model that has specifically tailored predictive relationships. Calibrating a local model requires more local data than calibrating a global model and may result in a patchwork of local models with sharp boundaries between them. If local data are available, then it can also be interesting to add these to the global model and analyse if this leads to locally improved (more accurate) predictions (Vitharana, Mishra, & Mapa, 2019). Therefore, this study aims to compare globally and locally calibrated SoilGrids models and assess the influence of adding local data to a globally calibrated SoilGrids model. To achieve this objective, three research questions will be answered.

1.1 Research questions

1. What is the model prediction performance of a globally calibrated SoilGrids model compared to a SoilGrids model calibrated on local soil data only?

2. What is the model prediction performance of the current SoilGrids model compared to a global SoilGrids model calibrated after adding local data?

3. How does the effect of adding local data on prediction accuracy depend on the size of the local dataset?

The methodology will be tested using Andhra Pradesh - India as case study area (chapter 2.5). The local SoilGrids model will be built according to the SoilGrids Framework, thereby using only soil observations from Andhra Pradesh and a local covariate feature selection. The soil properties chosen for this study are soil organic carbon content in % and soil pH in H2O.
2. Methodology

The research questions are answered using the SoilGrids framework as a basis. Chapter 2.1 describes the SoilGrids framework, which is used to answer the three research questions. Chapter 2.2, 2.3 and 2.4 each describes the method of a research question: what the model inputs and settings are, where predictions will be made and how the accuracy/influence of data of those predictions is determined. Chapter 2.5 explains the case study and chapter 2.6 finishes the methodology with the used materials and data.

2.1 The SoilGrids Framework

DSM often makes use of supervised machine learning (ML) algorithms to predict soil properties. Figure 1 shows the SoilGrids framework, which uses Random Forest (see appendix I, Theoretical background) as an ML algorithm. Each step in the framework is briefly explained below. See (Hengl et al., 2017) for a detailed step by step explanation of the SoilGrids DSM framework.

![SoilGrids statistical framework](image)

**Figure 1. SoilGrids statistical framework (Hengl et al., 2017).**

A. Soil observations

The soil samples and observations dataset (part of step A, future explained in section 2.6 Data and materials) contains a compilation of soil profiles and sample data used for model calibration. The soil samples dataset is derived from soil observations made all over the world and samples analysed in the lab by multiple organisations. ISRIC collects geo-referenced soil profile data from the world, harmonises the data, merges datasets and serves the result via the World Soil Information Service (Batjes et al., 2017). SoilGrids soil profiles contain measured values of soil properties (e.g. pH in H2O, sand, silt and clay), sample **depth** and coordinates of the measured location. Besides the soil observations, the soil
sample dataset may also include expert-based pseudo-observations. Some large areas that have extreme climatic conditions and/or have very restricted access are significantly under-sampled. To ensure that the model can represent those under-sampled areas, pseudo-observations are inserted and fill the gaps in the feature space so the dataset can be used for model training (Hengl et al., 2017).

B. Covariates
A covariate is a variable that is correlated with the soil property of interest. According to this definition, any variable that is measurable, considered to have a statistical relationship with the dependent variable and that is available at all observation and prediction points would qualify as a potential covariate. A covariate is thus a possible predictive or explanatory variable of the dependent variable (Fan, 2019). Hengl et al. (2017) list all the covariates used by SoilGrids. These include e.g. land cover classes (cultivated land, forests, grasslands, shrublands, wetlands, tundra, artificial surfaces and bare land cover), long-term averaged mean monthly hours under snow cover, global water table depth in meters, average soil and sedimentary-deposit thickness in meters. The full list of used covariates is shown in appendix II.

Covariates were generated using different remote sensing data repositories (e.g. MODIS land products), step B. Those data are stacked, so each location of the world has all covariate values (see Section 2.6). This covariate stack is used to extract the regression matrix (see below) used for training data and used as input for the model predictions.

C. Regression matrix
To be able to calibrate the model, each soil observation needs to be associated with information about the covariates. A regression matrix was built by extracting all covariates collocated with the soil profiles using a spatial overlay operation, step C. In this regression matrix, the sample depth of the soil observations is used as ‘depth’ covariate for model calibration.

D. Model calibration
Before the model is calibrated, first a correlation analyses and a random feature selection (RFE)(see appendix I, correlation analyses and RFE) are performed to gain the most optimal covariates to use in the Random Forest (RF, see appendix I; Random Forest) formula. RFE determines the covariate importance and the outcome of this function is a list with selected covariates which are most important for the RF predictions, without reducing the model performance. Next, the model parameters can be defined.

A machine learning algorithm learns patterns from existing data. RF learns parameters and structures directly from the training data and creates an ensemble of different tree models (Molnar, 2019). Hyperparameters are all model settings which cannot be learned directly from the training data. For SoilGrids the RF model hyperparameters which are set are mtry and num.trees. The hyperparameter mtry sets the number of covariables available for splitting at each tree node. The num.trees hyperparameter set the number of trees to grow. Larger number of trees produce more stable models and covariate importance estimates, but require more memory and longer run times (Liaw & Wiener, 2018). To estimate those hyperparameters, a random subset of the regression matrix (5 – 10% of the total size) is used to calibrate and validate a list of models where a predefined combination of
The optimal hyperparameters of the RF model were defined by the model with the lowest RMSE.

The regression matrix was used to calibrate the random forest model, step D. Per soil property, a formula is defined, where: soil property is a function of all optimal covariates. The formula, regression matrix and hyperparameters together are used to calibrate the SoilGrids model. For each soil property, a separate model is calibrated.

E. Prediction & Validation
The fully calibrated RF model can be used to make predictions for any location and any depth where the covariates are known, Step E. The covariate raster stack built in step B is used as input data. ISRIC generates SoilGrids predictions at six standard depth intervals; 0-5 cm, 5-15 cm, 15-30 cm, 30-60 cm, 60-100 cm and 100-200 cm. To make predictions for a depth interval layer, the prediction depth is added as a covariate to the covariate stack (sample depth is not part of the covariate stack). The prediction depth is the centre point of the depth interval layer to predict for and is calculated using the following formula

\[
\frac{\text{top of a layer} - \text{bottom of a layer}}{2} + \text{bottom of a layer}
\]

To determine the model performance (validation), the calibrated model was used to make predictions using 10-fold cross-validation (see appendix I, k-fold cross-validation. Each model was re-calibrated ten times using 90% of the data and predictions derived from the calibrated models are compared with observations of the remaining 10% to gain the model accuracy (Molnar, 2019). Assessments followed the same procedures as described in section 2.2.

2.1.1 SoilGrids rebuild
The original SoilGrids Framework was built in several different programs optimised for high-performance computing (HPC) and generation of tiled predictions (De Sousa, Poggio, Dawes, Kempen, & van den Bosch, 2020). To make this research executable in a non-HPC environment, it was decided to rebuild the SoilGrids framework completely in R. The SoilGrids framework was rebuilt as close as possible to the existing framework to minimise the influences on the model results. To be able to compare globally and locally calibrated SoilGrids models and assess the influence of adding local data to a globally calibrated SoilGrids model, there should not be any differences between the SoilGrids framework and the rebuilt DSM framework. Therefore, all model results were predicted using this remake of the SoilGrids framework.

2.2 Comparison of a global and local SoilGrids model
The global SoilGrids model, here called SGGlobal, and the local SoilGrids model, here called SGLocal, were created using the SoilGrids framework for each soil property of interest. For the SGGlobal model, global soil samples and global covariates were used to create the regression matrix, as shown in Table 1. The global soil sample dataset did not include the local soil samples used in the SGLocal model. For the covariates, I used the 129 covariates remaining after the correlation analysis (performed by ISRIC) as input for the RFE performed for each soil property model. The results of the RFE were used as covariate stack and as input for the model formula. The next step was to calibrating the model. The hyperparameters of the model were copied from the original SoilGrids, as the function to calculate those hyperparameters was tested. The optimal hyperparameters of the RF model were defined by the model with the lowest RMSE.
hyperparameters should be performed on an HPC and takes much of time (around ten days for all SoilGrids properties). The models should be as close as possible to the original SoilGrids and we assumed that for local calibration the hyperparameter results would not differ much from the global calibration results. Therefore, all models built during this research used the hyperparameters of the original SoilGrids. For SOC, mtry was set to 14 and num.trees to 250 for OC while for pH, mtry was set to 12 and num.trees to 150. With the hyperparameters set and the regression matrix complete, the model was calibrated and used for predicting. The model was used to make predictions for the local study area at a depth interval of interest and for the soil properties of interest.

The SG_local model was created using local soil samples and the global covariates clipped to the local extent. The procedure of creating the SG_local model was almost identical to SG_global. The only difference was; in the regression matrix, the standard depth of the local soil samples and a random number between 1-10 (later used as fold) was added.

Table 1. SoilGrids model settings to compare a global and local model.

<table>
<thead>
<tr>
<th>Model settings</th>
<th>SG_local</th>
<th>SG_global</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soil Observations</td>
<td>Local</td>
<td>Global</td>
</tr>
<tr>
<td>Used covariates</td>
<td>SG_local RFE results</td>
<td>SG_global RFE results</td>
</tr>
<tr>
<td>Prediction area</td>
<td>local study area</td>
<td>local study area</td>
</tr>
<tr>
<td>Soil Property</td>
<td>OC and soil pH</td>
<td>OC and soil pH</td>
</tr>
<tr>
<td>Prediction depth interval</td>
<td>5-15 cm</td>
<td>5-15 cm</td>
</tr>
</tbody>
</table>

Model comparison

The SG_global and SG_local model were assessed by comparing the used covariates and their relative importance. The importance of the variable is calculated using the Gini Importance. The Gini importance calculates each covariate importance as the sum of the number of splits (across all trees) that include the covariate, proportionally to the number of samples (soil observations) it splits (Strobl, Malley, & Tutz, 2009). The result can be shown as a bar plot with the covariates on the y-axis and the relative importance of the covariates on the x-axis. To be able to compare the results, the x-axes of SG_global was rescaled dividing the SG_global variable importance by the result of equation 3.

\[
\frac{\text{SG} \_\text{Global soil observations}}{\text{SG} \_\text{Local soil observations}}
\]

(3)

Variable importance plots were used to assess whether global and local variable importance are different, indicating local patterns. The ranger package also reports model calibration fit via the R^2 based on out-of-bag (OOB) samples. Those statistics were also compared per model and per soil property.
Map comparison
The results of the SG<sub>Local</sub> and SG<sub>Global</sub> predictions were compared by plotting them individually and plotting the difference between the two prediction maps by subtracting the SG<sub>Global</sub> results from the results SG<sub>Local</sub>. A visual scan of the difference map was done to check for abnormalities, such as extremes, and the mean of the difference map was calculated to assess whether one model gives systematically higher or lower predictions. Next, descriptive statistics of both model predictions were calculated and compared.

Accuracy
The accuracy of the prediction performance of both models was assessed using a test data corresponding to a subset of the local soil observations dataset that was set aside for testing. Details are provided in Section 2.6. Accuracy was assessed separately for each of the two soil properties. The accuracy of the prediction performance was gained by calculating the root mean square error (RMSE) to get the overall accuracy (Eq. 4), the mean error (ME) to quantify the prediction bias (Eq. 5), the R<sup>2</sup> to represent the fraction of explained variance (Eq. 6), and the prediction interval width as derived using QRF.

The RMSE and ME should be close to 0 for the best performance and indicate highest accuracy, as they represent the differences between the measured values and observed values. A model calibrates the data well if the differences between the observed values and the model's predicted values are small and unbiased. The R<sup>2</sup> is the fraction of the response variable variation that is explained by the model and it is expressed on a convenient 0 – 100% scale.

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (z(s_i) - \hat{z}(s_i))^2}
\]  
(4)

\[
ME = \frac{1}{n} \sum_{i=1}^{n} z(s_i) - \hat{z}(s_i)
\]  
(5)

where n is the total number of validation locations in the local study area. At each of these locations (s<sub>i</sub>), the difference between the actual value (z(s<sub>i</sub>)) and the predicted value (\hat{z}(s<sub>i</sub>)) is computed.

\[
R^2 = \left[1 - \frac{SSE}{SST}\right] \times 100
\]  
(6)

where SSE is the sum of squared errors at validation points, and SST is the total sum of squares. An R<sup>2</sup> of 100 indicates a perfect model where the model explains 100 % of the variation.

Scatter density plots of the predicted against observed values along with 1:1 lines were plotted. The closer the predictions are to the 1:1 line, the more accurate the model is. Lastly, for the two models the performance statistics, descriptive statistics and scatter plots were compared to assess the differences between model performances and conclude which model, the SG<sub>Global</sub> or the SG<sub>Local</sub> was most accurate for the local study area.
2.3 Influence of new soil observations

To assess the influence of new local soil observations on the global model, the new local soil observations were added to the global soil observations. This resulted in a new SoilGrids model, called the \(SG_{\text{Plus}}\) model. \(SG_{\text{Plus}}\) was calibrated on a new regression matrix holding the same covariates as the \(SG_{\text{Global}}\) model and the soil observations of the \(SG_{\text{Global}}\) and \(SG_{\text{Local}}\) models, see Table 2. For each soil property of interest, a prediction was made for all depth intervals of interest. The \(SG_{\text{Global}}\) model results of all three study areas were compared to \(SG_{\text{Plus}}\) model. The results of the \(SG_{\text{Plus}}\) and \(SG_{\text{Global}}\) predictions were assessed by plotting the difference between the two prediction maps. This was done by subtracting \(SG_{\text{Plus}}\) results from the \(SG_{\text{Global}}\) results. Assessments followed the same procedures as described in section 2.2.

It is possible that adding local data not only influences the predictions in the local area but also elsewhere in the world. This may happen in areas that have comparable environmental conditions as the local area (i.e., areas that have comparable covariate values). To derive which parts of the world have similar soil-forming factors, the Homosoil concept (Miller, 2012) was used. The Homosoil method searches for the smallest taxonomic distance of the ‘CLORPT’ soil-forming factors between a reference area and the region of interest, in this case, the local area. This includes climate, physiography, and parent materials (Mallavan, Minasny, & McBratney, 2010). Because this method does not explicitly take all covariates of the SoilGrids model into account, two areas were selected to review the results. One area with a small taxonomic distance is selected (called the “std study area”), and one with a large taxonomic distance (called the “ltd study area”). The ltd study area was used as a reference area, I expected the std study area to show some deviation in the difference maps, while the ltd area should show very little to no deviation. If the ltd study area did show a large deviation, it might be because Homosoil did not take all covariates into account.

Table 2. SoilGrids model settings to determine the influence of new soil observations.

<table>
<thead>
<tr>
<th>Model settings</th>
<th>(SG_{\text{Plus}})</th>
<th>(SG_{\text{Global}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soil Observations</td>
<td>Global + Local</td>
<td>Global</td>
</tr>
<tr>
<td>Used covariates</td>
<td>(SG_{\text{Global}}) RFE results</td>
<td>(SG_{\text{Global}}) RFE results</td>
</tr>
<tr>
<td>Prediction area</td>
<td>local, std &amp; ltd study area</td>
<td>local, std &amp; ltd study area</td>
</tr>
<tr>
<td>Soil Property</td>
<td>OC and soil pH</td>
<td>OC and soil pH</td>
</tr>
<tr>
<td>Prediction depth interval</td>
<td>0-5, 5-15 &amp; 15-30 cm depth layer</td>
<td>0-5, 5-15 &amp; 15-30 cm depth layer</td>
</tr>
</tbody>
</table>
2.4 Local data density

For traditional machine learning algorithms, accuracy typically increases according to a power-law (the more training data, the more accurate the model) and then reaches a plateau from whereon the accuracy of the predictions is just slightly changing. To analyse the influence of the local data density for the $SG_{Global}$ model, the prediction performance was evaluated for different newly added local data sample sizes. The goal hereby is to gain insight into the accuracy grown of the $SG_{Global}$ model for the local study area.

The local data samples were added to the $SG_{Global}$ model in eight ascending steps (Table 3). The last step was the results of RQ2. The $SG_{Global}$ model, including the new soil samples, is called $SG_{Plus}$. The model was re-calibrated seven times with the original $SG_{global}$ soil observations and the soil observations of the local area per ascending step (Table 4). The models are were used to make predictions for the local area and the local test dataset was used to gain the model accuracy. The results were presented as a learning curve. The prediction performance for the learning curve was calculated using the RMSE, calculated using from the test data.

<table>
<thead>
<tr>
<th>Run number</th>
<th>Number of new Soil samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>200</td>
</tr>
<tr>
<td>3</td>
<td>500</td>
</tr>
<tr>
<td>4</td>
<td>1000</td>
</tr>
<tr>
<td>5</td>
<td>2000</td>
</tr>
<tr>
<td>6</td>
<td>5000</td>
</tr>
<tr>
<td>7</td>
<td>10000</td>
</tr>
<tr>
<td>8</td>
<td>36898</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RQ3 Model settings</th>
<th>Number of new Soil samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>$SG_{Plus}$</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 3. Number of added local sample data used for assessing the effect of local data density.

Table 4. SoilGrids model settings to determine the influence of local data density.

2.5 Case study

The methods were tested using three different study areas. For the local study area, Andhra Pradesh, located in India (Figure 2), was chosen. This area was chosen because it represents a low sampling density area in $SG_{Global}$ and new soil observations were present. To test the difference between a global and a local model, new local soil samples were needed for the local study area. Within ISRIC there were 506704 new local soil samples of Andhra Pradesh pre-processed and in the same layout as the $SG_{global}$ soil samples ready to use. Nampula located in Mozambique was chosen as the std area. According to the Homosoil results (appendix III), this area is geographically most similar to the local study area. The Netherlands was chosen to represent the ltd study area. The Netherlands has a reasonable sample size in the $SG_{Global}$ model and it is taxonomically distinct from Andhra Pradesh.
Andhra Pradesh
The state Andhra Pradesh is located in the south-west of India (between 77° and 84° 40' East and 12° 41' and 22° North) with a total surface area of 160,000 km² and it is the fourth-biggest state in India. The state can be divided into three different zones. The Coastal Plains in the east is bordered by the Bay of Bengal and has a coastline of around 972 km. The Western Pediplains have considerable height differences. The elevation ranges from 0 to > 600 m above mean sea level. The Eastern Ghats follow the Coastal Plains closely. One outstanding area is the hill range in the North-East part. Here the elevation attains a height of 600 to 1200 m (P. Sudhakar, Reddy, Prasad, SatyaKumar, & Rao, 2017). Temperatures range from 12° to 30°C in winter to 20° to 41°C in summer. The monsoon season, which starts in July and continues till September, is the most extreme season where the state receives heavy rainfall (1150 mm). The State is dominated by cropland with large forests in the northern part and in the centre of the south (Rao & Wanmali, 2018). Figure 2 shows the soil map of Andhra Pradesh. The northern part is dominated by Nitisols, With Regosols near the sea. The southern part of the area is a mixture of Leptosols, Luvisols and Vertisols (Pike, 2018).

Netherlands
The Netherlands has a total surface area of 33.883 km² and is located between the North Sea to the north and west, Germany to the east and Belgium to the south. Most of the terrain consists of coastal lowlands, river deltas and reclaimed land, with some hills in the south-east. The elevation ranges from -7m till 321 meters above sea level. The Netherlands has a mild, maritime climate with generally warm summers and gentle winters. There is no rainy season in the Netherlands. Rain occurs throughout the whole year with spring as the driest season (Rowen & Heslinga, 2020). The centre and west part of the Netherlands is covered with thick layers of silt and gravel transported from the European mountains by the rivers Rhine and Maas (Cambisols and Fluvisols). Clay (Gleysols) is deposited in the sheltered lagoons.
behind the coastal dunes (Regosols) and Luvisols are mainly found in the southern part of the Netherlands. The rest of the Netherlands is covered by sand (podzols) with some Histosols between the Fluvisols and Podzols (Figure 3) (Maschinen, Investition, Beschaffungen, Ersatzbeschaffungen, & Mittelherkunft, 2012).

Nampula - Mozambique
Nampula is a province in the northern part of Mozambique and has a total area of 79,000 km². The Niassa Province borders Nampula to the north-west and west, the Zambezia Province to the south-west and the Indian Ocean is bordered to the east. The Ligonha River in the south-west of the area separates Nampula from Zambezia Province. In the west of are several hilly areas with mountains up to 1804m. Closer to the coastal area the land becomes less steep and near the coast, the rivers debouch into the Indian Ocean forming deltas at the northern and southern borders. The climate in Mozambique is tropical humid. There is a humid season from November to March and a dry season from April to October (Penvenne & Sheldon, 2020). Figure 4 shows the distribution of the major soil types in Nampula. The centre of the area is covered with Arenosols, with Lixisols to the east and west. The north-east of the area, from inland to the coast, is covered with Leptosols and Vertisols. The South-east of the area, also name from inland, is covered with Plansosols, Arenosols and Fluvisols (Maschinen et al., 2012).
2.6 Data and materials

Three data sources were used in this research; soil sample data, a stack of covariate layers and the ESA land cover map. ISRIC provided all data, and therefore not much further pre-processing was needed.

SoilGrids Soil observations

The SoilGrids soil observations dataset is a compilation of hundred heterogeneous point datasets from all over the world that have been brought together in a standardised database (WoSIS). The WoSIS database itself is developed and maintained by ISRIC. The datasets which contribute to the final SoilGrids observations dataset are provided by several external soil-related organisations, governmental organisation and several databases available within ISRIC (e.g. WISE, AFSP and SCOTER) (Batjes et al., 2017).

For organic carbon, there are 152.350 records with soil sample locations and 602.979 soil observations in the Soil Sample dataset. For Soil pH, there are 150.798 records with soil sample data and 659.473 observation spread over all continents (see Figure 5). Not all observations are sampled observation. Besides the soil observations, the soil sample dataset also includes expert-based pseudo-observations. Some large areas that have extreme climatic conditions and/or have very restricted access are significantly under-sampled (e.g. Greenland, the Sahara or northern parts of Russia) (Hengl et al., 2017). Soil samples are a mix of profile data (described and sampled horizons) for fixed sample depths (de Sousa, n.d.).
The soil data includes the profile location, sampling depth and the soil property values. An overview of the data is shown in table 5.

*Table 5. Head of the global soil observations data.*

<table>
<thead>
<tr>
<th>LAT</th>
<th>LON</th>
<th>PH</th>
<th>DEPTH</th>
</tr>
</thead>
<tbody>
<tr>
<td>80.783330</td>
<td>47.50000</td>
<td>6.6</td>
<td>2.5</td>
</tr>
<tr>
<td>80.783330</td>
<td>47.50000</td>
<td>6.7</td>
<td>12.5</td>
</tr>
<tr>
<td>80.783330</td>
<td>47.50000</td>
<td>7.3</td>
<td>30.0</td>
</tr>
<tr>
<td>35.20140</td>
<td>-91.98920</td>
<td>4.7</td>
<td>51.0</td>
</tr>
<tr>
<td>35.20140</td>
<td>-91.98920</td>
<td>4.9</td>
<td>32.0</td>
</tr>
</tbody>
</table>

*India soil observations*

The India soil observations were derived from the Soil Health Card Scheme program. The Soil Health Card scheme is an initiative launched by the Government of India in 2015. Soil Health Cards were distributed to farmers all around India. A soil sample from the topsoil (0-15 cm depth) of the farmers' field was analysed in the lab on 12 parameters including soil pH and soil organic carbon content (Mishra, Nair, Singh, Gazeley, & Kapoor, 2015). These data were published as open data and downloaded by ISRIC. ISRIC pre-processed the data and harmonises it to the SoilGrids framework.

The India dataset represents 506704 soil sample locations covering the entire Andhra Pradesh region and represents profile location, and the soil property value (table 6). Because we wanted to assess the effect of adding new data and wanted to add a reasonable number of additional data points, only 10% of the data was used as training data and the remaining data as test data. To ensure that the spatial distribution of the data was retained during subsampling of the training data, the subsamples were...
based on a density estimate of the original observations. The r function Point process random subsample (pp.subsample) was used for this (Evans, 2018). The default settings of this model were used. To merge the local (India) soil observations dataset and the SGlobal soil observations dataset, all covariates need a value for all locations. The column ‘sample depth’ is not present in the India dataset and needed to be added. Because the soil observations in the India dataset are sampled at a sample depth between 0-15 cm, the average soil sample depth of 7.5 was used as sample depth for all soil samples.

Table 6. Head of the local soil observations data.

<table>
<thead>
<tr>
<th>LAT</th>
<th>LON</th>
<th>OC</th>
</tr>
</thead>
<tbody>
<tr>
<td>15.35495</td>
<td>77.11725</td>
<td>0.30</td>
</tr>
<tr>
<td>13.87064</td>
<td>79.00641</td>
<td>0.27</td>
</tr>
<tr>
<td>16.64445</td>
<td>79.67471</td>
<td>0.31</td>
</tr>
<tr>
<td>15.61001</td>
<td>79.89583</td>
<td>1.09</td>
</tr>
<tr>
<td>15.58786</td>
<td>78.41065</td>
<td>0.05</td>
</tr>
<tr>
<td>14.37179</td>
<td>77.97894</td>
<td>0.89</td>
</tr>
<tr>
<td>16.41518</td>
<td>60.36449</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Soil properties and depths
The used soil properties for this research are; soil organic carbon (SOC) content in g/kg and soil pH in water. Organic carbon is together with pH, the best indicator of the health status of the soil structure (Maschinen et al., 2012). Therefore, the analysis was limited to these two soil properties.

This research focused on the 5-15 cm depth layer, as the new local soil sample data are enclosed in the selected depth interval. Research question 2 also focuses on the influence of new local soil samples on higher and lower depth intervals. Therefore, for research question 2; influence of new soil observations, the 0-5, 5-15 and 15-30 cm depth intervals are used, predicting at 2.5 cm, 10 cm and 22.5 cm depth.

Covariate stack
The original covariate stack consisted of 405 different covariates, which are primarily based on remote sensing data. These covariates were selected to represent factors of soil formation, according to Jenny, 1945. After the correlation analyses, performed by ISRIC, 129 covariates remained. Appendix II shows the table with all covariates and the description of their topic. All those covariates are stored in a raster stack with 250 x 250 m resolution and have the same spatial extent as the ESA land cover map.

ESA land cover map
The global soil mask map derived from the latest ESA land cover map (Defourny, 2017) was used to mask out areas where no predictions of soil properties can be made because there is no soil present at these locations. Those areas were represented as NoData and turn up white in all maps. The classes Urban (code 190), inland water (code 210), glacier (code 220) and bare surface (code 200) were masked out as well (figure 5). No predictions are made for permanent ice areas since they are subject to extreme
climatic conditions and therefore cannot be cultivated. The masked out areas are often under-represented in soil surveys, making it difficult to calibrate a reliable statistical model.

**Statistical software**
Analyses were conducted using freely available R software, which is a language and environment for statistical computing and graphics (The R Foundation, 2020). The advantage of using a scripting language is that the study is reproducible and reported in the scripts. The main Packages that were used during this research are shown in table 7.

*Table 7. Used R libraries to create the results for the research.*

<table>
<thead>
<tr>
<th>Library</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>raster</td>
<td>manipulating and using calculations on raster data (Etten et al., 2020).</td>
</tr>
<tr>
<td>sp</td>
<td>manipulating and plotting spatial data (Hijmans et al., 2020).</td>
</tr>
<tr>
<td>caret</td>
<td>to create the random forest (ranger) and the RFE (Max et al., 2020).</td>
</tr>
<tr>
<td>tidyverse</td>
<td>gain RFE functionalities (Wickham, 2019).</td>
</tr>
<tr>
<td>devtools</td>
<td>Provide input data and functions to run the HomoSoil script (Hester, 2020).</td>
</tr>
</tbody>
</table>
3 Results
In this chapter, the results are described per research question. Results of modelling and mapping soil organic carbon and pH are provided separately, starting with the soil organic carbon results.

3.1 Comparison of a global and local SoilGrids model
The SG\textsubscript{Global} model and SG\textsubscript{Local} model were compared firstly by analysing the variables which were most important in explaining the soil property and by comparing model statistics. Next, the prediction maps and difference maps were assessed using descriptive statistics and visual comparison. Lastly, the accuracy of the map was evaluated based on scatter plots and RMSE, ME and R-squared.

3.1.1 Soil Organic Carbon content
Model comparison
Figure 6 shows the used covariates ranked by variable importance of the SG\textsubscript{Global} model. The full name of the covariates can be found in appendix II. From the 129 covariates presented to the RFE function, the SG\textsubscript{Global} model selected 45 covariates in total to represent the global area, while the SG\textsubscript{Local} model used 29 covariates. Both the global and the local model identified 29 identical covariates required to build an accurate model. What strikes is, from the top 4 covariates used by SG\textsubscript{Global}, the SG\textsubscript{Local} model used only one; total yearly radiation (CLM\_WCL\_SRCYRSUM). The remaining three global variables are, soil sample depth (DEPTH), SD yearly snowfall prob. at 500 m (CLM\_ESA\_SYRSTD) and annual temperature range (max temperature of the warmest month – min temperature of the coldest month) (CLM\_WCL\_BIO07). From the four most important covariates of SG\textsubscript{Local}, two of them (total radiation for August; CLM\_WCL\_S08RAD and Digital Elevation model; MOR\_ENV\_DEMM) are also in the top 10 of the SG\textsubscript{Global} most important covariates. The other two, total monthly precipitation at 1 km for October (CLM\_WCL\_P10TOT) and precipitation of wettest quarter (CLM\_WCL\_BIO16), are in the bottom half of the SG\textsubscript{Global} covariate importance.

![Figure 6. Variable importance of the SG\textsubscript{Global} and SG\textsubscript{Local} SOC models ranked by SG\textsubscript{Global} variable importance.](image-url)
Map comparison
The SG<sub>Local</sub> and SG<sub>Global</sub> SOC prediction maps are shown on the right side of figure 7. The SG<sub>Local</sub> map (top right) predicts SOC percentage between 0% and 1% SOC in the south with some higher predictions to the north. The SG<sub>Global</sub> prediction map mainly shows predictions between 1.5% and 5% SOC, with some lower predictions in the south-west.

The differences map between the SG<sub>Global</sub> and the SG<sub>Local</sub> predictions is shown on the left side in Figure 7. Areas with relatively high SOC (between -2% and -3% SOC difference) in the north correspond to forest areas and the areas in the south to mountain areas. For those areas the SG<sub>Global</sub> model predicts a higher pH than the SG<sub>Local</sub> model. Most other areas representing a difference between -0.5% and -1.5% SOC. This implies that the SG<sub>Global</sub> model predicts higher SOC than the SG<sub>Local</sub> model, which also shows up in Table 8. The mean difference is -0.41% SOC, revealing that SG<sub>Global</sub> predicts overall higher SOC values. The absolute mean, which shows the mean deviation between the maps without taking into account which model predicts higher, is 1.41% SOC. If the observed statistics are compared to the predicted statistics of both models, SG<sub>Local</sub> seems to do a better job.

Table 8. Descriptive statistics of the SOC models vs Observed values in Andhra Pradesh in % SOC.

<table>
<thead>
<tr>
<th></th>
<th>SG&lt;sub&gt;Local&lt;/sub&gt;</th>
<th>SG&lt;sub&gt;Global&lt;/sub&gt;</th>
<th>Observed</th>
<th>Difference SG&lt;sub&gt;Local&lt;/sub&gt; - SG&lt;sub&gt;Global&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>0.07</td>
<td>0.76</td>
<td>0.02</td>
<td>-7.29</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>0.42</td>
<td>1.59</td>
<td>0.32</td>
<td>-1.71</td>
</tr>
<tr>
<td>Median</td>
<td>0.53</td>
<td>2.01</td>
<td>0.49</td>
<td>-1.27</td>
</tr>
<tr>
<td>Mean</td>
<td>0.71</td>
<td>2.13</td>
<td>0.58</td>
<td>-0.41</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>0.68</td>
<td>2.51</td>
<td>0.72</td>
<td>-0.94</td>
</tr>
<tr>
<td>Max.</td>
<td>2.47</td>
<td>9.23</td>
<td>2.60</td>
<td>0.95</td>
</tr>
<tr>
<td>Abs Mean</td>
<td></td>
<td></td>
<td></td>
<td>1.41</td>
</tr>
</tbody>
</table>
Figure 7. Soil pH prediction maps and difference map (top) of $SG_{\text{local}}$ (bottom left) and $SG_{\text{global}}$ (bottom right) for Andhra Pradesh - India.
**Accuracy**

To evaluate if \( \text{SG}_{\text{Local}} \) is more accurate than \( \text{SG}_{\text{Global}} \), scatter plots of the model predictions for the Organic Carbon models (Figure 8) were made. These plots show the predicted values vs the observed values and a 1:1 line. The more the dots are aligned around the line, the closer the predicted values are to the observed values. The observed values are based on the testing dataset of India, hence the scatter plots and the descriptive statistics (Table 9) are only representative for Andhra Pradesh. The legend shows the number of predictions per plot-pixel.

The \( \text{SG}_{\text{Global}} \) scatter plot shows that most of the dots are below the 1:1 line, indicating that the model predicts higher SOC values than observed. The scatterplot modelled with global data only, shows the densest area for predicted SOC between 1.2% and 1.9% SOC, while the densest area for observed SOC values is between 1.2% and 2.2% SOC. This also shows in the descriptive statistics (Table 9), where \( \text{SG}_{\text{Global}} \) scores much higher for each validation metric. Therefore, the difference map shows more negative than positive values. The scatter plot of the \( \text{SG}_{\text{Local}} \) model is more around the 1:1 line.

![Figure 8. Scatterplots of the global (left) and local (right) SOC model.](image)

<table>
<thead>
<tr>
<th></th>
<th>( \text{SG}_{\text{Local}} )</th>
<th>( \text{SG}_{\text{Global}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE (% SOC)</td>
<td>0.286</td>
<td>1.720</td>
</tr>
<tr>
<td>ME</td>
<td>0.008</td>
<td>1.497</td>
</tr>
<tr>
<td>( R^2 ) (% of the variation)</td>
<td>42.5</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Table 9. Descriptive statistics of the global and local SOC models with SOC in %.
3.1.2 Soil pH

**Model comparison**

After the RFE for the soil pH models, the $SG_{global}$ model selected 31 covariates and $SG_{local}$ model selected 29 covariates (Figure 9). 23 covariates were identically selected in both models. The top five covariates of the $SG_{global}$ model only have one covariate that was also selected for the $SG_{local}$ model; total monthly precipitation at 1 km in April (CLM_WCLP04TOT). The other four were not selected; global 30m tree cover (LUC_GFC_TRELY10), soil sample depth (DEPTH), total monthly precipitation at 1 km in October (CLM_WCL_P10TOT) and precipitation of wettest quarter (CLM_WCL_BIO16). Of the four most interesting peaks in $SG_{local}$, two are in the lowest 8 covariates of $SG_{global}$; Total monthly precipitation at 1 km for September (CLM_WCL_P09TOT) and Precipitation of Warmest Quarter (CLM_WCL_BIO18). $SG_{global}$ does not use one of the four most important covariates of the $SG_{local}$ model; Bioclimatic zones: zone 32 (ECO_USG_Z32).

![Figure 9. Variable importance of the global and local soil pH models.](image-url)
Map comparison
The left map in Figure 10 shows the pH difference map between $SG_{Local}$ minus $SG_{Global}$. The legend represents the difference between the two models in soil pH. Most of the map is represented by areas where $SG_{Local}$ predicts a higher pH than $SG_{Global}$. This also shows in the mean of the descriptive statistics table (Table 10), where $SG_{Global}$ has a mean of 6.73 pH and $SG_{Local}$ a mean of 7.37 pH. Other descriptive statistics show that the $SG_{Global}$ and $SG_{Local}$ model predicts 0.64 and 0.38 pH higher than the minimum soil observation. The maximum pH predicted with the $SG_{Global}$ model is 0.9 pH lower than the maximum prediction of soil pH, while $SG_{Local}$ has a 0.07 pH difference.

Table 10. Descriptive statistics of the Soil pH models vs Observed values in Andhra Pradesh.

<table>
<thead>
<tr>
<th></th>
<th>$SG_{Local}$</th>
<th>$SG_{Global}$</th>
<th>Observed</th>
<th>Difference $SG_{Local} - SG_{Global}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>5.18</td>
<td>5.44</td>
<td>4.80</td>
<td>-1.81</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>7.01</td>
<td>6.46</td>
<td>6.90</td>
<td>0.24</td>
</tr>
<tr>
<td>Median</td>
<td>7.29</td>
<td>6.80</td>
<td>7.41</td>
<td>0.51</td>
</tr>
<tr>
<td>Mean</td>
<td>7.37</td>
<td>6.73</td>
<td>7.38</td>
<td>0.52</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>7.79</td>
<td>7.05</td>
<td>7.90</td>
<td>0.81</td>
</tr>
<tr>
<td>Max.</td>
<td>9.05</td>
<td>8.01</td>
<td>9.12</td>
<td>2.22</td>
</tr>
<tr>
<td>Abs Mean</td>
<td></td>
<td></td>
<td></td>
<td>0.56</td>
</tr>
</tbody>
</table>
Figure 10. Soil pH prediction maps and difference map (top) of $SG_{local}$ (left) and $SG_{global}$ (right) for Andhra Pradesh.
Accuracy
Figure 11 shows scatter plots of the SG\textsubscript{Global} and SG\textsubscript{Local} models. Note that the observed values are based on the testing dataset of India, meaning that the scatter plots and the statistics are only representative for Andhra Pradesh. The scatter plot modelled with global data shows that the densest point area lies above the 1:1 line, meaning the model predicts lower soil pH values than observed. This also shows in the descriptive statistics (Table 10) where the SG\textsubscript{Global} model prediction mean is 0.85 pH units lower than the observed value mean.

The scatter plot modelled with local data only is close to the 1:1 line. The predicted values below 7.5 pH diverge a bit below the 1:1 line, indicating that the model predicts too high pH values, while the predicted values above 7.5 pH diverge a bit above the 1:1 line, indicating the model predicts too low. This also shows in the descriptive statistics (Table 11), the minimum value is higher than the observed value, while the maximum predictive value is lower than the observed value.

![Figure 11. Scatterplots of the global (left) and local (right) soil pH model.](image)

**Table 11. Descriptive statistics of the local and Global soil pH models.**

<table>
<thead>
<tr>
<th></th>
<th>SG\textsubscript{Local}</th>
<th>SG\textsubscript{Global}</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE (soil pH)</td>
<td>0.467</td>
<td>0.837</td>
</tr>
<tr>
<td>ME</td>
<td>-0.001</td>
<td>-0.490</td>
</tr>
<tr>
<td>R\textsuperscript{2} (% of the variation)</td>
<td>59.1</td>
<td>14.6</td>
</tr>
</tbody>
</table>
3.2 Influence of new soil observations

For each study area, the difference maps between SG\textsubscript{plus} and SG\textsubscript{Global} (SG\textsubscript{plus} minus SG\textsubscript{Global}) are shown per soil property at three different depths. The main highlights per study area and soil property are stated, starting with Andhra Pradesh, then Nampula and ending with The Netherlands. The section starts with the model statistics and variable importance.

3.2.1 model comparison

Figure 12 shows the variable importance for the SG\textsubscript{Global} and SG\textsubscript{Plus} models for SOC. Both models use the same RFE selected covariates, but the importance deviates. In the top four most important covariates from the SG\textsubscript{Global} models the CLM\_WCL\_SRDYSRUM (Total Yearly Solar radiation at 1 km), CLM\_ESA\_SYRST (SD yearly snowfall prob. at 500 m) and CLM\_WCL\_BIO07 (annual temperature range; maximum temp of the warmest month minus minimum temp of the coldest month) deviate most.

Table 12 shows the OOB error statistics for the SG\textsubscript{Global} and SG\textsubscript{Plus} models. The OOB error is based on all global observations. The RMSE is lower for the SG\textsubscript{Plus} model and the ME is even significantly lower for the SG\textsubscript{Plus}. The R\textsuperscript{2} however is almost equal.

![Variable importance of the SG\textsubscript{Global} and SG\textsubscript{Plus} SOC models based on the variable importance ranking of SG\textsubscript{Global}.

Table 12. OOB statistics for SG\textsubscript{Global} and SG\textsubscript{Plus} for SOC (%).

<table>
<thead>
<tr>
<th></th>
<th>SG\textsubscript{Global}</th>
<th>SG\textsubscript{Plus}</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE (% SOC)</td>
<td>3.056</td>
<td>2.950</td>
</tr>
<tr>
<td>ME</td>
<td>9.341</td>
<td>8.704</td>
</tr>
<tr>
<td>R\textsuperscript{2} (% of the variation)</td>
<td>71.7</td>
<td>72.1</td>
</tr>
</tbody>
</table>
Figure 12 shows the variable importance for the SG\textsubscript{Global} and SG\textsubscript{Plus} models for soil pH. Both models use the same RFE selected covariates, but the importance deviates. From the six most deviation covariate importance, five are the top five most important covariates from the SG\textsubscript{Global} model. Those covariates are; LUC\_GFC\_TRELY10 (Global 30m Tree Cover), Depth (sample depth), CLM\_WCL\_P04TOT (Total monthly precipitation at 1 km for April), CLM\_WCL\_P10TOT (Total monthly precipitation at 1 km for October) and CLM\_WCL\_BIO16 (Precipitation of Wettest Quarter). The other deviating covariate is CML\_WCL\_BIO9 (Mean Temperature of Driest Quarter).

Table 12 shows the OOB error statistics for the SG\textsubscript{Global} and SG\textsubscript{Plus} models. All model statistics are almost equal for both models, but the SG\textsubscript{Plus} model scores slightly better.

Table 13. OOB statistics for SG\textsubscript{Global} and SG\textsubscript{Plus} for soil pH.

<table>
<thead>
<tr>
<th></th>
<th>SG\textsubscript{Global}</th>
<th>SG\textsubscript{Plus}</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE (pH)</td>
<td>0.539</td>
<td>0.535</td>
</tr>
<tr>
<td>MSE</td>
<td>0.290</td>
<td>0.286</td>
</tr>
<tr>
<td>R(^2) (% of variation)</td>
<td>0.842</td>
<td>0.842</td>
</tr>
</tbody>
</table>

Figure 13. Variable importance of the SG\textsubscript{Global} and SG\textsubscript{Plus} soil pH models.
3.2.2 Andhra Pradesh

*Organic Carbon*

The 0-5 cm depth map (left map of Figure 14) shows mainly areas with a difference between 0% and -1% SOC. In the centre of the area and in the south-west part, the difference is between -1% and -3% SOC. In the south-east part, the difference is between 1% and 2% SOC. This also shows in the statistics table (Table 14) where the difference between the SGPlus and the SGGlobal mean is 0.35% SOC, the least difference of all three SOC depth maps. The minimum and median are similar, while the maximum deviates around 1% SOC.

The 5-15 cm depth map (middle map of Figure 14) shows mainly areas with a difference between -2% and -3% SOC. This means the SGGlobal model predicts higher SOC values than the SGPlus model. This also shows in the statistics table, where the mean of SGPlus is 0.71% SOC, and the mean of SGGlobal is 2.13 % SOC. This is the highest mean difference between all three SOC depth maps.

The 15-30 cm depth map (right map of Figure 14) shows mainly areas with a deviation between 0% and -1% SOC. The southern part of the area shows a deviation between -1% till -2% SOC, with in northern part a small area representing a difference between 0.5% and 1.5% SOC. The statistics table also shows that the SGGlobal predictions are higher than the SGLocal predictions, though the mean difference is smaller than the 5-15 cm depth map.

**Table 14. Statistics table of SOC (%) in Andhra Pradesh – India.**

<table>
<thead>
<tr>
<th></th>
<th>0 - 5 cm</th>
<th>5 - 15 cm</th>
<th>15 - 30 cm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SGPlus</td>
<td>SGGlobal</td>
<td>SGPlus</td>
</tr>
<tr>
<td>Min.</td>
<td>0.76</td>
<td>0.89</td>
<td>0.08</td>
</tr>
<tr>
<td>Median</td>
<td>2.37</td>
<td>2.75</td>
<td>0.63</td>
</tr>
<tr>
<td>Mean</td>
<td>2.55</td>
<td>2.90</td>
<td>0.71</td>
</tr>
<tr>
<td>Max.</td>
<td>11.1</td>
<td>10.00</td>
<td>7.22</td>
</tr>
</tbody>
</table>

*Figure 14. Difference maps of SOC (%) for three interval depths.*
Soil pH
Figure 1 shows the difference maps of soil pH for Andhra Pradesh. The difference maps for all depths show similar patterns, where the 0-5 cm depth map has the least deviation, meaning the difference between the model results is lower. The 5-15 cm and 15 – 30 cm depth maps have the most deviation, meaning the difference between the model results is larger. The largest difference areas, where the SGPlus model predicts higher, are located in the centre of the difference maps. Closer to the borders, the difference gets less and there are some areas near the border where the SGGlobal predicts higher soil pH. Table 1 shows that all the differences between the statistics are largest for the 5-15 cm depth map and smallest for the 0-5 cm depth map. The median, mean and max are for all SGPlus models at all tree depths higher than the SGGlobal model.

Table 1. Statistics table of soil pH for Andhra Pradesh – India.

<table>
<thead>
<tr>
<th></th>
<th>0 - 5 cm</th>
<th>5 - 15 cm</th>
<th>15 - 30 cm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SGPlus</td>
<td>SGGlobal</td>
<td>SGPlus</td>
</tr>
<tr>
<td>Min.</td>
<td>5.24</td>
<td>5.47</td>
<td>5.18</td>
</tr>
<tr>
<td>Median</td>
<td>7.22</td>
<td>6.75</td>
<td>7.29</td>
</tr>
<tr>
<td>Mean</td>
<td>7.15</td>
<td>6.70</td>
<td>7.26</td>
</tr>
<tr>
<td>Max.</td>
<td>8.69</td>
<td>7.00</td>
<td>9.05</td>
</tr>
</tbody>
</table>

Figure 15. Difference maps of soil pH for three interval depths, Andhra Pradesh – India.
3.2.3 Nampula

Organic Carbon

Figure 16 shows the results for Organic Carbon in Nampula. Overall, the three depth maps do not show substantial differences. In the 0 - 5 cm depth map, the most considerable difference is located in the north-east and south-eastern part of the area (near the coast). In the north-eastern part of the area, there is a small strip where the deviation is between -1% and -1.5% SOC. In the south-eastern part of the area, there is a small area where the deviation is between 1% and 2% SOC.

At the eastern border of the 5-15 cm depth difference map, there is a deviation between -1% and -3% SOC. In the southern point of the study area, there is a deviation between 0.5% and 1% SOC. In the 15 – 30 cm depth map, the same areas show the most considerable differences, where the deviation in the south is more spread. The statistic table for Organic Carbon (Table 16) also shows that the deviation between the two models for all depths is small. The maximum mean deviation is 0.5% SOC, and the min, max and median for each model and depth are almost equal.

Table 16. Statistical table of SOC (%) in Nampula – Mozambique.

<table>
<thead>
<tr>
<th></th>
<th>SGPlus</th>
<th>SGGlobal</th>
<th>SGPlus</th>
<th>SGGlobal</th>
<th>SGPlus</th>
<th>SGGlobal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>0.73</td>
<td>0.73</td>
<td>0.34</td>
<td>0.33</td>
<td>0.40</td>
<td>0.34</td>
</tr>
<tr>
<td>Median</td>
<td>1.57</td>
<td>1.54</td>
<td>1.31</td>
<td>1.28</td>
<td>1.05</td>
<td>1.02</td>
</tr>
<tr>
<td>Mean</td>
<td>1.70</td>
<td>1.66</td>
<td>1.38</td>
<td>1.36</td>
<td>1.13</td>
<td>1.08</td>
</tr>
<tr>
<td>Max</td>
<td>13.63</td>
<td>15.96</td>
<td>9.08</td>
<td>12.34</td>
<td>7.94</td>
<td>9.17</td>
</tr>
</tbody>
</table>

Figure 16. Difference maps of SOC (%) for three interval depths, Nampula – Mozambique.
Soil pH

Figure 17 shows that the difference map for soil pH for all three depth is almost equal. The areas with a difference higher than 0.25 pH show near the east border a slightly bigger area in the 15-30 cm depth map than the 0-5 cm depth map, but the areas with a difference higher than 0.25 pH in the western part of the area shows less dark red in the 15-30 cm depth map. This also shows in the statistic table for soil pH (Table 17), the mean of all predictions is equal for 0-5 and 5-15 cm depth and differs 0.01 pH of 15-30 cm depth.

Table 17. Statistical table of soil pH in Nampula – Mozambique.

<table>
<thead>
<tr>
<th>Depth</th>
<th>SGPlus</th>
<th>SGGlobal</th>
<th>SGPlus</th>
<th>SGGlobal</th>
<th>SGPlus</th>
<th>SGGlobal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>5.13</td>
<td>5.19</td>
<td>5.13</td>
<td>5.15</td>
<td>5.08</td>
<td>5.08</td>
</tr>
<tr>
<td>Median</td>
<td>6.07</td>
<td>6.07</td>
<td>6.06</td>
<td>6.06</td>
<td>6.04</td>
<td>6.03</td>
</tr>
<tr>
<td>Mean</td>
<td>6.08</td>
<td>6.08</td>
<td>6.07</td>
<td>6.07</td>
<td>6.05</td>
<td>6.04</td>
</tr>
<tr>
<td>Max</td>
<td>7.62</td>
<td>7.62</td>
<td>7.72</td>
<td>7.66</td>
<td>7.70</td>
<td>7.65</td>
</tr>
</tbody>
</table>

Figure 17. Difference maps of soil pH for three interval depths, Nampula – Mozambique.
3.2.4 Netherlands

**Organic Carbon**

The 0-5 depth difference map (Figure 18, left map) shows big areas with large differences in the northern and north-eastern part of the area and some in the western part, where the SG$_{\text{Plus}}$ model predicts between -1.5% and -2% SOC higher. The centre and southern part of the area show a significant amount of pixels indicating that the SG$_{\text{Global}}$ model predicts higher SOC values than the SG$_{\text{Local}}$ model. The statistics table (Table 18) also shows that the means are quite similar, but a slightly higher mean for SG$_{\text{Plus}}$; SG$_{\text{Plus}}$ 9.82% and SG$_{\text{Global}}$ 9.40% SOC.

The 5-15 cm Organic Carbon difference map shows an area of predominantly large differences (between 1.5% and 3% SOC) in the northern part of the Netherlands. This indicates that the SG$_{\text{Plus}}$ model predicts higher Organic Carbon than the SG$_{\text{Global}}$ model. Compared to the 0-5 and 15-30 cm depth maps, this prediction maps shows the least deviation, indicating that the model predictions are close to each other. The means of both maps are similar.

The 15-30 cm depth map shows bigger areas with higher negative differences (between -2% and -3% SOC) than the other depth maps. Those areas indicate that SG$_{\text{Global}}$ predicts a higher SOC %. Only the centre and southern part of the area show small dense areas of positive prediction values, which indicates that the SG$_{\text{Global}}$ model predicts higher SOC values that the SG$_{\text{Plus}}$ model. The statistics table shows that the median and mean both are around 1 % higher for SG$_{\text{Global}}$ than for SG$_{\text{Local}}$.

**Table 18. Statistical table of SOC (%) in the Netherlands.**

<table>
<thead>
<tr>
<th>Depth</th>
<th>SG$_{\text{Plus}}$</th>
<th>SG$_{\text{Global}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 5 cm</td>
<td>Min. 1.90</td>
<td>1.88</td>
</tr>
<tr>
<td></td>
<td>Median 9.88</td>
<td>9.41</td>
</tr>
<tr>
<td></td>
<td>Mean   9.82</td>
<td>9.40</td>
</tr>
<tr>
<td></td>
<td>Max. 37.71</td>
<td>36.18</td>
</tr>
<tr>
<td>5 - 15 cm</td>
<td>Min. 1.23</td>
<td>1.16</td>
</tr>
<tr>
<td></td>
<td>Median 5.09</td>
<td>4.85</td>
</tr>
<tr>
<td></td>
<td>Mean   5.48</td>
<td>5.18</td>
</tr>
<tr>
<td></td>
<td>Max. 36.78</td>
<td>34.87</td>
</tr>
<tr>
<td>15 - 30 cm</td>
<td>Min. 0.61</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>Median 6.63</td>
<td>7.49</td>
</tr>
<tr>
<td></td>
<td>Mean   7.11</td>
<td>8.11</td>
</tr>
<tr>
<td></td>
<td>Max. 42.44</td>
<td>43.66</td>
</tr>
</tbody>
</table>

*Figure 18. Difference maps of soil SOC (%) for three interval depths in the Netherlands.*
Soil pH
The 0-5 cm depth map (Figure 19, left map) shows the most diverge results of the three difference maps. Equally spread over the entire area, there are clusters of high difference areas (between 0.25 and 1 pH) and areas with low differences (around 0.0 pH). In between there are some outliers till -1 pH. Still, the overrepresented deviation between 0.5 and 1 pH indicates that the SGPlus model predicts higher pH values than the SGGlobal model. The statistical table (Table 19) only shows a difference of 0.04 pH in the mean of the predictions.

The difference maps for 5-15 cm and 15-30 cm depth (Figure 19, middle and right map) are quite similar. There is a cluster of high deviation (0.5 pH till 1.0 pH) near the western part, but no future outstanding areas with high differences. The rest of the area deviates between 0.5 pH and -0.5 pH deviation. The statistics table shows no significant differences between the two models.

Table 19. Statistical table of soil pH in the Netherlands.

<table>
<thead>
<tr>
<th>Depth Interval</th>
<th>SGPlus Min</th>
<th>SGGlobal Min</th>
<th>SGPlus Median</th>
<th>SGGlobal Median</th>
<th>SGPlus Mean</th>
<th>SGGlobal Mean</th>
<th>SGPlus Max</th>
<th>SGGlobal Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 5 cm</td>
<td>3.69</td>
<td>3.69</td>
<td>5.98</td>
<td>5.99</td>
<td>5.94</td>
<td>5.90</td>
<td>8.18</td>
<td>8.13</td>
</tr>
<tr>
<td>5 - 15 cm</td>
<td>3.75</td>
<td>3.74</td>
<td>6.01</td>
<td>6.01</td>
<td>6.13</td>
<td>6.12</td>
<td>8.07</td>
<td>8.05</td>
</tr>
<tr>
<td>15 - 30 cm</td>
<td>3.82</td>
<td>3.79</td>
<td>7.61</td>
<td>5.99</td>
<td>6.11</td>
<td>6.11</td>
<td>7.26</td>
<td>9.34</td>
</tr>
</tbody>
</table>

Figure 19. Difference maps between SGPlus and SGGlobal of soil pH for three interval depths – the Netherlands.
3.3 Local data density

Figure 21 shows a series of maps where, with each step, the amount of new local soil observations added to the $SG_{\text{Global}}$ model increases. The maps show the difference between the $SG_{\text{plus}n}$ model and the $SG_{\text{Global}}$ model. For each map predicted with more local observations as the previous one, the difference between the two models becomes larger. This starts at the centre of the area and spreads out more to the North and South. Most parts of the map contain deviation that indicate that the $SG_{\text{plus}n}$ model predicts a higher soil pH than the $SG_{\text{Global}}$ model (from 0 till -2). Still there are a few areas, mainly in the south near the borders and some in the north, where the $SG_{\text{Global}}$ model predicts higher.

Table 20 shows the model statistics for each prediction map and the observed values. Hereby noting that the observed values are randomly taken soil observations (around 37,000) and therefore do not cover the entire study area. It might be possible that the real minimum and maximum differ from the one stated here, but it gives a general overview of minimum values for min and max are. Figure 20 shows how the statistics slowly move closer to the observed values as $n$ increases, lifting steeper at the beginning of let line and flattening around n2000.

<table>
<thead>
<tr>
<th>n100</th>
<th>n200</th>
<th>n500</th>
<th>n1000</th>
<th>n2000</th>
<th>n5000</th>
<th>n10000</th>
<th>n36898</th>
<th>SGGlobal</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>5.34</td>
<td>5.50</td>
<td>5.37</td>
<td>5.41</td>
<td>5.31</td>
<td>5.43</td>
<td>5.15</td>
<td>5.18</td>
<td>5.44</td>
</tr>
<tr>
<td>Median</td>
<td>7.11</td>
<td>7.17</td>
<td>7.21</td>
<td>7.02</td>
<td>7.25</td>
<td>7.26</td>
<td>7.29</td>
<td>7.29</td>
<td>6.80</td>
</tr>
<tr>
<td>Mean</td>
<td>7.00</td>
<td>7.07</td>
<td>7.13</td>
<td>7.14</td>
<td>7.21</td>
<td>7.22</td>
<td>7.25</td>
<td>7.26</td>
<td>6.73</td>
</tr>
<tr>
<td>Max</td>
<td>8.54</td>
<td>8.38</td>
<td>8.60</td>
<td>8.69</td>
<td>8.73</td>
<td>8.81</td>
<td>8.91</td>
<td>9.05</td>
<td>8.01</td>
</tr>
<tr>
<td>abs. Mean diff</td>
<td>0.28</td>
<td>0.35</td>
<td>0.40</td>
<td>0.42</td>
<td>0.48</td>
<td>0.50</td>
<td>0.54</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 20. Learning curve of the predicted mean of the $SG_{\text{plus}n}$ model with respect to the mean observed soil observations, for soil pH in Andhra Pradesh – India, with the size of the dataset used for calibrating the model.
Figure 21. Difference maps with ascending new data samples used for calibrating the model for soil pH in Andhra Pradesh – India.
4 Discussion

This chapter briefly discusses all results, additional clarification is given to conflicting results and unexpected findings. Hereafter the limitations of the study are explained, and the chapter ends by summarising the importance of some results.

4.1 Comparison of a global and local SoilGrids model

Figure 7 and 10 show the SOC and soil pH prediction results of the $S_G$ and $S_L$. For each soil property the prediction results are of $S_G$ deviate from the $S_L$, which resulted in a deviation map with large deviations. Figures 8 and 11 and Tables 9 and 11 imply that the $S_L$ model is a better model than the $S_G$ model to make predictions for the Andhra Pradesh study area. The scatterplots of the $S_L$ model are closer to the 1:1 line, the RMSE, ME are closer to 0 and R-squared is higher for $S_L$. Tifafi et al. (2018) also showed large differences between the globally calibrated SoilGrids predictions and local reference data.

There could be two reasons, or a combination of both options, why $S_L$ predicts better for Andhra Pradesh than $S_G$. Firstly, the $S_L$ model uses other covariates which are better suited for Andhra Pradesh. Secondly, the number of soil observations to predict SOC and soil pH in the $S_G$ model are not representative enough to make accurate predictions for Andhra Pradesh. Even if there are only four soil observations present for Andhra Pradesh in the $S_G$ model, other soil observations with covariates equal to Andhra Pradesh in feature space, might compensate for the lack of local soil observations but they have to be present in the $S_G$ model and have similar covariates.

To analyse if the number of local soil observations causes the difference between $S_G$ and $S_G$, the $S_{G+}$ model results (Table 14 for SOC and Table 15 for soil pH) were compared to the $S_G$ model results (as the $S_{G+}$ model contains the $S_G$ and $S_L$ soil observations and uses the same covariates as the $S_G$ model). Tables 12 and 14 show the statistics for the tree SOC models and Tables 13 and 15 show the statistics for the soil pH models, all at the 5-15 cm depth interval. Those statistics show that for both models, the RMSE, ME and $R^2$ of the $S_{G+}$ model are significantly improved in comparison with the $S_G$ model. Other researches also showed that adding more data to an RF model improves the prediction accuracy (Caubet et al. 2019; Fassnacht et al., 2014). All together, we can conclude that most of the extreme deviation between the $S_G$ and $S_L$ model has to do with the number of local soil samples in the training data. Still, the $S_L$ model predictions are slightly better compared to the $S_{G+}$ model. This difference is the result of global modelling or local modelling. Influences might be from the global soil observations in $S_G$ model or the locally selected covariates from the $S_L$ model, as those are the only two differences between the models. To tell if the locally selected covariates have an influence on the difference between $S_G$ and $S_L$ model, the variable importance for both models are compared.

An interesting finding in the variable importance comparison between the $S_G$ and $S_L$ models is that the variable importance of both models is very different for both soil pH and SOC (Figure 6 for SOC and Figure 9 for soil pH). The variable importance is only meaningful if the model fits the data well (Ando, 2014). Here the models are calibrated at a different scale. The $S_G$ model predicts soil properties for a global scale quite well. However, when we only look at a local scale, the model does not perform very well. Therefore, the variable importance of the $S_G$ model does not tell us much for local scale. The Variable importance of the $S_L$ model, on the other hand, really shows what covariates are important for the local scale. Especially for the soil pH model covariates differ between a globally calibrated model and a locally calibrated model. Research of (Ando, 2014; Bolourchi, Moradi, Demirel, & Uysal, 2018) all show that selecting representative covariates is essential for an accurate prediction of local patterns. An example of a possible local pattern is shown in the $S_L$ prediction map.
of Figure 10 (bottom maps). The northeast part of \(SG_{\text{Local}}\) map shows more deviation in soil pH than the \(SG_{\text{Global}}\) prediction map. This may be explained by the deviation in variable importance between the two models. Based on this information we can also assume that the difference between the \(SG_{\text{Plus}}\) and \(SG_{\text{Local}}\) which was not explained by the new local soil samples is explained by the covariates.

Another interesting observation in Table 8 is that the minimum difference between \(SG_{\text{Global}}\) and \(SG_{\text{Local}}\) model is extremely high, namely \(-7.29\%\) SOC. This while the min and max SOC % are 0.07 and 2.47 for \(SG_{\text{Local}}\) and 0.76 and 9.23 for \(SG_{\text{Global}}\). When looking at the histogram of the occurrence (Figure 21) of prediction values, it is clear that predictions above 5.0 % SOC only occur in a few cases and therefore, do not influence further results.

### 4.2 Influence of new soil observations

One thing I expected is that adding more local data points to a model would improve the accuracy of a random forest model for the local area. The results of RQ1 show that more local points indeed improved the accuracy of the local study area for both soil properties. I also expected that the std study area (Nampula) would show substantial differences and the ltd study area (the Netherlands) would show little to no differences when adding new datapoints. However, the opposite occurred, especially for the SOC models. The std study area showed little to no differences for SOC predictions after adding new local data in India, while the ltd study area showed substantial differences.

This does not necessarily mean that other ltd and std areas also show this deviation. To see other areas also show this deviation future research is needed (see section 5.4) Also unexpected is that the 0 -5 cm and 5-15 cm depth interval layers show more deviation than the 5-10 cm depth interval layer, while this is the depth interval for which new local soil samples where added. The soil pH model showed for both Nampula and the Netherlands visually the same amount of deviation. The mean of all maps stayed the same however. To find out what caused the deviation, the SOC results of the Netherlands were compared to other studies and the variable importance plot was examined.

Because the SOC maps for the Netherlands have the largest differences between the two models, the results were compared to other work. Figure 19 Shows the SOC maps based on two different data sources in % SOC. Even if both maps look different at first sight, there are some similar patterns in the data. All values predicted higher than 11.63% SOC (dark brown) show up in the same areas with similar patterns. And at the centre of the map (the Veluwe) show similar patterns. The Costal area and the south-west of the area both show uniform colours, although the right map predicts lower than 1.74 % SOC and the left map predicts values between 1.74% and 2.91 % SOC.

![Prediction value occurrence for \(SG_{\text{Local}}\) and \(SG_{\text{Global}}\) for SOC.](image)
Figure 24. Estimated SOC (%) for 0 - 30 cm depth layer in the Netherlands according to the SGPlus model (left) and the SGGlobal model (right).

The maps in Figure 23 show the SOC in % for the depth interval layer of 0 – 30. To be able to compare the SGGlobal and SGPlus maps to those maps the weighted mean of the prediction maps for both models was calculated and plotted (Figure 24). When comparing the SGGlobal and SGPlus maps in Figure 24, both maps seem almost similar, while the difference maps show a lot of deviation (Figure 18). This probably happened because the 0-5cm and 5-15cm interval depth maps show mainly positive difference, while the 15-30 cm depth map mainly shows a negative difference, neutralising each other in the weighted mean map of figure 24. When comparing figure 24 to the maps shown in Figure 23, the predicted values higher than 11.63 % SOC seem to follow the same patterns in all four maps. The north-west part of the area shows different prediction values for both maps in figure 23 and the maps in figure 24. Figure 24 show a predicted value between 5.84% SOC and 11.63% SOC while the maps of figure 23 show a predicted value lower than 1.74% SOC and between 2.91 and 5.84% SOC. The south-east part of the area shows for figure 24 a predicted value lower than 1.74% SOC, the lowest predictions of the area. All together, the comparison did not show any explanation why both SOC models deviated so much.
The comparison of the variable importance of $SG_{\text{Plus}}$ and $SG_{\text{Global}}$ resulted in something interesting. Figures 12 and 13 show the variable importance of SOC and soil pH. The $SG_{\text{Plus}}$ showed different variable importance after adding the local soil observations to the model for both soil properties. The soil pH models showed the least deviation in variable importance and in OOB statistics. The deviation between the Soil pH maps was also smaller for all areas. The differences in variable importance probably caused the deviation between the models in Nampula and the Netherlands but do not explain why prediction results in the Netherlands deviated more than in Nampula. To examine why exactly the Netherlands showed higher differences, future research is needed (see 5.1; recommendations).

4.3 local data density

The results of the local data density analyses (Figure 22 and 23 and Table 21) show that adding more local observations increased the deviation between the $SG_{\text{Plus}}$ model and the $SG_{\text{Global}}$ model. From this we can tell that increasingly adding more local soil observations has a tremendous effect on the local predictions. The descriptive statistics of adding more local soil observations is also shown in Table 21, where Figure 23 shows the course of the mean predictions. Even though the minimum and maximum fluids when new soil samples are added, the median, mean and absolute mean difference all move slowly to the observed values. It is interesting to see that the mean prediction course starts off lifting steeper at the beginning of the line and flattens around $n2000$. This confirms our expectations that accuracy typically increases according to a power-law (the more training data, the more accurate the model) and then reaches a plateau from whereon the accuracy of the predictions is just slightly changing. Although accuracy is here measured as the mean and median of the predictions and not as RMSE, ME or $R^2$. It is interesting that even when a large number of new soil observations are added, the statistics in Table 21 do not reach the observed values. When the results of the comparison of a global and a local SoilGrids model (Table 11; the descriptive statistics of the Soil pH models vs Observed values in Andhra Pradesh), especially the $SG_{\text{Local}}$ results, are compared to the $SG_{\text{Plus}}$36898 and observed values, it is interesting to see what the influence of locally selected covariates is. Since $SG_{\text{Local}}$ is calibrated using all local soil observations and uses local covariates, and $SG_{\text{Plus}}$36898 is calibrated also using all local soil observations but uses globally selected covariates, the difference between the two model statistics (mean of 7.37 pH for $SG_{\text{Local}}$, 7.26 pH for $SG_{\text{Plus}}$36898 and 7.38 as observed mean pH) is the influence of locally selected covariates.

4.4 limitations of this study

For this study, it was not tested how many soil observations were already in the $SG_{\text{Global}}$ dataset with a small taxonomic distance to the soil observations in Andhra Pradesh and Nampula. This may influence how extreme the differences between the $SG_{\text{Global}}$ and $SG_{\text{Local}}$ are. The more std soil observations there are in $SG_{\text{Global}}$, the smaller the differences will be between the models.

It was also not tested what the influence of adding clustered data to the SoilGrids model has. Research showed that for other machine learning models, well-distributed sampling methods have a huge influence on the accuracy of the results (Caubet et al., 2019). Even if RF does not take sampling locations into account and is well known for its ability to handle skewed data and small numbers of observations, it might increase the accuracy of the model. To test this, further research is needed.

Because the SoilGrids DSM model was rebuilt, this may have caused some small deviations in the results. The original SoilGrids DSM predictions were compared with the $SG_{\text{Global}}$ results for the 5-15 cm depth layer for SOC and soil pH. It was expected that the results would be almost similar, but there was a large
difference between the two prediction maps. Even when the original trained SoilGrids DSM model was used to make predictions for the Andhra Pradesh study area, there was a deviation between the original and the rebuilt model results. A possible reason for this may be differences in the covariate stack used to make predictions. Due to time limits, this was not investigated. This could possibly influence the prediction results when being compared to other study results, but since the same data were used to build all SoilGrids DSM models, this did not influence the comparison between models.

It might happen that the covariates selected after the correlation analyses and RFE are still correlated. This was not taken into account during the interpretation of the results. If there is still a correlation between covariates, RF will randomly select one of the covariates as important and the other as less important. When two models each pick the other covariate as important, this will result in a differing variable importance plot.

The accuracy of the model results for the local data density research question is now measured in the mean of the prediction map. To gain better insight in the accuracy, it is better to use the RMSE or $R^2$. For the interpretation of this research, the mean was good enough, however the RMSE might show different results.
5 Conclusions and recommendations

This study aims to compare globally and locally calibrated SoilGrids models and assess the influence of adding local data to a globally calibrated SoilGrids model. To achieve this objective, the three research questions are answered in this chapter. This chapter ends with recommendations for future study.

5.1 Comparison of a global and local SoilGrids model

The \(SG_{\text{Global}}\) and the \(SG_{\text{Local}}\) model differ in each aspect considered in the comparison. In the model comparison, map comparison and accuracy, the \(SG_{\text{Local}}\) model performed better than the \(SG_{\text{Global}}\) model. This was due to the local soil samples used to calibrate \(SG_{\text{Local}}\) and the locally performed RFE. Together they caused the changes to the covariates selection and covariate importance for the random forest predictions, which had a direct effect on the accuracy.

5.2 Influence of new soil observations

When the local soil observations were added to the global model, the accuracy of the predictions increased substantially. It was expected that the areas close in covariate feature space would benefit most from local soil observations and the areas distant in feature space would benefit least. Results showed that it was not the case. In this research, the area distant in covariate feature space showed more deviation than the area close in feature space. Future research is needed to find out why the distant area showed such deviation. The OOB model statistics showed that adding the local soil observations had a positive influence on the model accuracy on predictions around the entire globe. To tell if the new local samples also had a positive influence on Nampula and the Netherlands specific, a test dataset is needed for future research.

5.3 Effect of local data density

The effect of adding local data on prediction accuracy first increased steadily as the size of the local data increased and then flattened off. This shows that the first set of new local data samples has the biggest effect on the local prediction accuracy and that adding more local soil observations to the model keeps increasing model the accuracy, but there is a tremendous amount of local soil samples needed to slowly rise the accuracy until it completely flattens out. From this point, only calibration with locally selected covariates can increase the model accuracy.

5.4 Recommendations

SoilGrids implementation

This study shows that it is important to realise that adding new data observations to the globally calibrated SoilGrids model influences model results in areas close in feature space, but also areas distant in feature space, according to the Homosoil principle. The OOB model statistics showed that adding local data did influence the model accuracy all over the globe in a positive way. Therefore, global modelling has its benefits, as soil observations from the entire globe can be used to make predictions for a local area. However, Other researches have shown that at some point having a good covariate selection is more important than having more available soil samples (Fassnacht et al., 2014). To make even more accurate predictions with a SoilGrids model, a combination between locally selected covariates and global soil observations can be used to make predictions for a local area. To find out if this combination is technical feasible, what the best local area size is and how to deal with border areas, future research is needed.
Influence of local data on other areas

The differences in variable importance probably caused the deviation between the models in Nampula and the Netherlands but did not explain why prediction results in the Netherlands deviated more than in Nampula. There was no logical explanation of why adding local data had such a high effect on areas which we expected to be far in feature space and such little effect on areas close in feature space. It might be that the selected areas from the Homosoil script were not as close in feature space as expected. The Homosoil script uses other covariates to select regions (based on Euclidean distance) than the covariates used for modelling. To find out if this was the case, a different selection method should be used to select the std and ltd study areas. The research report of A. Schoneveld, 2020 shows that Euclidean distance is not always the best method to calculate the distance in feature space and that other methods such as the Manhattan may calculate feature space distance better. Besides, could it be useful to introduce more study areas to investigate whether the deviation in the ltd and std study area will occur in different areas with different environmental conditions. For further research, more study areas may give further insights into how new local soil samples influences other areas.

Another possibility to find out why adding local data had such a high effect on predictions in the Netherlands, is analysing which variables have the most effect on the predictive model by performing a sensitivity analyses. To explain the relationship between model variables and predictions .. presents a method to look inside the black box of RF. It computes the contribution of each covariate to the RF model. The GINI variable importance used during this study is, according to Palczewska, 2013 often insufficient for the complete understanding of the relationship between covariates and the predicted value. Kuz’min et al., 2011 propose a new technique to calculate the contribution of a covariate. In this method, feature contribution is computed separately for each prediction and provides detailed information about relationships between variables and the predicted value (Palczewska, Palczewski, Marchese Robinson, & Neagu, 2013).

P. Grover, 2017 presents a methods to find for a given data point and associated prediction, which covariables (or combinations of covariables) explain this specific prediction. They use the treeinterpreter package in Python to show the sorted list of bias (mean of data at starting node) and individual node contributions for a given prediction (Grover, 2017). This local interpretation determines which covariables are used to come to that final prediction. This can be used to find out which covariables contributed most to the predictions with the largest deviation in all study areas. If the covariables used for the predictions in the Netherlands show similarity to the covariables used for predictions in Andhra Pradesh, then it could be an explanation why the predictions in the Netherlands show such high deviation.
6. References


de Sousa, L. (n.d.). How were the spatial predictions generated? Retrieved October 18, 2019, from https://www.isric.org/explore/soilgrids/faq-soilgrids#What_is_SoilGrids


Appendix

I Theoretical background

K-fold Cross-validation

Cross-validation is a resampling procedure used to estimate the accuracy of a machine learning model on unseen data. The parameter called k refers to the number of groups that a data sample will be split into. K-fold cross-validation shuffles the dataset randomly and splits the dataset into k groups (Brownlee, 2019). Each model is re-calibrated 10 times using 90% of the data and predictions derived from the calibrated models are compared with observations of the remaining 10% (de Sousa, n.d.). If k is set to 10, there will be 10 different test and training sets which are used to estimate the model accuracy. The dataset can be divided into k-groups by using folds. The dataset is equally divided and each data value in a fold is assigned the fold number (between 0 and k). The training dataset is used to calibrate the model and make predictions. Those predictions are evaluated on the corresponding test datasets. All the runs together are summarised and represent the accuracy of a model.

Random Forest

SoilGrids maps are produced using automated soil mapping based on the machine learning algorithm Random Forest (RF). RF has a simple yet powerful concept. It consists of a large number of individual decision trees that operate as a group. A simple example of a decision tree is shown in figure 25. Each time the path splits into two is called a node, and the question asked is based on the covariates (e.g. is red? Is probably based on the covariate; number colour and is underlined? On the covariate that tells if a number is underlined or not). At a node observations are split, observations that meet the criteria go down the Yes branch and ones that don’t go down the No branch. At each node the model will ask: What covariate will allow me to split the observations in a way that the ‘Yes’ groups are as different from each other as possible and the members of each resulting subgroup are as similar to each other as possible (Yiu, 2019)? The hyperparameter mtry sets the number of covariables available for splitting at each tree node. The ntree hyperparameter set the number of trees to grow. Larger number of trees produce more stable models and covariate importance estimates, but require more memory and longer run times (Liaw & Wiener, 2018).

![Figure 25. Random Forest decision tree example to predict numbers and their colours (Yiu, 2019).](image)

Each decision tree is trained on a different data subset, where sampling is done with replacement and uses a different covariates subset to reduce correlation between trees. All individual decision trees are
then used to make predictions (Breiman, 2001). The conditional mean of all those predictions is used to get the final RF regression prediction (Meinshausen, 2006). Because there is a low correlation between the trees, trees can produce ensemble predictions that are more accurate than any of the individual predictions. All trees together overcome errors of individual trees.

**correlation analyses**

For this research, it is important to use covariates which also have a low correlation. Reducing the number of covariates used in a model reduces calculation time and processing power needed. Besides, keeping two covariates which are highly correlated, can lead to incorrect conclusions when interpreting the variable importance. When a dataset has two highly correlated feature, RF can pick any of those covariates as important predictor and the other as not important, with no concrete preference for which covariate. If the RF models we want to compare in this study both pick another covariate as main predictor, conclusions based on that variable importance are then incorrect (Ando, 2014).

correlation analyses assessed bi-variate correlations between covariates with the Pearson correlation coefficient. If this coefficient was larger than 0.85, then one of the correlated covariates was excluded randomly.

**RFE**

Variables with high importance have a significant impact on the model prediction values. On the other side, variables with low importance might be omitted from a model, making model calibration and prediction simpler and faster.

Automatic feature selection methods can be was used to identify attributes covariates that are and are not required to build an accurate model to be used. An automatic method for feature selection provided by the caret R package is called Recursive Feature Elimination or RFE. RFE builds \( n \) models with different subsets of a dataset and states the most important variables based on the GINI index (see chapter 2.2 model comparison) (Kuhn, 2009). SoilGrids uses \( n = 4 \) to determine the covariates. This results in four differently selected covariate tables. If a covariate occurs in all those covariate tables, this covariate is selected for the final model. When the amount number of selected covariates for the final model is less than 25% of the input covariable, also the covariates that occurs three times are used in the final model (de Sousa, n.d.).

Table 22 and table 23 show the OOB accuracy of the SG\(_{local}\) model before and after the RFE. This shows that you can remove x variables that are not that significant and have similar or better performance in with much shorter training time.

**Table 22. SOC (%) OOB statistics for the SG\(_{local}\) model.**  **Table 23. soil pH OOB statistics for the SG\(_{local}\) model.**

<table>
<thead>
<tr>
<th></th>
<th>All covariates</th>
<th>RFE covariates</th>
<th></th>
<th>All covariates</th>
<th>RFE covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>0.29</td>
<td>0.29</td>
<td>RMSE</td>
<td>0.48</td>
<td>0.47</td>
</tr>
<tr>
<td>MSE</td>
<td>0.08</td>
<td>0.08</td>
<td>MSE</td>
<td>0.23</td>
<td>0.22</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>0.42</td>
<td>0.42</td>
<td>( R^2 )</td>
<td>0.58</td>
<td>0.59</td>
</tr>
</tbody>
</table>

**OOB error**

Each tree in the Random Forest is trained on a subset of the data, which is sampled with replacement from the original data. This results in around \(~2/3\) of distinct observations in each tree. The out-of-bag error is calculated on all the observations, but for calculating each row’s error, the method only considers trees that have not seen this row during training. This is similar to evaluating the model on a validation set.

Organic Carbon is together with pH, the best simple indicator of the health status of the soil. Moderate to high amounts of organic carbon are associated with fertile soils with a good structure (Maschinen et al., 2012). Soil organic carbon (SOC) is the largest carbon (C) stock in most terrestrial ecosystems, containing approximately 2344 Gt of organic C globally (Davidson & Janssens, 2006). Moreover, soil is recognised as the second largest C pool after the oceans and one of the most important components of the biosphere, delivering major ecosystem services and functions (Ogle & Paustian, 2005). SOC refers only to the carbon component of organic compounds. The main influences of the amount of SOC are soil type, climate and land/soil management. Where for soil, clay binds to organic matter, helping it from being decomposed, while sandy soils are coarse-textured which causes SOC to be rapidly decomposed. For climate SOC increase with rainfall, this is because increasing rainfall supports plant growth. Temperature also plays a part as decreasing temperatures slow the decomposer of SOC. Because SOC mainly exist in the top 0-10 cm of soils, land and soil management can have a big influence of SOC. Deforestation for example lay SOC bare for erosion, and transfers soil down the slope into the lower parts of the landscape, leaving the slope with low SOC content and the lower parts with increased SOC content (Griffin & Edwards, 2019).

The acidity or alkalinity of a substance is measured in pH units, a scale running from 0 to 14. A Soil pH lower than 7 is a high acidity soil, a pH of 7 is neutral and a pH higher than 7 is a high alkalinity soil. Most cultivated plants enjoy slightly acidic conditions with a pH of about 6.5, but some plants grow best under slightly acidic conditions (Mosaic, 2020). There are three main reasons when soil tends to become acidic. First is rainwater. Soils formed under low rainfall conditions tend to be natural as rainwater leaches basic ions away. Soils formed under conditions of high annual rainfall are more acidic than are soils formed under more arid conditions. Second is nitrogen fertilisers. Intensive farming can result in soil acidification over a number of years with nitrogen fertilisers or manures. Last main influencer of soil pH are plants. Decomposing organic matter and root respiration release Carbon dioxide which dissolves in soil water to form a weak organic acid (Bickelhaupt, 2020).
**Covariates**

- **Covariates**
- **Explanation**

**FULL CODE**

**DESCRIPTION**

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLM_CHE_BD02</td>
<td>Mean diurnal range [°C] at 1 km (based on CHLSA climate surfaces).</td>
</tr>
<tr>
<td>CLM_CHE_BD04</td>
<td>Temperature seasonality at 3 km (based on CHLSA climate surfaces).</td>
</tr>
<tr>
<td>CLM_CHE_BD08</td>
<td>Temperature seasonality at 3 km (based on CHLSA climate surfaces).</td>
</tr>
<tr>
<td>CLM_CHE_BD09</td>
<td>Mean Temperature of warmest quarter [°C] at 1 km (based on CHLSA climate surfaces).</td>
</tr>
<tr>
<td>CLM_CHE_BD13</td>
<td>Precipitation of wettest month [mm] at 1 km (based on CHLSA climate surfaces).</td>
</tr>
<tr>
<td>CLM_CHE_BD14T</td>
<td>Precipitation of wettest month [mm] at 1 km (based on CHLSA climate surfaces).</td>
</tr>
<tr>
<td>CLM_CHE_BD30AVG</td>
<td>Mean monthly precipitation at 1 km (based on CHLSA climate surfaces) for January.</td>
</tr>
<tr>
<td>CLM_CHE_BD40AVG</td>
<td>Mean monthly precipitation at 1 km (based on CHLSA climate surfaces) for February.</td>
</tr>
<tr>
<td>CLM_CHE_BD41AVG</td>
<td>Mean monthly precipitation at 1 km (based on CHLSA climate surfaces) for March.</td>
</tr>
<tr>
<td>CLM_CHE_BD42AVG</td>
<td>Mean monthly precipitation at 1 km (based on CHLSA climate surfaces) for April.</td>
</tr>
<tr>
<td>CLM_CHE_BD43AVG</td>
<td>Mean monthly precipitation at 1 km (based on CHLSA climate surfaces) for May.</td>
</tr>
<tr>
<td>CLM_CHE_BD44AVG</td>
<td>Mean monthly precipitation at 1 km (based on CHLSA climate surfaces) for June.</td>
</tr>
<tr>
<td>CLM_CHE_BD45AVG</td>
<td>Mean monthly precipitation at 1 km (based on CHLSA climate surfaces) for July.</td>
</tr>
<tr>
<td>CLM_CHE_BD46AVG</td>
<td>Mean monthly precipitation at 1 km (based on CHLSA climate surfaces) for August.</td>
</tr>
<tr>
<td>CLM_CHE_BD47AVG</td>
<td>Mean monthly precipitation at 1 km (based on CHLSA climate surfaces) for September.</td>
</tr>
<tr>
<td>CLM_CHE_BD48AVG</td>
<td>Mean monthly precipitation at 1 km (based on CHLSA climate surfaces) for October.</td>
</tr>
<tr>
<td>CLM_CHE_BD49AVG</td>
<td>Mean monthly precipitation at 1 km (based on CHLSA climate surfaces) for November.</td>
</tr>
<tr>
<td>CLM_CHE_BD50AVG</td>
<td>Mean monthly precipitation at 1 km (based on CHLSA climate surfaces) for December.</td>
</tr>
<tr>
<td>CLM_CHE_BD51AVG</td>
<td>Total annual precipitation at 1 km (based on CHLSA climate surfaces).</td>
</tr>
<tr>
<td>CLM_CHE_T04MIN</td>
<td>Minimum monthly temperature at 1 km (based on CHLSA Climate) for January.</td>
</tr>
<tr>
<td>CLM_CHE_T04MIN</td>
<td>Minimum monthly temperature at 1 km (based on CHLSA Climate) for February.</td>
</tr>
<tr>
<td>CLM_CHE_T04MIN</td>
<td>Minimum monthly temperature at 1 km (based on CHLSA Climate) for March.</td>
</tr>
<tr>
<td>CLM_CHE_T04MIN</td>
<td>Minimum monthly temperature at 1 km (based on CHLSA Climate) for April.</td>
</tr>
<tr>
<td>CLM_CHE_T05MIN</td>
<td>Minimum monthly temperature at 1 km (based on CHLSA Climate) for May.</td>
</tr>
<tr>
<td>CLM_CHE_T06MIN</td>
<td>Minimum monthly temperature at 1 km (based on CHLSA Climate) for June.</td>
</tr>
<tr>
<td>CLM_CHE_T07MIN</td>
<td>Minimum monthly temperature at 1 km (based on CHLSA Climate) for July.</td>
</tr>
<tr>
<td>CLM_CHE_T08MIN</td>
<td>Minimum monthly temperature at 1 km (based on CHLSA Climate) for August.</td>
</tr>
<tr>
<td>CLM_CHE_T09MIN</td>
<td>Minimum monthly temperature at 1 km (based on CHLSA Climate) for September.</td>
</tr>
<tr>
<td>CLM_CHE_T10MIN</td>
<td>Minimum monthly temperature at 1 km (based on CHLSA Climate) for October.</td>
</tr>
<tr>
<td>CLM_CHE_T11MIN</td>
<td>Minimum monthly temperature at 1 km (based on CHLSA Climate) for November.</td>
</tr>
<tr>
<td>CLM_CHE_T12MIN</td>
<td>Minimum monthly temperature at 1 km (based on CHLSA Climate) for December.</td>
</tr>
<tr>
<td>CLM_CHE_TYRMIN</td>
<td>Yearly averages of minimum monthly temperatures at 1 km (based on CHLSA Climate).</td>
</tr>
<tr>
<td>CLM_ESA_SG0AVG</td>
<td>Mean monthly snowfall prob. at 500 m (based on MODIS snow product) for January.</td>
</tr>
<tr>
<td>CLM_ESA_SG0AVG</td>
<td>Mean monthly snowfall prob. at 500 m (based on MODIS snow product) for February.</td>
</tr>
<tr>
<td>CLM_ESA_SG0AVG</td>
<td>Mean monthly snowfall prob. at 500 m (based on MODIS snow product) for March.</td>
</tr>
<tr>
<td>CLM_ESA_SG0AVG</td>
<td>Mean monthly snowfall prob. at 500 m (based on MODIS snow product) for April.</td>
</tr>
<tr>
<td>CLM_ESA_SG0AVG</td>
<td>Mean monthly snowfall prob. at 500 m (based on MODIS snow product) for May.</td>
</tr>
<tr>
<td>CLM_ESA_SG0AVG</td>
<td>Mean monthly snowfall prob. at 500 m (based on MODIS snow product) for June.</td>
</tr>
<tr>
<td>CLM_ESA_SG0AVG</td>
<td>Mean monthly snowfall prob. at 500 m (based on MODIS snow product) for July.</td>
</tr>
<tr>
<td>CLM_ESA_SG0AVG</td>
<td>Mean monthly snowfall prob. at 500 m (based on MODIS snow product) for August.</td>
</tr>
<tr>
<td>CLM_ESA_SG0AVG</td>
<td>Mean monthly snowfall prob. at 500 m (based on MODIS snow product) for September.</td>
</tr>
<tr>
<td>CLM_ESA_SG0AVG</td>
<td>Mean monthly snowfall prob. at 500 m (based on MODIS snow product) for October.</td>
</tr>
<tr>
<td>CLM_ESA_SG0AVG</td>
<td>Mean monthly snowfall prob. at 500 m (based on MODIS snow product) for November.</td>
</tr>
<tr>
<td>CLM_ESA_SG0AVG</td>
<td>Mean monthly snowfall prob. at 500 m (based on MODIS snow product) for December.</td>
</tr>
<tr>
<td>CLM_ESA_SG0AVG</td>
<td>Mean yearly snowfall prob. at 500 m (based on MODIS snow product).</td>
</tr>
<tr>
<td>CLM_ESA_SG0AVG</td>
<td>Mean yearly snowfall prob. at 500 m (based on MODIS snow product).</td>
</tr>
<tr>
<td>CLM_MOD_C00AVG</td>
<td>Long-term averaged monthly cloud cover Jan.</td>
</tr>
<tr>
<td>CLM_MOD_C00AVG</td>
<td>Long-term averaged monthly cloud cover Feb.</td>
</tr>
<tr>
<td>CLM_MOD_C00AVG</td>
<td>Long-term averaged monthly cloud cover Mar.</td>
</tr>
<tr>
<td>CLM_MOD_C00AVG</td>
<td>Long-term averaged monthly cloud cover Apr.</td>
</tr>
<tr>
<td>CLM_MOD_C00AVG</td>
<td>Long-term averaged monthly cloud cover May.</td>
</tr>
<tr>
<td>CLM_MOD_C00AVG</td>
<td>Long-term averaged monthly cloud cover Jun.</td>
</tr>
<tr>
<td>CLM_MOD_C00AVG</td>
<td>Long-term averaged monthly cloud cover Jul.</td>
</tr>
<tr>
<td>CLM_MOD_C00AVG</td>
<td>Long-term averaged monthly cloud cover Aug.</td>
</tr>
<tr>
<td>CLM_MOD_C00AVG</td>
<td>Long-term averaged monthly cloud cover Sep.</td>
</tr>
<tr>
<td>CLM_MOD_C00AVG</td>
<td>Long-term averaged monthly cloud cover Oct.</td>
</tr>
<tr>
<td>CLM_MOD_C00AVG</td>
<td>Long-term averaged monthly cloud cover Nov.</td>
</tr>
<tr>
<td>CLM_MOD_C00AVG</td>
<td>Long-term averaged monthly cloud cover Dec.</td>
</tr>
<tr>
<td>CLM_MOD_C00AVG</td>
<td>Long-term averaged monthly cloud cover Jan.</td>
</tr>
<tr>
<td>CLM_MOD_C00AVG</td>
<td>Long-term averaged monthly cloud cover Feb.</td>
</tr>
<tr>
<td>CLM_MOD_C00AVG</td>
<td>Long-term averaged monthly cloud cover Mar.</td>
</tr>
<tr>
<td>CLM_MOD_C00AVG</td>
<td>Long-term averaged monthly cloud cover Apr.</td>
</tr>
<tr>
<td>CLM_MOD_C00AVG</td>
<td>Long-term averaged monthly cloud cover May.</td>
</tr>
<tr>
<td>CLM_MOD_C00AVG</td>
<td>Long-term averaged monthly cloud cover Jun.</td>
</tr>
<tr>
<td>CLM_MOD_C00AVG</td>
<td>Long-term averaged monthly cloud cover Jul.</td>
</tr>
<tr>
<td>CLM_MOD_C00AVG</td>
<td>Long-term averaged monthly cloud cover Aug.</td>
</tr>
<tr>
<td>CLM_MOD_C00AVG</td>
<td>Long-term averaged monthly cloud cover Sep.</td>
</tr>
<tr>
<td>CLM_MOD_C00AVG</td>
<td>Long-term averaged monthly cloud cover Oct.</td>
</tr>
<tr>
<td>CLM_MOD_C00AVG</td>
<td>Long-term averaged monthly cloud cover Nov.</td>
</tr>
<tr>
<td>CLM_MOD_C00AVG</td>
<td>Long-term averaged monthly cloud cover Dec.</td>
</tr>
<tr>
<td>CLM_MOD_C00AVG</td>
<td>Cloud cover seasonality concentration.</td>
</tr>
</tbody>
</table>
CLM_MOD_LST01AVG Long-term averaged mean monthly surface temperature (nighttime) MODIS January. Derived using a stack of MOD09L1A2 LST images.

CLM_MOD_LST02AVG Long-term averaged mean monthly surface temperature (nighttime) MODIS February. Derived using a stack of MOD09L1A2 LST images.

CLM_MOD_LST03AVG Long-term averaged mean monthly surface temperature (nighttime) MODIS March. Derived using a stack of MOD09L1A2 LST images.

CLM_MOD_LST04AVG Long-term averaged mean monthly surface temperature (nighttime) MODIS April. Derived using a stack of MOD09L1A2 LST images.

CLM_MOD_LST05AVG Long-term averaged mean monthly surface temperature (nighttime) MODIS May. Derived using a stack of MOD09L1A2 LST images.

CLM_MOD_LST06AVG Long-term averaged mean monthly surface temperature (nighttime) MODIS June. Derived using a stack of MOD09L1A2 LST images.

CLM_MOD_LST07AVG Long-term averaged mean monthly surface temperature (nighttime) MODIS July. Derived using a stack of MOD09L1A2 LST images.

CLM_MOD_LST08AVG Long-term averaged mean monthly surface temperature (nighttime) MODIS August. Derived using a stack of MOD09L1A2 LST images.

CLM_MOD_LST09AVG Long-term averaged mean monthly surface temperature (nighttime) MODIS September. Derived using a stack of MOD09L1A2 LST images.

CLM_MOD_LST10AVG Long-term averaged mean monthly surface temperature (nighttime) MODIS October. Derived using a stack of MOD09L1A2 LST images.

CLM_MOD_LST11AVG Long-term averaged mean monthly surface temperature (nighttime) MODIS November. Derived using a stack of MOD09L1A2 LST images.

CLM_MOD_LST12AVG Long-term averaged mean monthly surface temperature (nighttime) MODIS December. Derived using a stack of MOD09L1A2 LST images.

CLM_MOD_LST01MAX Long-term maximum mean monthly surface temperature (nighttime) MODIS January. Derived using a stack of MOD09L1A2 LST images.

CLM_MOD_LST02MAX Long-term maximum mean monthly surface temperature (nighttime) MODIS February. Derived using a stack of MOD09L1A2 LST images.

CLM_MOD_LST03MAX Long-term maximum mean monthly surface temperature (nighttime) MODIS March. Derived using a stack of MOD09L1A2 LST images.

CLM_MOD_LST04MAX Long-term maximum mean monthly surface temperature (nighttime) MODIS April. Derived using a stack of MOD09L1A2 LST images.

CLM_MOD_LST05MAX Long-term maximum mean monthly surface temperature (nighttime) MODIS May. Derived using a stack of MOD09L1A2 LST images.

CLM_MOD_LST06MAX Long-term maximum mean monthly surface temperature (nighttime) MODIS June. Derived using a stack of MOD09L1A2 LST images.

CLM_MOD_LST07MAX Long-term maximum mean monthly surface temperature (nighttime) MODIS July. Derived using a stack of MOD09L1A2 LST images.

CLM_MOD_LST08MAX Long-term maximum mean monthly surface temperature (nighttime) MODIS August. Derived using a stack of MOD09L1A2 LST images.

CLM_MOD_LST09MAX Long-term maximum mean monthly surface temperature (nighttime) MODIS September. Derived using a stack of MOD09L1A2 LST images.

CLM_MOD_LST10MAX Long-term maximum mean monthly surface temperature (nighttime) MODIS October. Derived using a stack of MOD09L1A2 LST images.

CLM_MOD_LST11MAX Long-term maximum mean monthly surface temperature (nighttime) MODIS November. Derived using a stack of MOD09L1A2 LST images.

CLM_MOD_LST12MAX Long-term maximum mean monthly surface temperature (nighttime) MODIS December. Derived using a stack of MOD09L1A2 LST images.

CLM_MOD_PWAVG Long-term averaged mean annual precipitation Water Vapor in cm for months January and February. Derived using a stack of MOD09G_L2 monthly images from NED.

CLM_MOD_PW01 Long-term averaged mean monthly MODIS Precipitable Water Vapor in cm for months March and April. Derived using a stack of MOD09G_L2 monthly images from NED.

CLM_MOD_PW02 Long-term averaged mean monthly MODIS Precipitable Water Vapor in cm for months May and June. Derived using a stack of MOD09G_L2 monthly images from NED.

CLM_MOD_PW03 Long-term averaged mean monthly MODIS Precipitable Water Vapor in cm for months July and August. Derived using a stack of MOD09G_L2 monthly images from NED.

CLM_MOD_PW04 Long-term averaged mean monthly MODIS Precipitable Water Vapor in cm for months September and October. Derived using a stack of MOD09G_L2 monthly images from NED.

CLM_MOD_PW05 Long-term averaged mean monthly MODIS Precipitable Water Vapor in cm for months November and December. Derived using a stack of MOD09G_L2 monthly images from NED.

CLM_MOD_PWAVG Long-term averaged mean annual precipitation Water Vapor in cm for months January and February. Derived using a stack of MOD09G_L2 monthly images from NED.

CLM_WCI_BI001 BI01 = Annual Mean Temperature

CLM_WCI_BI002 BI02 = Mean Diurnal Range (Mean of monthly (max temp - min temp))

CLM_WCI_BI003 BI03 = Isothermality (BI02/BI01)*100

CLM_WCI_BI004 BI04 = Temperature Seasonality (standard deviation *100)

CLM_WCI_BI005 BI05 = Maximum Temperature of Warmest Month

CLM_WCI_BI006 BI06 = Min Temperature of Coldest Month

CLM_WCI_BI007 BI07 = Temperature Annual Range (BI06- BI05)

CLM_WCI_BI008 BI08 = Mean Temperature of Wettest Quarter

CLM_WCI_BI009 BI09 = Mean Temperature of Driest Quarter

CLM_WCI_BI010 BI10 = Mean Temperature of Warmest Quarter

CLM_WCI_BI011 BI11 = Mean Temperature of Coldest Quarter

CLM_WCI_BI012 BI12 = Annual Precipitation

CLM_WCI_BI013 BI13 = Precipitation of Wettest Month

CLM_WCI_BI014 BI14 = Precipitation of Driest Month

CLM_WCI_BI015 BI15 = Precipitation Seasonality (Coefficient of Variation)

CLM_WCI_BI016 BI16 = Precipitation of Wettest Quarter

CLM_WCI_BI017 BI17 = Precipitation of Driest Quarter

CLM_WCI_BI018 BI18 = Precipitation of Warmest Quarter

CLM_WCI_BI019 BI19 = Precipitation of Coldest Quarter

CLM_WCI_P02TOT Total monthly precipitation at 1 km (based on WorldClim v2 climate) for January.

CLM_WCI_P02TOT Total monthly precipitation at 1 km (based on WorldClim v2 climate) for February.

CLM_WCI_P03TOT Total monthly precipitation at 1 km (based on WorldClim v2 climate) for March.

CLM_WCI_P04TOT Total monthly precipitation at 1 km (based on WorldClim v2 climate) for April.

CLM_WCI_P05TOT Total monthly precipitation at 1 km (based on WorldClim v2 climate) for May.

CLM_WCI_P06TOT Total monthly precipitation at 1 km (based on WorldClim v2 climate) for June.

CLM_WCI_P07TOT Total monthly precipitation at 1 km (based on WorldClim v2 climate) for July.

CLM_WCI_P08TOT Total monthly precipitation at 1 km (based on WorldClim v2 climate) for August.

CLM_WCI_P09TOT Total monthly precipitation at 1 km (based on WorldClim v2 climate) for September.

CLM_WCI_P10TOT Total monthly precipitation at 1 km (based on WorldClim v2 climate) for October.

CLM_WCI_P11TOT Total monthly precipitation at 1 km (based on WorldClim v2 climate) for November.

CLM_WCI_P12TOT Total monthly precipitation at 1 km (based on WorldClim v2 climate) for December.

CLM_WCI_S01RAD Solar radiation at 1 km (based on WorldClim v2 climate) for January.

CLM_WCI_S02RAD Solar radiation at 1 km (based on WorldClim v2 climate) for February.

CLM_WCI_S03RAD Solar radiation at 1 km (based on WorldClim v2 climate) for March.

CLM_WCI_S04RAD Solar radiation at 1 km (based on WorldClim v2 climate) for April.

CLM_WCI_S05RAD Solar radiation at 1 km (based on WorldClim v2 climate) for May.

CLM_WCI_S06RAD Solar radiation at 1 km (based on WorldClim v2 climate) for June.

CLM_WCI_S07RAD Solar radiation at 1 km (based on WorldClim v2 climate) for July.

CLM_WCI_S08RAD Solar radiation at 1 km (based on WorldClim v2 climate) for August.

CLM_WCI_S09RAD Solar radiation at 1 km (based on WorldClim v2 climate) for September.

CLM_WCI_S10RAD Solar radiation at 1 km (based on WorldClim v2 climate) for October.

CLM_WCI_S11RAD Solar radiation at 1 km (based on WorldClim v2 climate) for November.

CLM_WCI_S12RAD Solar radiation at 1 km (based on WorldClim v2 climate) for December.

CLM_WCI_SR01AVG Average Yearly Solar radiation at 1 km (based on WorldClim v2 climate).

CLM_WCI_SR02AVG Average Yearly Solar radiation at 1 km (based on WorldClim v2 climate).

CLM_WCI_SR03AVG Average Yearly Solar radiation at 1 km (based on WorldClim v2 climate).
CLM_WCL_T04MAX Maximum monthly temperature at 1 km (based on WorldClim v2 Climate) for April.
CLM_WCL_T04MIN Minimum monthly temperature at 1 km (based on WorldClim v2 Climate) for April.
CLM_WCL_T05AVG Average monthly temperature at 1 km (based on WorldClim v2 Climate) for May.
CLM_WCL_T05MAX Maximum monthly temperature at 1 km (based on WorldClim v2 Climate) for May.
CLM_WCL_T05MIN Minimum monthly temperature at 1 km (based on WorldClim v2 Climate) for May.
CLM_WCL_T06AVG Average monthly temperature at 1 km (based on WorldClim v2 Climate) for June.
CLM_WCL_T06MAX Maximum monthly temperature at 1 km (based on WorldClim v2 Climate) for June.
CLM_WCL_T06MIN Minimum monthly temperature at 1 km (based on WorldClim v2 Climate) for June.
CLM_WCL_T07AVG Average monthly temperature at 1 km (based on WorldClim v2 Climate) for July.
CLM_WCL_T07MAX Maximum monthly temperature at 1 km (based on WorldClim v2 Climate) for July.
CLM_WCL_T07MIN Minimum monthly temperature at 1 km (based on WorldClim v2 Climate) for July.
CLM_WCL_T08AVG Average monthly temperature at 1 km (based on WorldClim v2 Climate) for August.
CLM_WCL_T08MAX Maximum monthly temperature at 1 km (based on WorldClim v2 Climate) for August.
CLM_WCL_T08MIN Minimum monthly temperature at 1 km (based on WorldClim v2 Climate) for August.
CLM_WCL_T09AVG Average monthly temperature at 1 km (based on WorldClim v2 Climate) for September.
CLM_WCL_T09MAX Maximum monthly temperature at 1 km (based on WorldClim v2 Climate) for September.
CLM_WCL_T10AVG Average monthly temperature at 1 km (based on WorldClim v2 Climate) for October.
CLM_WCL_T10MIN Minimum monthly temperature at 1 km (based on WorldClim v2 Climate) for October.
CLM_WCL_T11AVG Average monthly temperature at 1 km (based on WorldClim v2 Climate) for November.
CLM_WCL_T11MAX Maximum monthly temperature at 1 km (based on WorldClim v2 Climate) for November.
CLM_WCL_T11MIN Minimum monthly temperature at 1 km (based on WorldClim v2 Climate) for November.
CLM_WCL_T12AVG Average monthly temperature at 1 km (based on WorldClim v2 Climate) for December.
CLM_WCL_T12MAX Maximum monthly temperature at 1 km (based on WorldClim v2 Climate) for December.
CLM_WCL_T12MIN Minimum monthly temperature at 1 km (based on WorldClim v2 Climate) for December.

CLM_V01APP water vapor pressure at 1 km (based on WorldClim v2 Climate) for January.
CLM_V02APP water vapor pressure at 1 km (based on WorldClim v2 Climate) for February.
CLM_V03APP water vapor pressure at 1 km (based on WorldClim v2 Climate) for March.
CLM_V04APP water vapor pressure at 1 km (based on WorldClim v2 Climate) for April.
CLM_V05APP water vapor pressure at 1 km (based on WorldClim v2 Climate) for May.
CLM_V06APP water vapor pressure at 1 km (based on WorldClim v2 Climate) for June.
CLM_V07APP water vapor pressure at 1 km (based on WorldClim v2 Climate) for July.
CLM_V08APP water vapor pressure at 1 km (based on WorldClim v2 Climate) for August.
CLM_V09APP water vapor pressure at 1 km (based on WorldClim v2 Climate) for September.
CLM_V10APP water vapor pressure at 1 km (based on WorldClim v2 Climate) for October.
CLM_V11APP water vapor pressure at 1 km (based on WorldClim v2 Climate) for November.
CLM_V12APP water vapor pressure at 1 km (based on WorldClim v2 Climate) for December.

CLM_W03SPD Wind speed at 1 km (based on WorldClim v2 Climate) for January.
CLM_W02SPD Wind speed at 1 km (based on WorldClim v2 Climate) for February.
CLM_W03SPD Wind speed at 1 km (based on WorldClim v2 Climate) for March.
CLM_W04SPD Wind speed at 1 km (based on WorldClim v2 Climate) for April.
CLM_W05SPD Wind speed at 1 km (based on WorldClim v2 Climate) for May.
CLM_W06SPD Wind speed at 1 km (based on WorldClim v2 Climate) for June.
CLM_W07SPD Wind speed at 1 km (based on WorldClim v2 Climate) for July.
CLM_W08SPD Wind speed at 1 km (based on WorldClim v2 Climate) for August.
CLM_W09SPD Wind speed at 1 km (based on WorldClim v2 Climate) for September.
CLM_W10SPD Wind speed at 1 km (based on WorldClim v2 Climate) for October.
CLM_W11SPD Wind speed at 1 km (based on WorldClim v2 Climate) for November.
CLM_W12SPD Wind speed at 1 km (based on WorldClim v2 Climate) for December.

ECO_USG_Bioc12 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_201 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_202 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_203 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_204 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_205 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_206 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_207 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_208 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_209 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_210 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_211 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_212 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_213 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_214 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_215 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_216 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_217 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_218 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_219 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_220 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_221 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_222 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_223 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_224 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_225 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_226 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_227 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_228 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_229 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_230 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_231 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_232 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_233 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_234 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_235 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_236 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_237 Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_238  Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
ECO_USG_239  Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
GEO_GLM_01  Average soil and sedimentary deposit thickness in meters. Estimated using physical landform models.
GEO_GLM_010  based on the global lithology map
GEO_GLM_012  based on the global lithology map
GEO_GLM_014  based on the global lithology map
GEO_GLM_016  based on the global lithology map
GEO_GLM_018  based on the global lithology map
GEO_GLM_020  based on the global lithology map
GEO_GLM_022  based on the global lithology map
GEO_GLM_024  based on the global lithology map
GEO_GLM_026  based on the global lithology map
GEO_GLM_028  based on the global lithology map
GEO_GLM_030  based on the global lithology map
GEO_GLM_032  based on the global lithology map
GEO_GLM_034  based on the global lithology map
GEO_GLM_036  based on the global lithology map
GEO_GLM_038  based on the global lithology map
GEO_GLM_040  based on the global lithology map
GEO_GLM_042  based on the global lithology map
GEO_GLM_044  based on the global lithology map
GEO_GLM_046  based on the global lithology map
GEO_GLM_048  based on the global lithology map
GEO_GLM_050  based on the global lithology map
GEO_GLM_052  based on the global lithology map
GEO_GLM_054  based on the global lithology map
GEO_GLM_056  based on the global lithology map
GEO_GLM_058  based on the global lithology map
GEO_GLM_060  based on the global lithology map
GEO_GLM_062  based on the global lithology map
GEO_GLM_064  based on the global lithology map
GEO_GLM_066  based on the global lithology map
GEO_GLM_068  based on the global lithology map
GEO_GLM_070  based on the global lithology map
GEO_GLM_072  based on the global lithology map
GEO_GLM_074  based on the global lithology map
GEO_GLM_076  based on the global lithology map
GEO_GLM_078  based on the global lithology map
GEO_GLM_080  based on the global lithology map
GEO_GLM_082  based on the global lithology map
GEO_GLM_084  based on the global lithology map
GEO_GLM_086  based on the global lithology map
GEO_GLM_088  based on the global lithology map
GEO_GLM_090  based on the global lithology map
GEO_GLM_092  based on the global lithology map
GEO_GLM_094  based on the global lithology map
GEO_GLM_096  based on the global lithology map
GEO_GLM_098  based on the global lithology map
GEO_GLM_100  based on the global lithology map
GEO_GLM_102  based on the global lithology map
GEO_GLM_104  based on the global lithology map
GEO_GLM_106  based on the global lithology map
GEO_GLM_108  based on the global lithology map
GEO_GLM_110  based on the global lithology map
GEO_GLM_112  based on the global lithology map
GEO_GLM_114  based on the global lithology map
GEO_GLM_116  based on the global lithology map
GEO_USG_GUA  Bioclimatic zones based on the USGS's A New Map of Global Ecological Land Units.
LUC_LCE_01  ESA land cover map 2010
LUC_LCE_05  ESA land cover map 2015
LUC_LCE_10  ESA land cover map 2010 class 1000
LUC_LCE_15  ESA land cover map 2010 class 1000
LUC_LCE_12  ESA land cover map 2010 class 1011
LUC_LCE_14  ESA land cover map 2010 class 1012
LUC_LCE_16  ESA land cover map 2010 class 1020
LUC_LCE_18  ESA land cover map 2010 class 1030
LUC_LCE_20  ESA land cover map 2010 class 1040
LUC_LCE_22  ESA land cover map 2010 class 1050
LUC_LCE_24  ESA land cover map 2010 class 1060
LUC_LCE_26  ESA land cover map 2010 class 1061
LUC_LCE_28  ESA land cover map 2010 class 1062
LUC_LCE_29  ESA land cover map 2010 class 1070
LUC_LCE_30  ESA land cover map 2010 class 1071
LUC_LCE_32  ESA land cover map 2010 class 1072
LUC_LCE_34  ESA land cover map 2010 class 1080
LUC_LCE_36  ESA land cover map 2010 class 1081
LUC_LCE_38  ESA land cover map 2010 class 1082
LUC_LCE_40  ESA land cover map 2010 class 1090
LUC_LCE_42  ESA land cover map 2010 class 1100
LUC_LCE_44  ESA land cover map 2010 class 1110
LUC_LCE_46  ESA land cover map 2010 class 1120
LUC_LCE_48  ESA land cover map 2010 class 1130
LUC_LCE_50  ESA land cover map 2010 class 1140
LUC_LCE_52  ESA land cover map 2010 class 1150
LUC_LCE_54  ESA land cover map 2010 class 1160
LUC_LCE_56  ESA land cover map 2010 class 1170
LUC_LCE_58  ESA land cover map 2010 class 1180
LUC_LCE_60  ESA land cover map 2010 class 1190
LUC_LCE_62  ESA land cover map 2010 class 1200
LUC_LCE_64  ESA land cover map 2010 class 1201
LUC_LCE_66  ESA land cover map 2010 class 1202
LUC_LCE_68  ESA land cover map 2010 class 1220
LUC_GFC_BARYL10  Global 30m Bare Ground (circa 2010)
LUC_GFC_TREY10  Global 30m Tree Cover (circa 2010)
LUC_GLC_C01  Cultivated land cover for year 2010 based on GloCover30
LUC_GLC_C02  Forests cover for year 2010 based on GloCover30
LUC_GLC_C03  Grasslands cover for year 2010 based on GloCover30
LUC_GLC_C04  Shrublands cover for year 2010 based on GloCover30
LUC_GLC_C05  Wetland cover for year 2010 based on GloCover30
LUC_GLC_C07  Tundra cover for year 2010 based on GloCover30
LUC_GLC_C08  Artificial Surfaces cover for year 2010 based on GloCover30
LUC_GLC_C09  Bareland cover for year 2010 based on GloCover30
LUC_GLC_C011  Binary: Cultivated land cover for year 2010 based on GloCover30
LUC_GLC_C021  Binary: Forests cover for year 2010 based on GloCover30
LUC_GLC_C031  Binary: Grasslands cover for year 2010 based on GloCover30
LUC_GLC_C041  Binary: Shrublands cover for year 2010 based on GloCover30
LUC_GLC_C051  Binary: Wetland cover for year 2010 based on GloCover30
LUC_GLC_C071  Binary: Tundra cover for year 2010 based on GloCover30
LUC_GLC_C081  Binary: Artificial Surfaces cover for year 2010 based on GloCover30
LUC_GLC_C091  Binary: Bareland cover for year 2010 based on GloCover30
LUC_LDS_MNG  Landsat-based estimated distribution of Mangroves published in Giri et al. (2011)
MOR_ENV_DMEM  DEM based on 100 m resolution from EarthEnv-DEMR. No additional post-processing has been applied to remove vegetation and filter noise.
MOR_ENV_DMEM  DEM merged elevation based on 300 m resolution DEMs from ViewFinderPanoramas SRTM DEM, SRTMGL3, and GMTED2010.
MOR_MRG_CBD  Local downslope Curvature based on DEMMR5G derived in SAGA GIS.
MOR_MRG_CBD  Local upslope Curvature based on DEMMR5G derived in SAGA GIS.
MOR_MRG_CRV  Downslope Curvature based on DEMMR5G derived in SAGA GIS.
MOR_MRG_DVM  Deviation from Mean Value (surface roughness) based on DEMMR5G. Derived in SAGA GIS using a 13 by 13 search radius.
MOR_MRG_DVM  Deviation from Mean Value (surface roughness) based on DEMMR5G. Derived in SAGA GIS using a 9 by 9 search radius.
MOR_MRG_MNN  Melton Ruggedness Number derived in SAGA GIS.
MOR_MRG_NEG  Negative Topographic Openness based on DEMMR5G
MOR_MRG_POS  Positive Topographic Openness based on DEMMR5G
MOR_MRG_TPI  Terrain slope based on DEMMR5G derived in SAGA GIS and expressed in radians x 100.
MOR_MRG_TPI  Topographic Position Index is the difference to the mean calculation (residual analysis) proposed by Wilson & Gallant (2000).
SAGA Wetness Index based on DEM/MIRGS. SAGA TWI is based on a modified catchment area calculation, which does not think of the flow as very thin film.

As a result it predicts for cells situated in valley floors with a small vertical distance to a channel a more realistic, higher potential soil moisture compared to the standard TWI calculation.

Multiresolution Index of Valley Bottom Flatness (MIRGVBF) based on DEM/MIRGS. Derived in SAGA GIS at 500 m, then downscaled to 250 m. Computationally very intensive operation.

Valley depth based on DEM/MIRGS i.e. vertical distance to a channel network base level derived in SAGA GIS.

Landform class: Breaks/Foothills

Landform class: Flat Plains

Landform class: High Mountains/Deep Canyons

Landform class: Hills

Landform class: Low Hills

Landform class: Low Mountains

Landform class: Smooth Plains

LandBand 3 (red) for year 2000

LandBand 3 (red) for year 2014

LandBand 4 (NIR) for year 2000

LandBand 4 (NIR) for year 2014

LandBand 5 (SWIR) for year 2000

LandBand 5 (SWIR) for year 2014

LandBand 7 (SWIR) for year 2000

LandBand 7 (SWIR) for year 2014

Mean yearly MODIS NIR band 4

Mean yearly MODIS MIR band 4

Mean yearly MODIS NIR band 4

Mean yearly MODIS MIR band 4

Normalized Difference Water Index in 2010 computed with NIR,00 and SWL,00

Normalized Difference Water Index in 2014 computed with NIR,00 and SWL,14

Normalized Difference Water Index in 2010 computed with NIR,00 and SWL,00

Normalized Difference Water Index in 2014 computed with NIR,00 and SWL,14

Normalized Difference Water Index in 2010 computed with NIR,00 and SWL,00

Normalized Difference Water Index in 2014 computed with NIR,00 and SWL,14

Normalized Difference Water Index in 2010 computed with NIR,00 and SWL,00

Normalized Difference Water Index in 2014 computed with NIR,00 and SWL,14

Normalized Difference Water Index in 2010 computed with NIR,00 and SWL,00

Normalized Difference Water Index in 2014 computed with NIR,00 and SWL,14

Long-term averaged mean monthly MODIS Enhanced Vegetation Index (EV1) for months January and February. Derived using a stack of MODIS EV1 images.

Long-term averaged mean monthly MODIS Enhanced Vegetation Index (EV1) for months March and April. Derived using a stack of MODIS EV1 images.

Long-term averaged mean monthly MODIS Enhanced Vegetation Index (EV1) for months May and June. Derived using a stack of MODIS EV1 images.

Long-term averaged mean monthly MODIS Enhanced Vegetation Index (EV1) for months July and August. Derived using a stack of MODIS EV1 images.

Long-term averaged mean monthly MODIS Enhanced Vegetation Index (EV1) for months September and October. Derived using a stack of MODIS EV1 images.

Long-term averaged mean monthly MODIS Enhanced Vegetation Index (EV1) for months November and December. Derived using a stack of MODIS EV1 images.

EVI

Max. Dominance of EVI combinations between adjacent pixels

Range of EVI

Minimum mean MODIS EVI

S0 yearly MODIS EVI

Net Primary Productivity (2000-2015 average)

Net Primary Productivity in 2015

Net Primary Productivity in 2000

Global Inundation Extent from Multi-Satellites - Downscaled to 15 arc-seconds

Global Inundation Extent from Multi-Satellites - Downscaled to 15 arc-seconds

Global Inundation Extent from Multi-Satellites - Downscaled to 15 arc-seconds

Global Inundation Extent from Multi-Satellites - Downscaled to 15 arc-seconds

Global Inundation Extent from Multi-Satellites - Downscaled to 15 arc-seconds

Global Inundation Extent from Multi-Satellites - Downscaled to 15 arc-seconds

Surface water change

Global surface water maximum extent

Occurrence probability

Global Water Table Depth in meters based on Fan and Miguez-Macho (2015)
III Homosoil results

Similar regions as Andhra Pradesh, India

Similarity

High

Low