JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2020MS002242

# Global Prediction of Soil Saturated Hydraulic Conductivity Using Random Forest in a Covariate-Based GeoTransfer Function (CoGTF) Framework

**Surya Gupta[1], Peter Lehmann[1], Sara Bonetti[2,4], Andreas Papritz[1], and Dani Or[1,3]**

[1]Department of Environmental Systems Science, Soil and Terrestrial Environmental Physics, ETH Zürich, Zürich, Switzerland, [2]Bartlett School of Environment, Institute for Sustainable Resources, Energy and Resources, University College London, London, UK, [3]Division of Hydrologic Sciences, Desert Research Institute, Reno, NV, USA, [4]Soil Physics and Land Management Group, Wageningen University, Wageningen, The Netherlands

**Key Points:**

- Climate, vegetation, and terrain affect spatial patterns of saturated hydraulic conductivity (Ksat)
- The effect of these environmental covariates on Ksat is quantified using remote sensing data and machine learning
- We introduce Covariate-based GeoTransfer Functions to improve Ksat predictions based on pedotransfer functions

**Supporting Information:**

Supporting Information may be found in the online version of this article.

**Correspondence to:**

S. Gupta,
surya.gupta@usys.ethz.ch

**Citation:**

Gupta, S., Lehmann, P., Bonetti, S., Papritz, A., & Or, D. (2021). Global prediction of soil saturated hydraulic conductivity using random forest in a Covariate-based GeoTransfer Function (CoGTF) framework. *Journal of Advances in Modeling Earth Systems*, *13*, e2020MS002242. https://doi.org/10.1029/2020MS002242

Received 8 JUL 2020
Accepted 4 JAN 2021



**Abstract** Saturated hydraulic conductivity (Ksat) is a key soil hydraulic parameter for representing infiltration and drainage in land surface models. For large scale applications, Ksat is often estimated from pedotransfer functions (PTFs) based on easy-to-measure soil properties like soil texture and bulk density. The reliance of PTFs on data from uniform arable lands and the omission of soil structure limits the applicability of texture-based predictions of Ksat in vegetated lands. To include effects of terrain, climate, and vegetation in the derivation of a new global Ksat map at 1 km resolution, we harness technological advances in machine learning and availability of remotely sensed surrogate information. For model training and testing, a global compilation of 6,814 geo-referenced Ksat measurements from the literature was used. The accuracy assessment based on spatial cross-validation shows a concordance correlation coefficient (CCC) of 0.16 and a root mean square error (RMSE) of 1.18 for log10 Ksat values in cm/day (CCC = 0.79 and RMSE = 0.72 for non-spatial cross-validation). The generated maps of Ksat represent spatial patterns of soil formation processes more distinctly than previous global maps of Ksat based on easy-to-measure soil properties. The validation of the model indicates that Ksat could be modeled without bias using Covariate-based GeoTransfer Functions (CoGTFs) that harness spatially distributed surface and climate attributes, compared to soil information based PTFs. The relatively poor performance of all models in the validation (low CCC and high RMSE) highlights the need for the collection of additional Ksat values to train the model for regions with sparse data.

**Plain Language Summary** The soil saturated hydraulic conductivity (Ksat) defines how fast water infiltrates into and percolates through the soil. To model water flow at large scales, accurate maps of Ksat are needed. Usually, Ksat is not measured directly but deduced from well-known basic soil properties (e.g., soil texture, bulk density). However, these estimates neglect the influence of vegetation and climate on formation of soil structures that control Ksat. To improve global predictions of Ksat, we use a new spatially referenced Ksat data collection and apply machine learning to exploit correlations between Ksat and other properties (e.g., soil information, terrain, climate, and vegetation). These correlations are then implemented at global scale using maps of all relevant properties (so called *"environmental covariates"*) that were measured by remote sensing. We call this new approach to predictive Ksat mapping *"Covariate-based GeoTransfer Function"* (CoGTF) to highlight differences with other maps that neglect spatial correlation with soil formation processes and that are based only on soil data (so called *"pedotransfer functions"*, PTFs). We show that the new maps based on CoGTF perform better than approaches based on PTFs.

## 1. Introduction

The description of water, energy, and carbon fluxes between the land surface and the atmosphere relies heavily on the availability of soil hydraulic data (Fashi et al., 2016; Gutmann & Small, 2007; Montzka et al., 2017). At global scale, maps of soil hydraulic properties at ever increasing resolution are required for building Land Surface Models (Montzka et al., 2017). A prominent soil hydraulic property is the soil saturated hydraulic conductivity (Ksat) that affects the partitioning of rainfall between surface runoff and

infiltration (Zimmermann et al., 2013), and plays a critical role in a variety of hydrological and climatological applications (Fatichi et al., 2020; Gutmann & Small, 2007; Or, 2019).

For large scale applications (regional and global), soil hydraulic parameters are often estimated from easy-to-measure soil properties (e.g., texture, organic carbon content, bulk density) by means of pedotransfer functions (PTFs, Bouma, 1989; Santra & Das, 2008). For example, Dai et al. (2019) have recently produced 1 km resolution global maps of soil hydraulic properties using the median values of respective predictions by multiple PTFs. Likewise, Zhang and Schaap (2017) developed a global map of Ksat based on the Rosetta 3 PTF, employing artificial neural networks and bootstrap sampling (an extension of Schaap et al., 2001), and making use of two data sets from the USA (Ahuja et al., 1989; Rawls et al., 1982) and the Unsaturated Soil Hydraulic Database, UNSODA (Leij et al., 1996; Nemes et al., 2001). Similarly, Simons et al. (2020) have recently launched a new global Ksat map called HiHydroSoil v2.0 based on the PTF developed by Tóth et al. (2015) for various European regions.

Maps based on PTFs have several limitations. They are usually developed for specific geographic regions and thus only represent local conditions of soil forming processes (e.g., Jorda et al., 2015; Khlosi et al., 2016; Nemes et al., 2005; Saxton & Rawls, 2006; Tomasella & Hodnett, 1998; Wösten et al., 1999). This hinders their transferability across large geographical regions (Vereecken et al., 2016). In addition, PTFs generally ignore soil structure and pedogenic information and rely heavily on soil textural information (Fatichi et al., 2020), limiting their applicability in soils characterized by aggregation and formation of biopores. Moreover, PTFs are generally defined as a function of clay content, without consideration of the effect of different clay minerals on soil hydraulic properties (Hodnett & Tomasella, 2002).

The availability of highly resolved remote sensing (RS) environmental covariates offers new opportunities for injecting local information into the modeling of Ksat. Examples of the potential usefulness of such covariates are reported by several studies. For example, Obi et al. (2014) developed PTFs for many soil hydraulic properties using terrain attributes. Similarly, Sharma et al. (2006) combined PTFs with vegetation and topography indices (deduced from a digital elevation model, DEM), while Jana and Mohanty (2011) showed that the introduction of topographic attributes and information on vegetation (i.e., leaf area index, LAI) along with in situ soil basic properties could improve predictions of soil hydraulic properties.

In this paper, we hypothesize that Ksat predictions could be improved using a combination of soil variables and RS covariate layers integrated by using a machine learning (ML) framework. We profit from the advancement in RS techniques (providing spatial information on different ecological parameters with unprecedented resolution) to improve the predictions for soil hydraulic parameters and bridge the gap between site-specific soil properties and landscape variability. We apply concepts of predictive soil mapping using a large data set of Ksat measurements much larger than those previously used by Zhang and Schaap (2017) and Dai et al. (2019) and local information (soil, vegetation, climate) to define *"Covariate-based GeoTransfer Functions"* (CoGTFs) and to provide global estimates of Ksat values. To highlight the impact of geo-referencing soil properties and environmental covariates (i.e., the basic principles of predictive soil mapping), we use the term GTF and not PTF. We compare mapping accuracy using global and regional assessment including visual interpretation of produced spatial predictions. We show how this method (providing novel covariate-based maps of Ksat) could be used to overcome some of the limitations of traditional PTFs.

Our specific objectives are:

1. to improve the accuracy and spatial detail of global Ksat maps by harnessing state-of-the-art global remote sensing data products at 1 km spatial resolution,
2. to generate global maps of Ksat at different soil depths (0, 30, 60, and 100 cm),
3. to identify the key environmental covariates spatially correlated with Ksat.

We first describe the model training for Ksat mapping using a random forest (RF) ML algorithm and compare the results against maps generated with Rosetta 3, HiHydroSoil v2.0, and the map by Dai et al. (2019). Then, we validate the new CoGTF, Rosetta 3, HiHydroSoil v2.0 maps and the map of Dai et al. (2019) with an independent data set. We finally show the importance of using environmental covariates to capture spatial patterns of Ksat and improve the accuracy of predicted soil hydraulic properties.

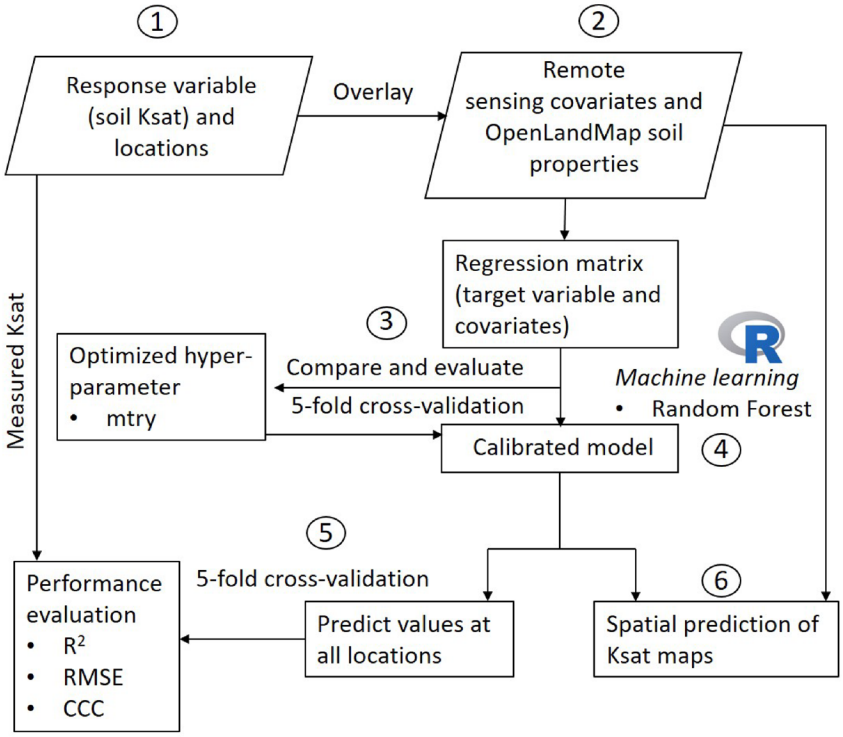


**Figure 1.** Computational workflow used to generate the new CoGTF Ksat map. See text for more details about the specific steps ($R^2$ is the coefficient of determination, RMSE the root mean square error, and CCC the concordance correlation coefficient).

## 2. Material and Methods

### 2.1. Covariate-based GeoTransfer Function (CoGTF) Framework

We propose here an integrated predictive soil modeling framework where soil variables are combined with environmental covariates using the RF method (Figure 1). A central hypothesis in this study is that environmental covariates could be harnessed to improve the global mapping of Ksat (Jana & Mohanty, 2011). The basis for such hypothesis is the dominant role of climate, topography, and vegetation in soil formation and thus in shaping local hydraulic transport properties. We refer to our approach as CoGTF to highlight the use of a large number of different environmental covariates, in contrast to traditional PTFs, generally based on soil properties only.

The CoGTF maps were produced using six principal steps:

1. prepare a geo-referenced data set of the response variable Ksat,
2. overlay the observation sites of the response and the environmental covariates (including predictions of basic soil properties), and produce a regression matrix,
3. optimize the main hyper-parameter (*mtry*) of the RF approach,
4. fit the RF model,
5. evaluate the performance of the RF model,
6. produce spatial predictions of the response variable (Ksat).

We implemented the spatial predictions and the creation of Ksat maps in the R environment for statistical computing (R Core Team, 2020) and provide our code at https://github.com/ETHZ-repositories/Ksat\text{_}mapping\text{_}2020/.

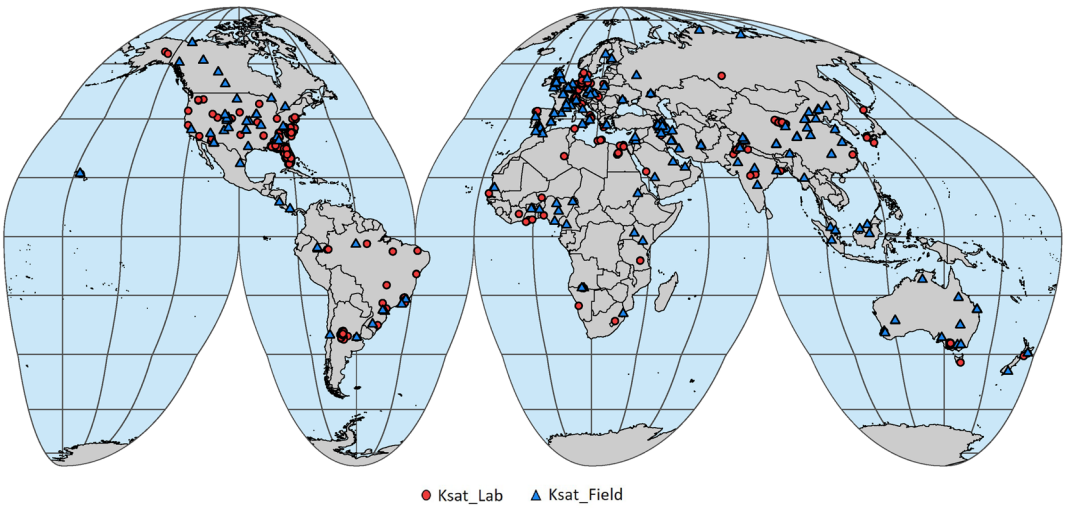


**Figure 2.** Spatial distribution of measured Ksat values (6,814 samples in total) used to produce the global Ksat map. Colors refer to laboratory (red) and field (blue) measurements. The map is presented in the Goode equal-area homolosine projection. For more details and access to the Ksat data see Gupta et al. (2020).

### 2.2. Training Data

Our first task was to enlarge the Ksat database beyond the ≈ 1,300 measurements used to train Rosetta 3 and the PTFs employed by Dai et al. (2019). HiHydroSoil v2.0 was based on PTF of Tóth et al. (2015) using about 3,500 Ksat values from Europe that are not publicly available. To this end, we collected geo-referenced Ksat data from the literature. The Ksat values were log-transformed ( $\log_{10}$Ksat), and cm/day was selected as standard unit. A detailed description of the data collection and processing is provided in Gupta et al. (2020). We managed to compile a total of 13,258 Ksat measurements from 1,908 sites across the globe as shown in Figure 2. Most training data are from North America (especially Florida), followed by Europe, Asia, South America, Africa, and Australia. The Ksat database (SoilKsatDB) includes both field (4,131) and lab (9,155) measurements.

To limit the over-representation of Florida (mainly arable land not representative of soils with natural vegetation) and allow a geographical balance with other national data sets, we randomly selected only 1% of the 6,532 Florida samples, so that a total of 6,814 Ksat measurements (from 821 sites) were finally used for Ksat mapping (the effect of sub-sampling Florida data is discussed in the Supplementary Information, SI). With respect to climatic regions, 4,298 Ksat values are from temperate regions and 1,109, 789, 582, and 36 from arid, tropical, boreal, and polar regions, respectively. The RF model was developed using all 6,814 Ksat measurements (on which results in sections 3.1 and 3.2 are based). To compare the CoGTF predictions with predictions by Rosetta 3, the map of Dai et al. (2019), and HiHydroSoil v2.0, we split the data into calibration and validation sets and re-trained our model with the calibration set only (see sec. 2.6 for details).

### 2.3. Soil and Environmental Covariates

For Ksat modeling at global scale, we used four soil properties (sand, clay, soil depth, and bulk density) and 24 layers with environmental covariates, all globally available from OpenLandMap.org (Hengl et al., 2019). The covariates were selected to represent ecological conditions essential for soil formation according to Jenny (1994). The covariates can be divided into five groups:

1. *Climate*: mean annual precipitation, temperature, temperature seasonality, maximum temperature of the warmest month, minimum temperature of the coldest month, precipitation of wettest month, precipitation of driest month (Chelsa products, Karger et al., 2017), cloud fraction (Wilson & Jetz, 2016), diffuse irradiation, direct irradiation, annual land surface temperature, monthly precipitation and its standard deviation (Brocca et al., 2019).

2. *Terrain*: landscape metrics (slope, aspect, and topographic wetness index, see Yamazaki et al., 2017). This group further included a lithological map (Hengl, 2018).
3. *Surface reflectance*: Landsat and MODIS data sets (red, near infra-red NIR, and short wave infra-red SWIR, see Hansen et al., 2013), snow probability (Buchhorn et al., 2017) and regularly flooded wetlands (Tootchi et al., 2019).
4. *Vegetation*: annual fraction of absorbed photosynthetically active radiation (FAPAR, Baret et al., 2016).
5. *Soil*: sand and clay contents as well as bulk density for different soil depths (matching the sampling depth of Ksat) from OpenLandMap (Hengl et al., 2019). Soil depth was used as further covariate to model the change of Ksat with depth (cf., Hengl & MacMillan, 2019).

A detailed list of all the covariates is provided in Table S1. All covariate layers were resampled to the standard grid with a spatial resolution of 1 km, covering latitudes between −62.0 and 87.37°. We did not map Antarctica as this continent is dominantly covered with permanent ice and lacked training points.

### 2.4. Computational Details on RF

After preparing the Ksat data and extracting all covariate values for locations with Ksat measurements from OpenLandMap.org (Hengl et al., 2019), a regression matrix was formed, and the RF model was fitted using the R packages *caret* (Kuhn, 2020) and *ranger* (Wright & Ziegler, 2017). The goodness-of-fit of the RF model was evaluated by the out-of-bag error (OOB, Hastie et al., 2009, section 15.3.1). Partial dependence plots (Hastie et al., 2009, section 10.13.2) were generated by the R packages *hexbin* (Carr et al., 2020), *lattice* (Sarkar, 2008) and *viridis* (Garnier, 2018).

The optimal value of the most sensitive hyper-parameter *mtry* of RF was determined by spatial five-fold cross-validation (CV, Lovelace et al., 2019, section 11.4), and default values of *ranger* were used for the remaining hyper-parameters, e.g., number of trees, minimal node size, maximum tree depth, splitting rule. For spatial CV the Earth surface was divided into 5° by 5° blocks as shown in Figure S1 of SI. We then formed the CV partition by randomly assigning the blocks to five subsets such that each contained about 20% of the data. The RF model was then fitted five times, leaving a subset out at a time and using it as test set. This CV procedure was repeated three times, and predictions of the three CV repetitions were averaged per Ksat measurement. The final CV results are shown for the optimal *mtry* in a hexbin plot with a LOWESS-line (Locally Weighted Scatterplot Smoothing, Cleveland, 1981) added to reveal conditional bias of the Ksat predictions. We also ran three repetitions of *non-spatial* five-fold CV with random allocation of single measurements to the five CV subsets to further assess predictive accuracy.

The relative importance of the covariates for modeling Ksat was assessed by the node impurity, which, for RF regression problems, is computed as the decrease of residual sum of squares (RSS) when a particular covariate splits the data at the nodes of a tree (Hastie et al., 2009, sections 10.13.1, 15.3.2). The variable that provides maximum decline in RSS (and consequently increase in node purity) is considered as the most important variable, the variable with the second largest RSS decrease is considered second most important, and so on.

### 2.5. Criteria to Assess Predictive Accuracy

The accuracy of the CV predictions and validations study (section 2.6) was evaluated using bias (BIAS), root mean square error (RMSE), coefficient of determination ($R^2$), and the concordance correlation coefficient (CCC).

BIAS and RMSE are defined as:

$$\text{BIAS} = \sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)}{n} \tag{1}$$

$$\text{RMSE} = \sqrt{\frac{\text{SSE}}{n}} \tag{2}$$

where

$$\mathrm{SSE} = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

is the sum of squared errors between the predictions $\hat{y}_i$ and the measurements $y_i$, and $n$ is the total number of observations.

$R^2$ is defined as:

$$R^2 = \left[1 - \frac{\mathrm{SSE}}{\mathrm{SST}}\right] \tag{3}$$

where

$$\mathrm{SST} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

is the total sum of squares and $\overline{y}$ the arithmetic mean of the measurements.

The concordance correlation coefficient (CCC, Lawrence & Lin, 1989), a further measure of the agreement between observed and predicted Ksat values, is given by:

$$\mathrm{CCC} = \frac{2 \cdot \rho \cdot s_{\hat{y}} \cdot s_y}{s_{\hat{y}}^2 + s_y^2 + \mathrm{BIAS}^2} \tag{4}$$

where $\rho$ is the Pearson correlation coefficient between $\hat{y}_i$ and $y_i$, and $s_{\hat{y}}$ and $s_y$ are the respective standard deviations. CCC is equal to 1 for a perfect model with $\hat{y}_i = y_i$ in which case we have BIAS = 0, $\rho = 1$ and $s_{\hat{y}} = s_y$.

### 2.6. Validation of Accuracy of Ksat Maps

The accuracy of predictions of Ksat by CoGTF and the three PTF-based approaches was evaluated with a subset of the Ksat data that was selected as follows. The Earth surface was again divided into 5° blocks as for spatial CV (Figure S1 of SI). For a fair comparison, Ksat measurements in blocks in North America or Europe were dropped because Rosetta 3 and the PTF used in HiHydroSoil v2.0 were mostly trained with data from these regions. We partitioned the blocks randomly into five subsets, such that each contained about 20% of the remaining 2,497 Ksat measurements. Each of these subsets was then used as test data for the five repetitions of the validation procedure.

In each repetition, we extracted predictions of Ksat from the Rosetta 3, HiHydroSoil v2.0, and the Dai et al. (2019) maps for locations with test data, and we re-computed CoGTF predictions of Ksat afresh by excluding the test measurements while fitting the RF model and optimizing *mtry*. Note that there are several variants of Rosetta 3, differing in the soil information used to build the neural networks (Zhang et al., 2019). We used the variant H3w, which is based on sand, silt, clay percentage and bulk density and is often chosen as standard variant (see map in Zhang & Schaap, 2019). Because we did not have all Ksat maps for all depths and HiHydroSoil v2.0 did not provide Ksat predictions for all test data, we focused on 0–30 cm depth interval. We considered in the five repetitions only 369, 230, 353, 363, and 348 measurements, respectively, that came from soil depth 0–30 cm and for which all the maps gave predictions. For each repetition we computed the accuracy criteria as used for CV (Table S2). This procedure warranted that the test data was independent from the data used to compute the Ksat predictions and adhered to the principles underlying spatial (cross-) validation.

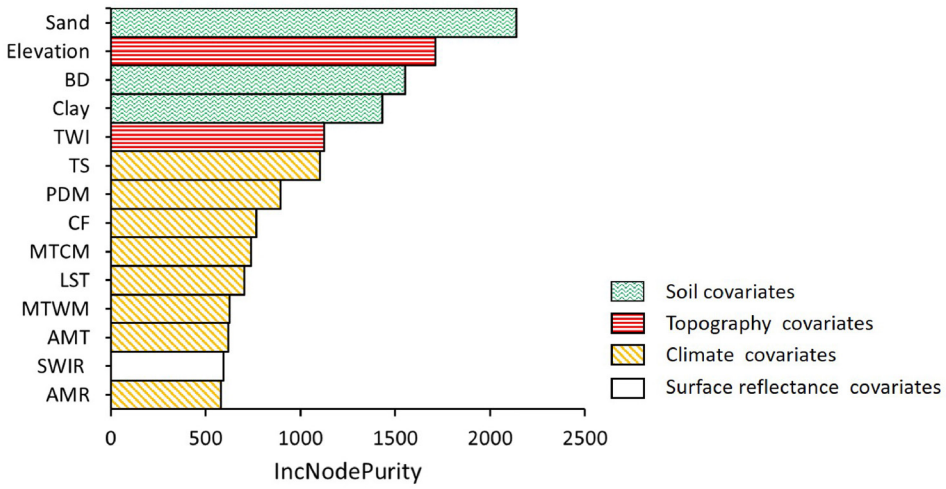


**Figure 3.** Relative importance of the covariates for modeling Ksat by the RF model. The abscissa displays the average increase in node purity (the larger the value, the more important is a covariate). The 14 most important covariates are shown here: sand, clay content and bulk density (BD) belong to the soil covariates. Elevation and topographic wetness index are terrain covariates. Temperature seasonality, precipitation of driest month (PDM), cloud fraction (CF), minimum temperature of coldest month, annual average land surface temperature (LST), maximum temperature of warmest month, mean annual temperature, and mean annual rainfall belong to the climate category. Shortwave infrared (SWIR) Landsat-7 band is from the surface reflectance group.

## 3. Results

### 3.1. Variable Importance and Accuracy of CoGTF Model

Figure 3 lists the most important covariates for Ksat modeling. The abscissa displays the increase in node purity: The higher the value, the more important is a covariate. As displayed in Figure 3, sand content emerged as the most important covariate, followed by elevation (important for soil formation and water flow), clay content, and bulk density. Climate covariates become predominant after the fifth covariate.

The CoGTF model fitted the logarithms of the Ksat measurements reasonably well (out-of-bag RMSE = 0.72 $\log_{10}$ cm/day, CCC = 0.80). The results of spatial and non-spatial CV are presented by hexbin density plots in Figures 4a and 4b, respectively. For spatial CV, the line of LOWESS fell onto the 1:1-line for Ksat predictions

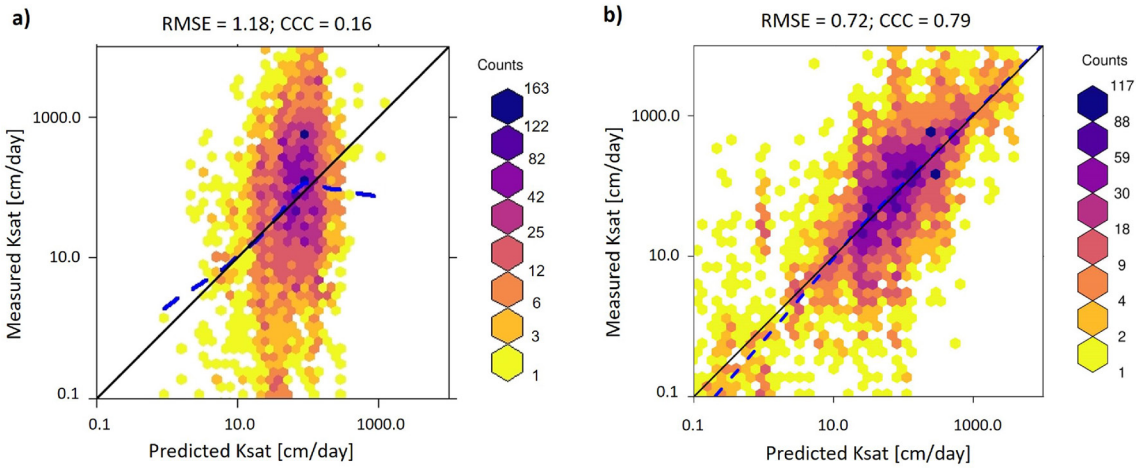


**Figure 4.** Correlation between measurements and (a) spatial and (b) non-spatial cross-validation (CV) predictions of Ksat based on the CoGTF model. The color code represents the number of observations in each hexagonal bin. The solid black line is the 1:1 line, and the blue dashed line is the LOWESS (locally weighted scatter plot smoothing) curve. RMSE is the root mean square error in $\log_{10}$ (cm/day), CCC concordance correlation coefficient. *Note*. that some extreme measurements are not displayed with the chosen axis scale.

between 10 and 350 cm/day. Hence predictions were conditionally unbiased in this range, and a slight bias was visible only for the sparse extreme predictions. For non-spatial CV, the line of LOWESS was below the 1:1 line for predictions less than 10 cm/day, indicating here a slight positive conditional bias.

Spatial CV results showed a substantially smaller accuracy than goodness-of-fit, with RMSE and BIAS equal to 1.18 and 0.116 $\log_{10}$ (cm/day), respectively, and CCC equal to 0.16. The criteria for non-spatial CV were close to those for goodness-of-fit (RMSE = 0.72, BIAS = −0.0039, both in $\log_{10}$ (cm/day), and CCC = 0.79). The strong contrast between spatial and non-spatial CV results warns against overly optimistic assessment of prediction accuracy by non-spatial CV.

### 3.2. Global CoGTF Map of Ksat

Global Ksat maps were produced for soil depths equal to 0, 30, 60, and 100 cm as proposed by the GlobalSoilMap standard (Arrouays et al., 2014). Figure 5a shows the CoGTF map of Ksat at 0 cm soil depth, while results for other soil depths are provided in Figures S2 and S3 in the SI. Predicted Ksat at 0 cm depth varies between 0.05 and 31,600 cm/day. High Ksat values were predicted for the Sahel and the west coast of South America, while low Ksat values were produced in the Oklahoma, Illinois, and Missouri states of America, Europe and parts of Asia (mainly India and northeastern part of China). In general, Ksat values decreased with depth, with the most significant reduction observed in North and South America, China, India, and Russia (see Figure S2 in the SI).

Figure 6a compares the probability density function (PDF) of the global CoGTF Ksat predictions with the PDFs of the 6,814 measured and fitted Ksat values. The predictions showed a narrower distribution compared to both measurements and fitted values. This smoothing is a well-known characteristic of mean square prediction methods that are trained by minimizing squared errors. For such methods, to which also RF belongs, the optimal predictor is the conditional expectation of the prediction target given the data (Hastie et al., 2009, section 2.4), and its variance is less than the variance of the training data ("law of total variance").

### 3.3. Comparison With Other Global Ksat Maps

Figure 5 compares the CoGTF Ksat map with the Rosetta 3, Dai et al. (2019) and HiHydroSoil v02.0 maps. The main differences between CoGTF and Rosetta 3 are the low Ksat values predicted by Rosetta 3 for tropical regions and the abrupt change in Rosetta 3 predictions in high latitude regions (northern North America, Scandinavia, Russia), as a consequence of the strong sensitivity of Rosetta 3 predictions to bulk density. In general, Rosetta 3 predicted smaller Ksat values than CoGTF for most regions of the world (see also Figure 6b), except for the northern high latitude regions and areas with coarse-textured soils such as the Sahara and the Middle East. Likewise, the Dai et al. (2019) and HiHydroSoil v2.0 maps also showed lower Ksat than CoGTF, in particular in the tropical regions. These shifts are clearly visible in Figure 6b, which compares the PDFs of global Ksat predictions by the four approaches with the PDF of the measurements. The medians of globally predicted $\log_{10}$Ksat were equal to 2.00 (CoGTF), 1.62 (Rosetta 3), 1.65 (Dai et al., 2019) and 1.17 (HiHydroSoil v2.0), respectively.

Figure 7 compares a subset of 4,614 Ksat measurements from 0 to 30 cm soil depth with predictions obtained by the four approaches. Note that for CoGTF, the predictions are a subset of the values displayed in Figure 4a. For Rosetta 3, Dai et al. (2019) and HiHydroSoil v2.0, the blue dashed LOWESS curves are mostly above the 1:1 lines. Hence these approaches suffer from substantial negative conditional biases. Additionally, such predictions also had larger RMSE and smaller CCC, signaling that Rosetta 3, Dai et al. (2019), and HiHydroSoil v2.0 predicted Ksat values with larger error than CoGTF.

### 3.4. Validation of Accuracy of CoGTF, Rosetta 3, HiHydroSoil v2.0, and Dai et al. (2019) Maps

Table 1 compares the accuracy of Ksat predictions by CoGTF, Rosetta 3, Dai et al. (2019), and HiHydroSoil v2.0 for the validation data (see section. 2.6). CoGTF predicted the validation data more accurately than the other methods. In particular, the magnitude of the BIAS of CoGTF was clearly smaller: on the original scale of the measurements, CoGTF under-predicted the validation data by a factor of 1.5, whereas

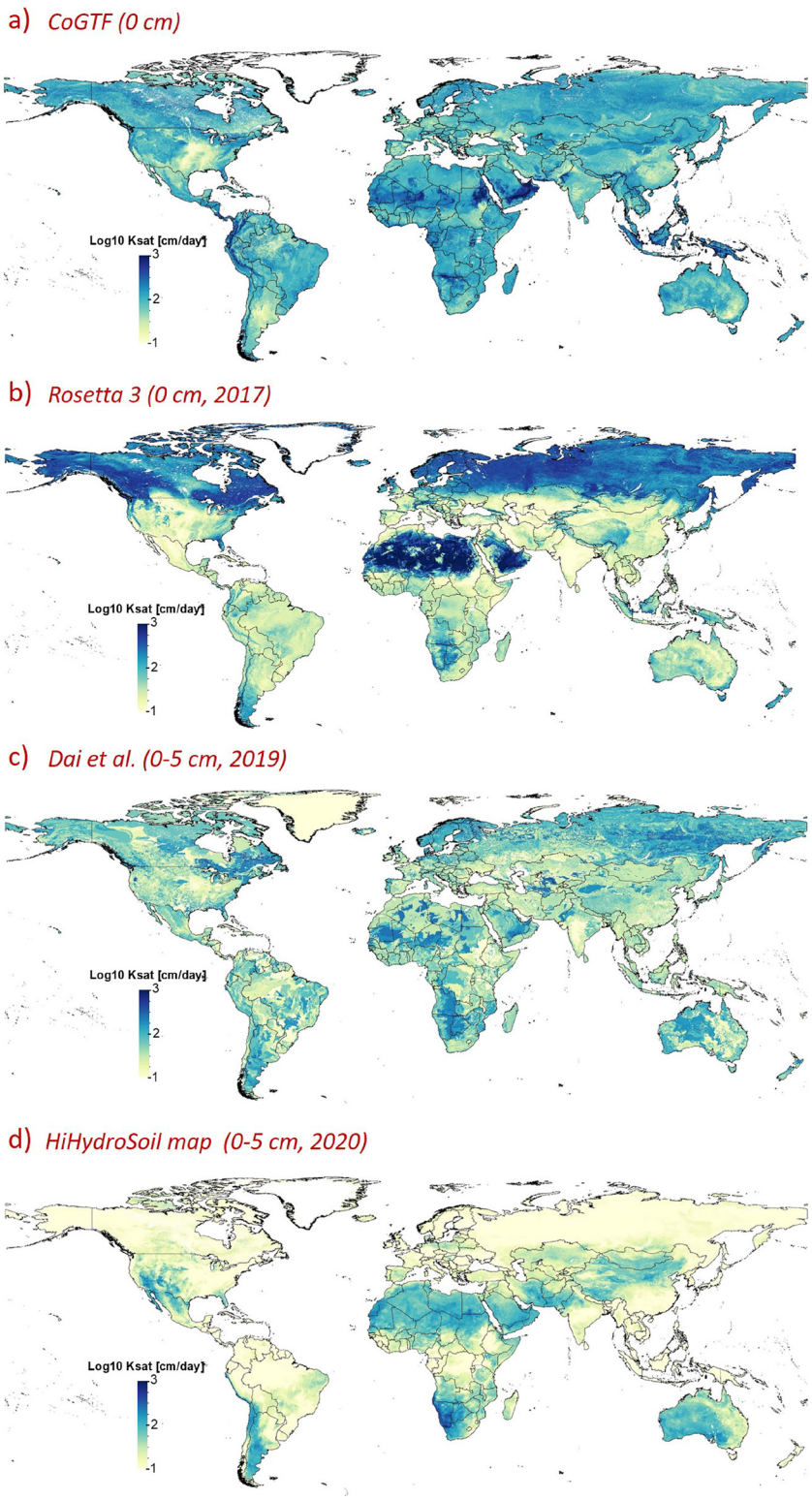


**Figure 5.** Global maps of topsoil Ksat produced by (a) CoGTF, (b) Rosetta 3 PTF (Zhang & Schaap, 2019), (c) Dai et al. (2019), and (d) HiHydroSoil v2.0 (Simons et al., 2020). The scale of Ksat was truncated at minimum and maximum values of 10 and 1,000 cm/day to enhance the spatial variation of Ksat displayed in the maps.

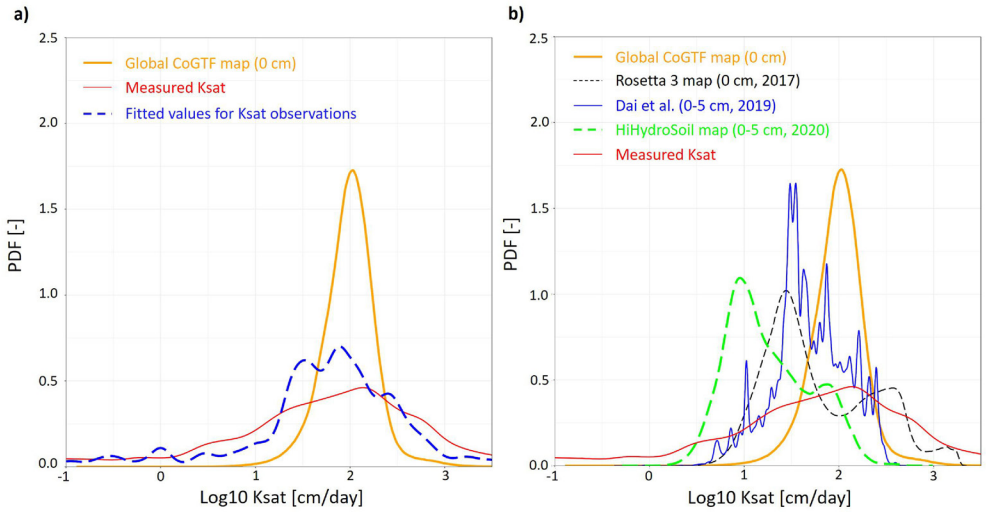


**Figure 6.** (a) Probability density functions (PDFs) of 6,814 measured (red) and fitted Ksat values (by CoGTF, blue), and of global GoGTF map (orange). (b) PDFs of global maps: CoGTF (orange), Rosetta 3 (black), Dai et al. (2019) (blue), and HiHydroSoil v2.0 (green) along with with PDF of Ksat measurements (red).

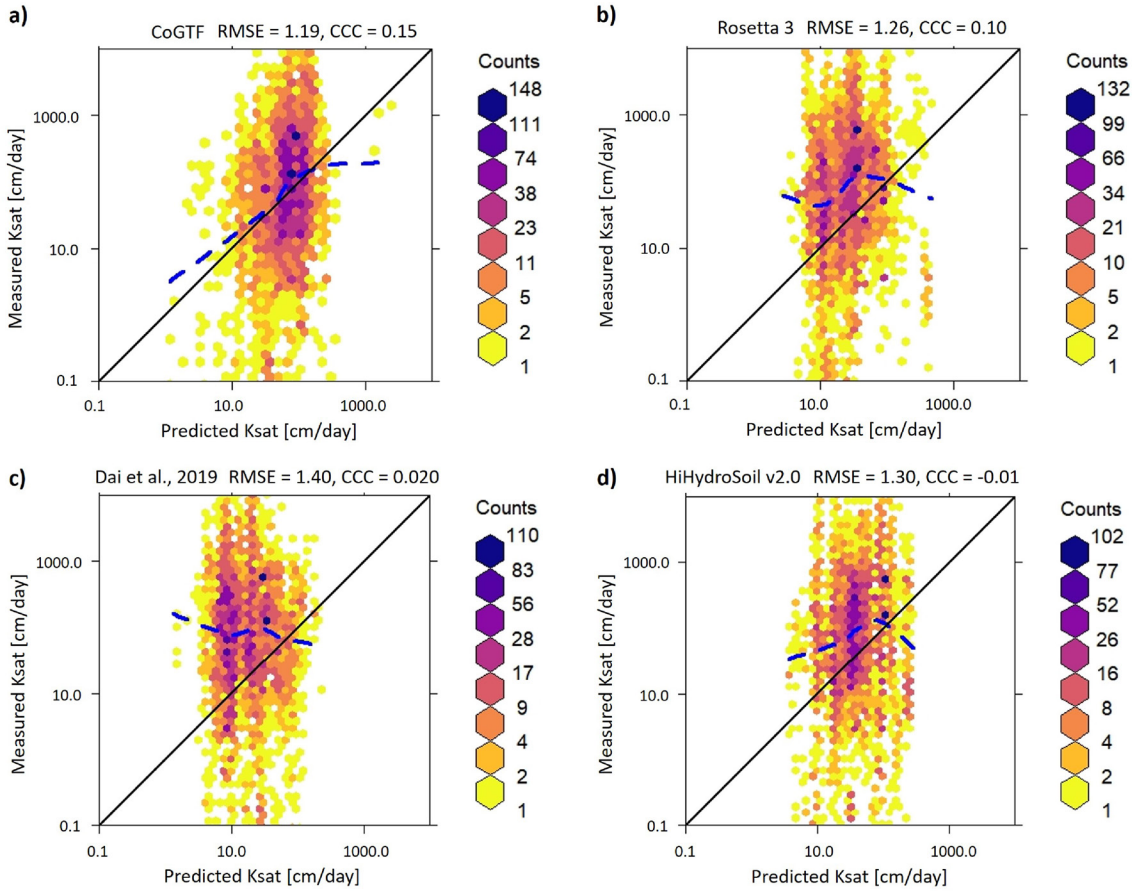


**Figure 7.** Correlations between 4,614 measurements of Ksat from 0 to 30 cm soil depth and (a) CoGTF (spatial CV results), (b) Rosetta 3 (Zhang & Schaap, 2019), (c) Dai et al. (2019) and (d) HiHydroSoil v2.0 (Simons et al., 2020) predictions. The solid black lines are 1:1 lines and the blue dashed lines are the LOWESS (locally weighted scatterplot smoothing) curves. RMSE is the root mean square error in $\log_{10}$ (cm/day), CCC concordance correlation coefficient. *Note*. that some extreme measurements are not displayed with the chosen axis scale.

**Table 1**
*Root Mean Square Error (RMSE), BIAS, Coefficient of Determination ($R^2$) and Concordance Correlation Coefficient (CCC) of Predictions of $\log_{10}$Ksat (units cm/day) by CoGTF, Rosetta 3, Dai et al. (2019) and HiHydroSoil v2.0 for the Validation Data (Five Subsets of Ksat Measurements From 0 to 30 cm Soil Depth, see Table S2)*

| Models | RMSE | BIAS | $R^2$ | CCC |
|---|---|---|---|---|
| CoGTF | 0.94 | −0.18 | −0.004 | 0.11 |
| Rosetta 3 | 1.15 | −0.69 | −0.52 | 0.09 |
| Dai et al. (2019) | 1.07 | −0.52 | −0.31 | 0.11 |
| HiHydroSoil v2.0 | 1.25 | −0.65 | −0.78 | −0.08 |

*Note*. The CoGTF model was trained without the validation data sets (see Section 2.6).

Rosetta 3, Dai et al. (2019), and HiHydroSoil v2.0 under-predicted by factors of 4.9, 3.3, and 4.4, respectively. Furthermore, the reported RMSE for the log-transformed data correspond to coefficients of relative predictive variation of 8.7 (CoGTF), 14.1 (Rosett 3), 11.7 (Dai et al., 2019), and 17.7 (HiHydroSoil v2.0) on the original scale of the measurements.

## 4. Discussion

### 4.1. Characteristics of the CoGTF Global Ksat Map

In this paper, we have produced global estimates of Ksat by linking terrain, climate, vegetation, and soil spatial covariates to measured Ksat values, thereby injecting local information (usually ignored by traditional PTFs) into spatial Ksat modeling. We refer to this as the Covariate-based GeoTransfer Functions (CoGTF) approach. The newly developed global CoGTF maps (Figure 5 and Figure S2 of SI) featured high Ksat values in the northern regions of South America, the central regions of Africa, and Southeast Asia (mainly Indonesia, Malaysia, Myanmar, Philippines, Singapore, and Thailand), probably due to high rainfall, hot temperature, dense vegetation, and related factors, which all foster intense soil structure formation. Likewise, high values of Ksat up to about 100–300 cm/day were predicted for desert regions such as the Thar desert in India, Rub' al Khali desert in Saudi Arabia, deserts of northern and southern Africa and central Australia, where sandy soils with high permeability prevail. Colombia, Peru, and parts of Brazil showed high Ksat values as well, probably due to high soil organic carbon content (SOC, Allison, 1973), which may be linked to distinct soil aggregation (Beare et al., 1994). This hypothesis is supported by high SOC values predicted by OpenLandMap (Hengl et al., 2019) for these regions. Similar results were reported by Belk et al. (2007), with Ksat values in a tropical Brazilian forest ranging between 100 and 1,000 cm/day at the soil surface.

Our results show that rainfall, temperature, and their variations are the most important climate covariates for Ksat mapping (Figure 3). They do not only act as catalyst in soil chemical reactions (weathering of soil minerals), but also determine type and biomass of vegetation, which are both important for soil structure formation (see partial dependence plots in Figure S4 of SI). In this study, FAPAR (used as a proxy for vegetation) turned out to be of minor relevance for Ksat mapping (see Figure 3). The lack of evidence that vegetation indices like FAPAR correlate with Ksat may be related to the common practice to sample soil for Ksat measurements at locations without large roots. Additional Ksat measurements from such sites in forested areas may reveal the relevance of vegetation indices.

Central parts of India, the eastern parts of Australia and USA and parts of China showed low Ksat values, likely due to the presence of soils rich in clay that reduces the soil permeability (see as well discussion on role of clay mineral type in text S1 of SI). The western part of USA, some Middle East countries (e.g., Iran, Turkey) and northern parts of Algeria had low Ksat values that may be related to high elevation, low rainfall, sparse vegetation, and thus less intense soil structure formation. Many studies have recognized the indirect influence of elevation on soil properties (Leij et al., 2004). Similarly, land-use e.g., forest or pasture directly impacts Ksat. For example, Chandler et al. (2018) showed that forests display larger soil hydraulic conductivity values compared to pastures.

In contrast to the Rosetta 3 map, the new CoGTF map did not show spatial dominance of large Ksat values in the Sahara region (see Figure 5). This may be caused by a lack of Ksat measurements in this region. To overcome this limitation, an option is to add expert-guess, pseudo-observations of Ksat in order to minimize extrapolation effects in under-sampled geographic areas lacking field observations (Hengl et al., 2017). To explore this opportunity, we added 100 pseudo-observations at random locations in deserts and re-trained the CoGTF model with this extended data set. Figure S5 of SI shows the resulting map with higher Ksat predictions in Sahara, but no noteable differences elsewhere. However, without observational evidence of dominating high Ksat in this region, we suggest the use of the map of Figure 5.

**Table 2**
*Root Mean Square Error (RMSE), Coefficient of Determination ($R^2$), BIAS and Concordance Correlation Coefficient (CCC) of $\log_{10}$Ksat (Units cm/day) Predictions From Three Repetitions of 5-Fold Spatial Cross-validation for CoGTF Models That Included Different Sets of Covariates (p is the Number of Covariates, mtry the Optimal Parameter Value Determined by Spatial CV and Non-spatial CV)*

| Covariate Set | *p* | *mtry* | RMSE | $R^2$ | BIAS | CCC |
|---|---|---|---|---|---|---|
| Spatial Cross-validation | | | | | | |
| Soil | 4 | 1 | 1.22 | 0.04 | 0.12 | 0.12 |
| Vegetation, Terrain and Climate | 24 | 2 | 1.21 | 0.05 | 0.14 | 0.10 |
| Soil, Vegetation, Terrain and Climate | 28 | 6 | 1.18 | 0.10 | 0.11 | 0.16 |
| Non-spatial Cross-validation | | | | | | |
| Soil | 4 | 2 | 0.75 | 0.63 | −0.001 | 0.77 |
| Vegetation, Terrain and Climate | 24 | 16 | 0.73 | 0.65 | −0.004 | 0.78 |
| Soil, Vegetation, Terrain and Climate | 28 | 6 | 0.72 | 0.66 | −0.003 | 0.79 |

### 4.2. Effects of the Spatially Clustered Florida Data

Out of the available 13,267 Ksat measurements, we used only 6,814 for Ksat mapping to avoid a distortion of the predictions by the clustered data from Florida. The full data set contained 6,532 observations from Florida, but we used only 1% of them for global mapping. Figure S6 of SI compares the map computed with all 13,267 Ksat measurements with the map trained only on 6,814 measurements. The difference between these maps (Figure S6c) shows a clear impact in regions with sandy soils such as the Sahara, the central part of Africa and the Middle East with significantly higher Ksat predictions when all Florida points were included. A similar effect was observed in parts of South America and Australia. On the other hand, southern Africa and the higher northern latitudes showed higher Ksat values when only 1% of the Florida data were used. The observed effects of using the complete data set from Florida can be explained by prevalence of high sand content and high Ksat values in the Florida data (see Figure S7 of SI).

### 4.3. Gain of Prediction Accuracy by Using Environmental Covariates in CoGTF Model

We included environmental covariates in the spatial modeling of Ksat that are commonly believed to capture the effects of vegetation, terrain, and climate on soil formation and, in turn, on hydraulic soil properties (Hao et al., 2019; Ottoni et al., 2018). To investigate the predictive power of these covariates, we compared spatial CV results for CoGTF models, fitted (a) only to soil covariates from OpenLandMap, (b) only to environmental covariates (vegetation, terrain, climate), and (c) to both types of covariates (standard CoGTF model). The resulting maps are shown in Figure S8 of SI, and Table 2 reports the spatial CV results.

The CoGTF model that included only the soil covariates from OpenLandMap provided the least accurate predictions, while the standard CoGTF model with all covariates resulted to be the most accurate. The model using only the environmental covariates fared intermediate. The differences in RMSE were small, and the reduction resulting from considering the environmental covariates was modest. At first sight, it therefore appears that little can be gained by including the environmental covariates in the modeling. However, one must bear in mind that the OpenLandMap soil covariates were themselves predicted from a similar set of environmental covariates. It is therefore not surprising that the model including only the soil covariates performed quite well. Adhering to a paradigm of predictive modeling, we nevertheless prefer the model with the best predictive performance over a model chosen based on some possibly preconceived notions.

The non-spatial cross validation results are also shown to further illustrate the difference between two approaches (spatial vs. non-spatial CV). The results shows a significant difference between RMSE and $R^2$ with over-optimistic model performance for non-spatial CV. Therefore, it is highly recommended to use the spatial cross validation for spatial mapping (Ploton et al., 2020).

### 4.4. Use of the Global CoGTF Ksat Maps and Future Developments

Although the CoGTF map provided more accurate and less biased predictions than previously published, a validation RMSE equal to 0.94 $\log_{10}$ (cm/day) revealed a still very limited overall accuracy of our Ksat predictions. Furthermore, CoGTF predictions varied much less than the training data, an inherent characteristic of mean square prediction methods, as mentioned above. Another property of RF method further contributes to the smoothing: its predictions are means of response values for the subsets of the data that form the "leaves" of the trees. Hence the predictions are always limited to the range of the training data and they vary less than them. The PDF of the CoGTF predictions was therefore narrower than the PDF of the Ksat measurements. This is not a severe problem if Ksat is of direct interest. However, if Ksat values are fed into a nonlinear transport model, then smoothed Ksat predictions potentially induce strong biases in the model output. The PDFs of the predictions by the other approaches were wider and multi-modal. In particular, the PDF of the global Rosetta 3 predictions had three peaks. The lower and middle peak in the distribution might be the result of the over-representation of loamy and sandy soil samples in the data set used to train Rosetta 3 (see Table 2 in Zhang & Schaap, 2019), and the small upper peak was likely related to uniformly high Ksat predictions for the Sahara region.

The global CoGTF maps provide information on Ksat at different depths for regional and global scale studies. The CoGTF maps presented here have a spatial resolution of 1 km. This resolution can likely be improved in the near future, considering various initiatives to estimate soil and environmental covariate information with higher spatial resolution. In addition, the maps can be improved if more comprehensive Ksat data will become available.

## 5. Summary and Conclusions

Soil saturated hydraulic conductivity (Ksat) is an important soil parameter in Earth system and land surface models that require spatially distributed soil hydraulic information at global scale. The major limitations of currently available maps of Ksat are that (1) they were developed using only a limited number of Ksat measurements, mainly from temperate regions, (2) they were derived only from basic soil properties, thus ignoring the effect of biologically induced soil structure formation and weathering processes affecting the clay mineralogy, and (3) they do not yet benefit from the wealth of environmental covariates nowadays available. Therefore, we proposed a new global map of Ksat obtained by linking a new data set of measured Ksat values (6,814 samples) with 24 environmental covariates and three soil properties (sand and clay content, bulk density) to inject local-scale spatial information on vegetation, climate, and topography into Ksat predictions. The new map combines geo-referenced information of soil properties and environmental covariates and is called Covariate-based GeoTransfer Functions (CoGTF) map.

We used the random forest ML algorithm to fit the Ksat models. The accuracy of the CoGTF map was assessed by spatial cross-validation CCC and RMSE. The CCC and RMSE (in $\log_{10}$ (cm/day)) were 0.16 and 1.18, respectively (0.79 and 0.72 for non-spatial cross-validation). The CoGTF global Ksat map was compared with other global Ksat maps that were computed with PTFs that ignore environmental covariates (Rosetta 3, Dai et al. (2019) and HiHydroSoil v2.0). A major difference between the CoGTF and the other maps was the much lower Ksat values predicted for tropical regions compared to the CoGTF Ksat map. The tropical regions are expected to have rather high Ksat values due to soil formation processes that differ from those in temperate regions and presence of more inactive, non-swelling, clay minerals (kaolinite). The effects of active and inactive clay minerals on Ksat are captured by the CoGTF map because the formation of clay minerals depends on precipitation, temperature and dense vegetation. The CoGTF map, Rosetta 3 map, HiHydroSoil v2.0, and the map of Dai et al. (2019) were validated using test data that were not used to train the models, and the result showed that the CoGTF map performed better than the other models. Considering that also the model performance of CoGTF (using environmental covariates and the best available Ksat data set) is relatively poor (CCC = 0.11), the application of global Ksat maps is associated with uncertainty especially in regions without supporting training data. The presented systematic validation of the different models show how critical the collection of additional Ksat data is to improve the model performance and to reproduce global patterns. Despite the uncertainty related to the large RMSE and small CCC value, the small (absolute) BIAS value of CoGTF (0.18) compared to 0.52–0.69 for PTF models supports

the incorporation of environmental covariates. We thus propose to transition from PTFs based only on soil texture and bulk density for the prediction of Ksat to CoGTF models that exploit comprehensive sets of covariates, which characterize conditions of soil formation. The study provides a blueprint for how geo-referenced covariates could be used within the ML framework to improve Ksat predictive mapping. Moreover, the resulting CoGTF global maps are readily updateable as more information (i.e., covariates and Ksat measurements) become available.

## Data Availability Statement

The data sets produced in this study are available at https://doi.org/10.5281/zenodo.3934853.

**Acknowledgments**
The study was supported by ETH Zurich (Grant ETH-18 18-1). We are grateful for the scientific collaboration with Tom Hengl (OpenGeoHub Foundation, Wageningen, the Netherlands) and fruitful discussions on this paper. We want to acknowledge the www.openlandmap.org portal for providing the soil properties map. We would like to thank Samuel Bickel and Simone Fatichi (ETH Zurich) for insightful discussions. We also want to thank Zhongwang Wei for providing the Rosetta 3 Ksat maps.

## References

Ahuja, L. R., Cassel, D. K., Bruce, R. R., & Barnes, B. B. (1989). Evaluation of spatial distribution of hydraulic conductivity using effective porosity data. *Soil Science*, *148*(6), 404–411. https://doi.org/10.1097/00010694-198912000-00002

Allison, F. E. (1973). *Soil organic matter and its role in crop production*. Elsevier.

Arrouays, D., McBratney, A. B., Minasny, B., Hempel, J. W., Heuvelink, G. B. M., MacMillan, R. A., et al. (2014). The GlobalSoilMap project specifications. In *GlobalSoilMap Basis of the global spatial soil information system* (pp. 9–12). CRC Press.

Baret, F., Weiss, M., Verger, A., & Smets, B. (2016). *Atbd for Lai, Fapar and Fcover from Proba-V products at 300m resolution (Geov3). INRA*. Retrieved from https://land.copernicus.eu/global/sites/cgls.vito.be/files/products/ImagineS\text{_}RP2.1\text{_}ATBD-LAI300m\text{_}I1.73.pdf

Beare, M. H., Hendrix, P. F., & Coleman, D. C. (1994). Water-stable aggregates and organic matter fractions in conventional- and no-Tillage soils. *Soil Science Society of America Journal*, *58*(3), 777–786. https://doi.org/10.2136/sssaj1994.03615995005800030020x

Belk, E. L., Markewitz, D., Rasmussen, T. C., Carvalho, E. J. M., Nepstad, D. C., & Davidson, E. A. (2007). Modeling the effects of throughfall reduction on soil water content in a Brazilian Oxisol under a moist tropical forest. *Water Resources Research*, *43*(8), W08432. https://doi.org/10.1029/2006WR005493

Bouma, J. (1989). Using soil survey data for quantitative land evaluation. In *Advances in soil science* (pp. 177–213). Springer.

Brocca, L., Filippucci, P., Hahn, S., Ciabatta, L., Massari, C., Camici, S., et al. (2019). SM2RAIN-ASCAT (2007–2018): Global daily satellite rainfall from ASCAT soil moisture. *Earth System Science Data Discussion*, 1–31.

Buchhorn, M., Bertels, L., Smets, B., Lesiv, M., & Wur, N. (2017). *Copernicus global land operations "vegetation and energy"*.

Carr, D., & ported by Nicholas Lewin-Koh, Maechler, M., & contains copies of lattice functions written by Deepayan Sarkar. (2020). *hexbin: Hexagonal binning routines [Computer software manual]*. Retrieved from https://CRAN.R-project.org/package=hexbin(Rpackageversion1.28.1

Chandler, K. R., Stevens, C. J., Binley, A., & Keith, A. M. (2018). Influence of tree species and forest land use on soil hydraulic conductivity and implications for surface runoff generation. *Geoderma*, *310*, 120–127. https://doi.org/10.1016/j.geoderma.2017.08.011

Cleveland, W. S. (1981). LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician*, *35*(1), 54. https://doi.org/10.2307/2683591

Dai, Y., Xin, Q., Wei, N., Zhang, Y., Shangguan, W., Yuan, H., et al. (2019). A global high-resolution data set of soil hydraulic and thermal properties for land surface modeling. *Journal of Advances in Modeling Earth Systems*.

Fashi, F. H., Gorji, M., & Shorafa, M. (2016). Estimation of soil hydraulic parameters for different land-uses. *Model. Earth Syst. Environ.*, *2*(4), 1–7. https://doi.org/10.1007/s40808-016-0229-0

Fatichi, S., Or, D., Walko, R., Vereecken, H., Young, M. H., Ghezzehei, T. A., et al. (2020). Soil structure is an important omission in earth system models. *Nature Communications*, *11*(1), 1–11. https://doi.org/10.1038/s41467-020-14411-z

Garnier, S. (2018). *Viridis: Default color maps from "matplotlib" [Computer software manual]*. Retrieved from https://CRAN.R-project.org/package=viridis(Rpackageversion0.5.1

Gupta, S., Hengl, T., Lehmann, P., Bonetti, S., & Or, D. (2020). SoilKsatDB: A global compilation of soil saturated hydraulic conductivity measurements. *Zenodo*. https://doi.org/10.5281/zenodo.3752721

Gutmann, E. D., & Small, E. E. (2007). A comparison of land surface model soil hydraulic properties estimated by inverse modeling and pedotransfer functions. *Water Resources Research*, *43*(5). https://doi.org/10.1029/2006wr005135

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., et al. (2013). High-resolution global maps of 21st-century forest cover change. *Science*, *342*(6160), 850–853. https://doi.org/10.1126/science.1244693

Hao, M., Zhang, J., Meng, M., Chen, H. Y., Guo, X., Liu, S., & Ye, L. (2019). Impacts of changes in vegetation on saturated hydraulic conductivity of soil in subtropical forests. *Scientific Reports*, *9*(1), 1–9. https://doi.org/10.1038/s41598-019-44921-w

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning; data mining, inference and prediction* (2nd ed.). Springer.

Hengl, T. (2018). Global landform and lithology class at 250 m based on the USGS global ecosystem map. *Zenodo*. Retrieved from https://doi.org/10.5281/zenodo.1461428

Hengl, T., Collins, T., Wheeler, I., & MacMillan, R. (2019). Everybody has a right to know What's happening with the planet: Toward a Global Commons. *Zenodo*. https://doi.org/10.5281/zenodo.3274294

Hengl, T., de Jesus, J. M., Heuvelink, G. B., Gonzalez, M. R., Kilibarda, M., Blagotić, A., et al. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS One*, *12*(2), e0169748. https://doi.org/10.1371/journal.pone.0169748

Hengl, T., & MacMillan, R. A. (2019). *Predictive soil mapping with R* (p. 370). OpenGeoHub foundation. ISBN:978-0-359-30635-0. Retrieved from www.soilmapper.org

Hodnett, M., & Tomasella, J. (2002). Marked differences between van Genuchten soil water-retention parameters for temperate and tropical soils: A new water-retention pedo-transfer functions developed for tropical soils. *Geoderma*, *108*(3–4), 155–180. https://doi.org/10.1016/s0016-7061(02)00105-2

Jana, R. B., & Mohanty, B. P. (2011). Enhancing PTFs with remotely sensed data for multi-scale soil water retention estimation. *Journal of Hydrology*, *399*(3–4), 201–211. https://doi.org/10.1016/j.jhydrol.2010.12.043

Jenny, H. (1994). *Factors of soil formation: A system of quantitative pedology*. Courier Corporation.

Jorda, H., Bechtold, M., Jarvis, N., & Koestel, J. (2015). Using boosted regression trees to explore key factors controlling saturated and near-saturated hydraulic conductivity. *European Journal of Soil Science*, *66*(4), 744–756. https://doi.org/10.1111/ejss.12249

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., et al. (2017). Climatologies at high resolution for the earth's land surface areas. *Scientific Data*, *4*, 170122. https://doi.org/10.1038/sdata.2017.122

Khlosi, M., Alhamdoosh, M., Douaik, A., Gabriels, D., & Cornelis, W. M. (2016). Enhanced pedotransfer functions with support vector machines to predict water retention of calcareous soil. *European Journal of Soil Science*, *67*(3), 276–284. https://doi.org/10.1111/ejss.12345

Kuhn, M. (2020). *Caret: Classification and regression training [computer software manual]*. Retrieved from https://CRAN.R-project.org/package=caret(Rpackageversion6.0-86

Lawrence, I., & Lin, K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 255–268.

Leij, F., Alves, W., Van Genuchten, M. T., & Williams, J. (1996). *The UNSODA unsaturated soil hydraulic database; User's Manual, Version 1.0. Rep. EPA/600/R-96, 95, 103*.

Leij, F. J., Romano, N., Palladino, M., Schaap, M. G., & Coppola, A. (2004). Topographical attributes to predict soil hydraulic properties along a hillslope transect. *Water Resources Research*, *40*(2), W02407. https://doi.org/10.1029/2002WR001641

Lovelace, R., Nowosad, J., & Muenchov, J. (2019). *Geocomputation with R*. CRC Press. https://doi.org/10.1201/9780203730058

Montzka, C., Herbst, M., Weihermüller, L., Verhoef, A., & Vereecken, H. (2017). A global data set of soil hydraulic properties and sub-grid variability of soil water retention and hydraulic conductivity curves. *Earth System Science Data*, *9*(2), 529–543. https://doi.org/10.5194/essd-9-529-2017

Nemes, A., Rawls, W. J., & Pachepsky, Y. A. (2005). Influence of organic matter on the estimation of saturated hydraulic conductivity. *Soil Science Society of America Journal*, *69*(4), 1330–1337. https://doi.org/10.2136/sssaj2004.0055

Nemes, A., Schaap, M., Leij, F., & Wösten, J. (2001). Description of the unsaturated soil hydraulic database UNSODA version 2.0. *Journal of Hydrology*, *251*(3–4), 151–162. https://doi.org/10.1016/s0022-1694(01)00465-6

Obi, J. C., Ogban, P. I., Ituen, U. J., & Udoh, B. T. (2014). Development of pedotransfer functions for coastal plain soils using terrain attributes. *Catena*, *123*, 252–262. https://doi.org/10.1016/j.catena.2014.08.015

Or, D. (2019). The tyranny of small scales–on representing soil processes in global land surface models. *Water Resources Research*, *55*. https://doi.org/10.1029/2018wr024050. https://doi.org/10.1029/2019wr024846

Ottoni, M. V., Ottoni Filho, T. B., Schaap, M. G., Lopes-Assad, M. L. R., & Rotunno Filho, O. C. (2018). Hydrophysical database for Brazilian soils (hybras) and pedotransfer functions for water retention. *Vadose Zone Journal*, *17*(1), 1–17. https://doi.org/10.2136/vzj2017.05.0095

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., et al. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, *11*(1), 1–11. https://doi.org/10.1038/s41467-020-18321-y

Rawls, W. J., Brakensiek, D. L., & Saxtonn, K. (1982). Estimation of soil water properties. *Transactions of the ASAE*, *25*(5), 1316–1320.

R Core Team. (2020). *R: A language and environment for statistical computing [computer software manual]*. Vienna, Austria. Retrieved from https://www.R-project.org/

Santra, P., & Das, B. S. (2008). Pedotransfer functions for soil hydraulic properties developed from a hilly watershed of Eastern India. *Geoderma*, *146*(3–4), 439–448. https://doi.org/10.1016/j.geoderma.2008.06.019

Sarkar, D. (2008). *Lattice: Multivariate data visualization with r*. Springer. Retrieved from http://lmdvr.r-forge.r-project.org ISBN 978-0-387-75968-5.

Saxton, K. E., & Rawls, W. J. (2006). Soil water characteristic estimates by texture and organic matter for hydrologic solutions. *Soil Science Society of America Journal*, *70*(5), 1569–1578. https://doi.org/10.2136/sssaj2005.0117

Schaap, M. G., Leij, F. J., & Van Genuchten, M. T. (2001). Rosetta: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. *Journal of Hydrology*, *251*(3–4), 163–176. https://doi.org/10.1016/s0022-1694(01)00466-8

Sharma, S. K., Mohanty, B. P., & Zhu, J. (2006). Including topography and vegetation attributes for developing pedotransfer functions. *Soil Science Society of America Journal*, *70*(5), 1430–1440. https://doi.org/10.2136/sssaj2005.0087

Simons, G., Koster, R., & Droogers, P. (2020). *Hihydrosoil v2. 0-high resolution soil maps of global hydraulic properties*. Retrieved from https://www.futurewater.nl/wp-content/uploads/2020/10/HiHydroSoil-v2.0-High-Resolution-Soil-Maps-of-Global-Hydraulic-Properties.pdf

Tomasella, J., & Hodnett, M. G. (1998). Estimating soil water retention characteristics from limited data in Brazilian Amazonia. *Soil Science*, *163*(3), 190–202. https://doi.org/10.1097/00010694-199803000-00003

Tootchi, A., Jost, A., & Ducharne, A. (2019). Multi-source global wetland maps combining surface water imagery and groundwater constraints. *Earth System Science Data*, *11*, 189–220. https://doi.org/10.5194/essd-11-189-2019

Tóth, B., Weynants, M., Nemes, A., Makó, A., Bilas, G., & Tóth, G. (2015). New generation of hydraulic pedotransfer functions for Europe. *European Journal of Soil Science*, *66*(1), 226–238. https://doi.org/10.1111/ejss.12192

Vereecken, H., Schnepf, A., Hopmans, J. W., Javaux, M., Or, D., Roose, T., et al. (2016). Modeling soil processes: Review, key challenges, and new perspectives. *Vadose Zone Journal*, *15*(5). https://doi.org/10.2136/vzj2016.01.0001.letter.ed

Wilson, A. M., & Jetz, W. (2016). Remotely sensed high-resolution global cloud dynamics for predicting ecosystem and biodiversity distributions. *PLoS Biology*, *14*(3), 1–20. https://doi.org/10.1371/journal.pbio.1002415

Wösten, J., Lilly, A., Nemes, A., & Le Bas, C. (1999). Development and use of a database of hydraulic properties of European soils. *Geoderma*, *90*(3–4), 169–185. https://doi.org/10.1016/s0016-7061(98)00132-3

Wright, M. N., & Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. https://doi.org/10.18637/jss.v077.i01

Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J. C., et al. (2017). A high-accuracy map of global terrain elevations. *Geophysical Research Letters*, *44*(11), 5844–5853. https://doi.org/10.1002/2017gl072874

Zhang, X., Zhu, J., Zhu, O., Matocha, C., & Edwards, D. (2019). Effect of macroporosity on pedotransfer function estimates at the field scale. *Vadose Zone Journal*, *18*(1), 1–15. https://doi.org/10.2136/vzj2018.08.0151

Zhang, Y., & Schaap, M. G. (2017). Weighted recalibration of the Rosetta pedotransfer model with improved estimates of hydraulic parameter distributions and summary statistics (Rosetta 3). *Journal of Hydrology*, *547*, 39–53. https://doi.org/10.1016/j.jhydrol.2017.01.004

Zhang, Y., & Schaap, M. G. (2019). Estimation of saturated hydraulic conductivity with pedotransfer functions: A review. *Journal of Hydrology*, *575*, 1011–1030. https://doi.org/10.1016/j.jhydrol.2019.05.058

Zimmermann, A., Schinn, D. S., Francke, T., Elsenbeer, H., & Zimmermann, B. (2013). Uncovering patterns of near-surface saturated hydraulic conductivity in an overland flow-controlled landscape. *Geoderma*, *195–196*, 1–11. https://doi.org/10.1016/j.geoderma.2012.11.002