Journal of Hydrology 593 (2021) 125840

Contents lists available at ScienceDirect

Journal of Hydrology

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

Research papers

Root zone soil moisture estimation with Random Forest

Coleen Carranza [a,*], Corjan Nolet [b], Michiel Pezij [c,d], Martine van der Ploeg [e]

[a] Soil Physics and Land Management Group, Wageningen University, Wageningen, The Netherlands
[b] FutureWater, Wageningen, The Netherlands
[c] Department of Water Engineering and Management, University of Twente, Enschede, The Netherlands
[d] HKV, Botter 1129, 8232 JN Lelystad, The Netherlands
[e] Hydrology and Quantitative Water Management Group, Wageningen University, Wageningen, The Netherlands

ARTICLE INFO

This manuscript was handled by C. Corradini, Editor-in-Chief, with the assistance of Wei Hu, Associate Editor

Keywords:
Root zone soil moisture
Random Forest
Prediction
Process-based modeling

ABSTRACT

Accurate estimates of root zone soil moisture (RZSM) at relevant spatio-temporal scales are essential for many agricultural and hydrological applications. Applications of machine learning (ML) techniques to estimate root zone soil moisture are limited compared to commonly used process-based models based on flow and transport equations in the vadose zone. However, data-driven ML techniques present unique opportunities to develop quantitative models without having assumptions on the processes operating within the system being investigated. In this study, the Random Forest (RF) ensemble learning algorithm, is tested to demonstrate the capabilities and advantages of ML for RZSM estimation. Interpolation and extrapolation of RZSM on a daily timescale was carried out using RF over a small agricultural catchment from 2016 to 2018 using in situ measurements. Results show that RF predictions have slightly higher accuracy for interpolation and similar accuracy for extrapolation in comparison with RZSM simulated from a process-based model combined with data assimilation. RF predictions for extreme wet and dry conditions were, however, less accurate. This was inferred to be due to infrequent sampling of such conditions that led to poor learning in the trained RF model and to incomplete representation of relevant subsurface processes at the study sites in the RF covariates. Since RF does not depend on parameters required to estimate subsurface water flow, it is more advantageous than a process-based model in data-poor regions where soil hydraulic parameters are incomplete or missing, especially when the primary goal is only the estimation of soil moisture states.

1. Introduction

Root zone soil moisture (RZSM) is an important environmental variable that impacts hydrological processes relevant for agriculture and climate-related studies. It is one of the main drivers for agricultural productivity (Rigden et al., 2020) and serves as an indicator for crop water stress, which is valuable for drought monitoring (Bolten et al., 2009). Outside the hydrological cycle, RZSM dynamics play a role in quantifying soil carbon fluxes (e.g. Kurc and Small, 2007).

Accurate estimates of RZSM are necessary in order to have a better understanding of agricultural and environmental processes it controls. Direct RZSM measurements can be obtained from in situ sensors installed along the soil profile or at specific depths (Vereecken et al., 2008; Dobriyal et al., 2012). Achieving distributed spatial measurements of RZSM can be a challenge because installation of sensors at the subsurface can be a tedious task and are likely to disturb the soil properties. It has become relatively common to extract RZSM from surface soil moisture (SSM), which may be in situ or satellite-derived (Ulaby et al., 1996), since they are more easily obtained. Satellite-derived SSM has the advantage of providing spatially distributed soil moisture while in situ measurements offer higher temporal frequency (second or minutes) compared to satellites, which only provide snapshots at regular time intervals (days or weeks).

Analytical solutions are applied in cases when direct RZSM measurements are lacking or insufficient. These methods are based on theoretical or empirical relations between environmental variables controlling RZSM state. Arguably, the most common approach is to apply process-based hydrological models which are based on conceptual understanding of the system (e.g. Cordova and Bras, 1981; Porporato et al., 2004. These models employ numerical solutions of flow and transport equations in unsaturated porous media (Feddes et al., 1988). Information on soil hydraulic properties, either measured directly or

* Corresponding author at: Soil Physics and Land Management Group, Wageningen University, Wageningen, The Netherlands.
E-mail address: coleen.carranza@wur.nl (C. Carranza).

https://doi.org/10.1016/j.jhydrol.2020.125840
Received 21 August 2020; Received in revised form 30 October 2020; Accepted 30 November 2020
Available online 10 December 2020

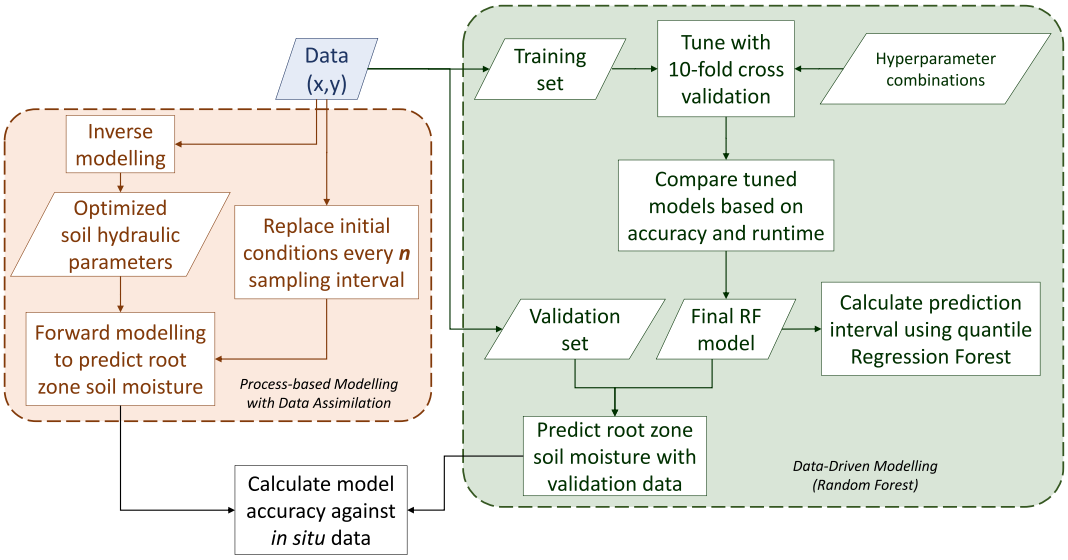

Fig. 1. Flowchart for comparison of data-driven modeling using Random Forest and process-based modeling in Hydrus-1D combined with data assimilation.

from pedo-transfer functions (Schaap et al., 2001; Van Looy et al., 2017), are required to estimate water movement across a chosen flow domain. It may be necessary to optimize soil hydraulic parameters, for instance using inverse modeling (e.g. Ritter et al., 2003), in order to improve model simulation accuracy. The prevailing meteorological conditions, as well as variables that describe vegetation growth, are necessary to determine the amount of water entering and exiting a given flow domain. In the last couple of decades, data assimilation methods have been applied to improve process-based model estimates (Houser et al., 1998; Pezij et al., 2019), which may take advantage of satellite-derived SSM information.

Data driven methods to estimate RZSM include time series analysis (TSA) and machine learning (ML) techniques. These methods aim to extract knowledge by evaluating patterns or variability in that data and further stimulate actions that are dictated by the data. In the context of RZSM estimation, data-driven methods implicitly incorporate and evaluate all the interacting processes that produced a given RZSM state. TSA methods, such as the application of an exponential filter (Wagner et al., 1999; Albergel et al., 2008), a cumulative distribution function (cdf-matching, Gao et al., 2019; Zhuang et al., 2020), or transfer-functions (Pezij et al., 2020) primarily utilize surface soil moisture data to derive a functional relation with RZSM. However, calibration of functional parameters may be necessary each time it is applied to a different study area in order to obtain high accuracy. ML algorithms build mathematical models based on training sets and covariates to extract information from data. Furthermore, they are tuned to handle diverse and large volumes of data sets, which may be relevant for large scale studies or for operational (water) management. In hydrology and climate studies, advances in ML techniques have been applied predominately for prediction and forecasting of environmental variables (e.g. Shiri et al., 2017; Ali et al., 2019; Kratzert et al., 2019), sensitivity or optimization of model parameters (e.g. Spear et al., 2020; Teweldebrhan et al., 2020), and uncertainty estimation (e.g. Shrestha et al., 2009; Kayastha et al., 2014). Application of ML in soil hydrology have also started to gain attention in the last couple of decades. For instance, ML techniques have been applied to estimate model-derived RZSM using Artificial Neural Networks (Kornelsen and Coulibaly, 2014) or satellite-derive SSM using Support Vector Machines (Ahmad et al., 2010). Furthermore, ML allows up- or downscaling of soil moisture obtained from satellite data (Srivastava et al., 2013; Zhang et al., 2017). Comparison of ML models have been made for forecasting of soil moisture using values at discrete soil moisture depths (Prasad et al., 2018) or soil layers (Matei et al., 2017) at regional scales. Interestingly, SSM has also been estimated from in situ measurements of soil moisture at deeper layers using ML (Coopersmith et al., 2016). In a comparison study, Karandish and Šimnek (2016) showed that ML may provide a useful alternative to process-based models using limited input data. ML has recently been applied to optimize soil hydraulic parameters such as the saturated hydraulic conductivity (Araya and Ghezzehei, 2019). Shiri et al. (2020) showed that ML, specifically Random Forest (RF), can accurately simulate subsurface wetting fronts due to drip irrigation in agricultural areas.

In this study, the main goal is to demonstrate the applicability of Random Forest (RF), an ensemble ML model, to estimate RZSM within a agricultural small catchment. Among the advantages of RF outlined by Tyralis et al. (2019) is that it produces consistent predictions and it reduces the variance without increasing the bias of the predictions. Furthermore, RF was selected among the multitude of ML models available in order to balance prediction accuracy with interpretability of the results with respect to the input variables. This is facilitated by investigation of the variable importance list which enumerates the covariates with the greatest influence on RF prediction accuracy. Other ML models, especially deep learning methods, have succeeded in achieving high prediction accuracies but is still challenged by interpretability and ease of relating results to model covariates (Montavon et al., 2018; Reichstein et al., 2019). So far, there are still limited studies applying RF in soil hydrology, particularly in estimations of RZSM. Applying such a data-driven method will ensure that all the processes operating in the system under study are incorporated in the predictive RF model developed. In addition, there has been proliferation of RZSM in situ measurements in the last couple of decades from various soil moisture monitoring networks worldwide (e.g. International Soil Moisture Network (ISMN, Dorigo et al., 2011) which provides an excellent opportunity to capitalize on ML techniques. In this study, an almost three year long dataset of daily measurements in agricultural fields were used for RF modeling in two ways: 1) interpolation at randomly selected points within the time series and 2) extrapolation of future RZSM state based on past values. A comparison is then made between the RF results and a process-based model in order to assess the capabilities of a data-driven method. A pore-flow model with data assimilation via direct insertion of in situ measurements was applied to simulate RZSM at the study sites.

2. Materials and methods

As an overview, Random Forest (RF) was applied for interpolation

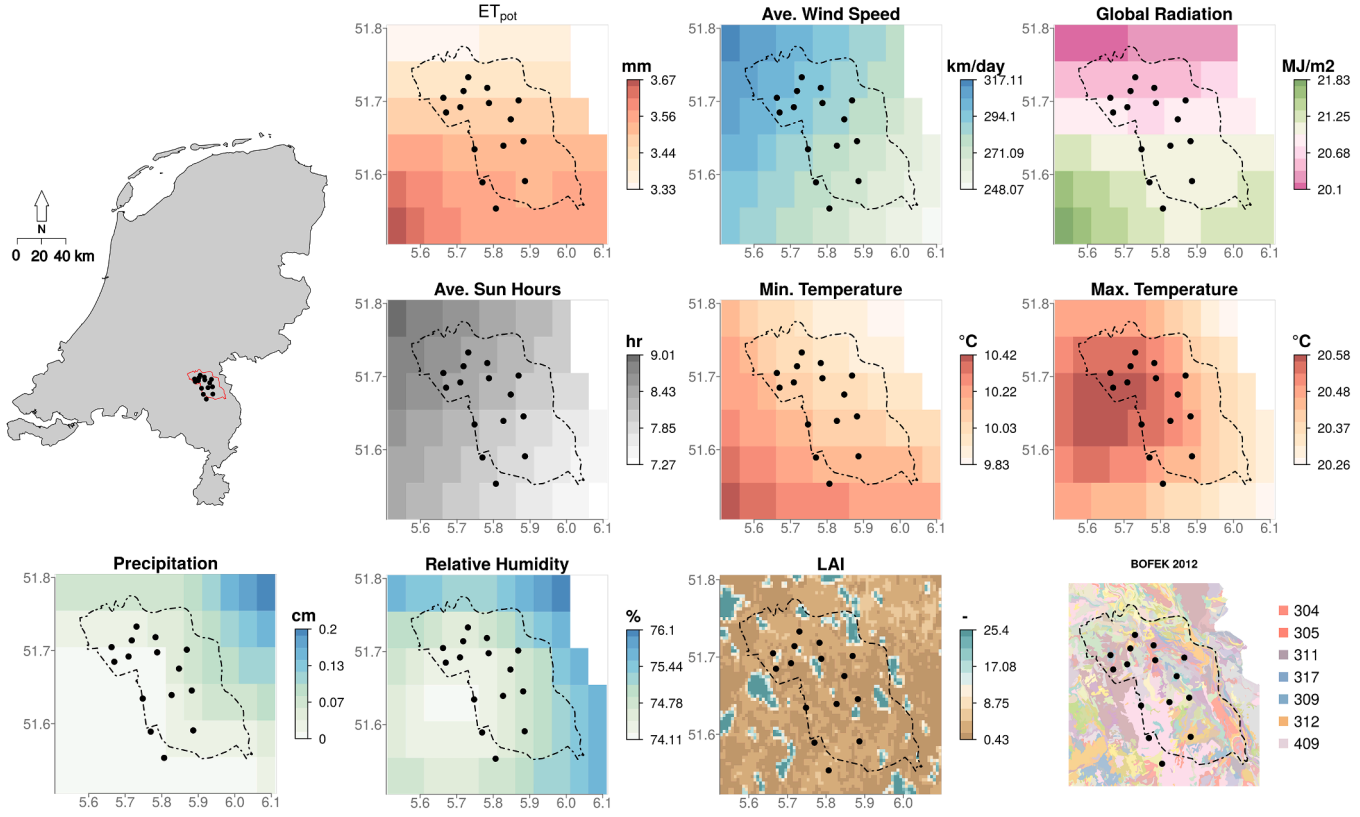

Fig. 2. Site characteristics. Top left map shows the location of the Raam catchment (red area) and soil moisture stations (black dots). The covariates for Random Forest are shown, including meteorological data, Leaf area index (LAI), and soil hydraulic group from BOFEK2012. The images shown are snapshots of the datasets on July 3, 2016, except for BOFEK2012.

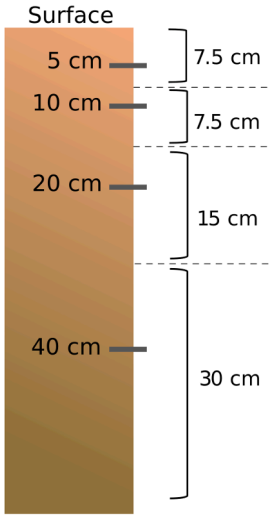

Fig. 3. Schematic diagram of installation setup at each station. Root zone soil moisture is calculated as the zone-weighted depth-average values based on the measurements and associated soil thicknesses (see Eq. (1)). For each measurement depth (5, 10, 20, and 40 cm), its associated soil thickness is based on the midway distance between two adjacent measurement points.

and extrapolation of root zone soil moisture (RZSM) within a small catchment. For comparison of results, RZSM were also simulated using a process-based (PB) pore-flow model which was combined with data assimilation. For both methods, steps to optimize model (hyper) parameters were applied to improve model performance and to allow objective comparison of the results. Briefly, a selection of hyperparameters were tuned for RF while the soil hydraulic parameters were optimized via inverse modeling for PB. Data assimilation via direct insertion of in situ measurements was further applied to improve PB model results. The succeeding sections describe in detail the datasets used and methods applied while a summary is shown in Fig. 1.

2.1. Raam soil moisture network

The Raam catchment is located in the southeastern portion of the Netherlands which holds mostly sandy soils. A total of 15 operational soil moisture stations are distributed across the whole catchment (Fig. 2). At each station, soil moisture and temperature sensors (Decagon EC-H20 5TM) were installed at 5, 10, 20, 40, and 80 cm depths and measurements were recorded every 15 min. The soil moisture stations were located in agricultural fields, which are the characteristic land cover type within the catchment area. The most common crop type at the stations is grass, followed by corn, potato, sugar beet, and other vegetable crops (Table S1). A more detailed description of the Raam soil moisture network is provided in Benninga et al. (2018).

Measurements down to 40 cm depth were integrated over a 60 cm averaging depth to calculate root zone soil moisture (Fig. 3). This was chosen in order to have a uniform root zone across the study sites which have varying crop types. For grass fields, the active root zone may only be up to 20 cm because of its shallow rooting system while for crops such as corn or potato, the root zone can extend beyond 1 m. Nevertheless, the depth used for the analysis generally captures the active root zone for the crops at the study sites. Furthermore, the methods applied in this study could also be customized for other depths that would suitably represent the root zone depths. Root zone soil moisture θ_{rz} is given by:

$$\theta_{rz} = \frac{\sum_{j=1}^{n} \theta_j \Delta z_j}{z} \quad (1)$$

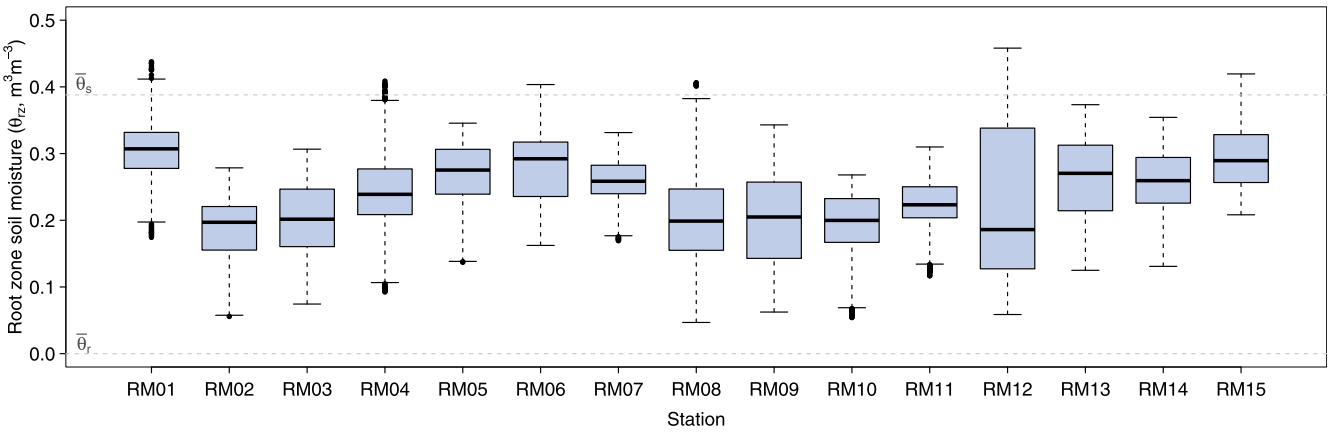

Fig. 4. Boxplots showing the distribution of root zone soil moisture (θ_{rz}) across the study sites. The average saturated ($\overline{\theta}_s$) and residual water contents ($\overline{\theta}_r$) of soils (from BOFEK2012) at the sites are indicated for comparison.

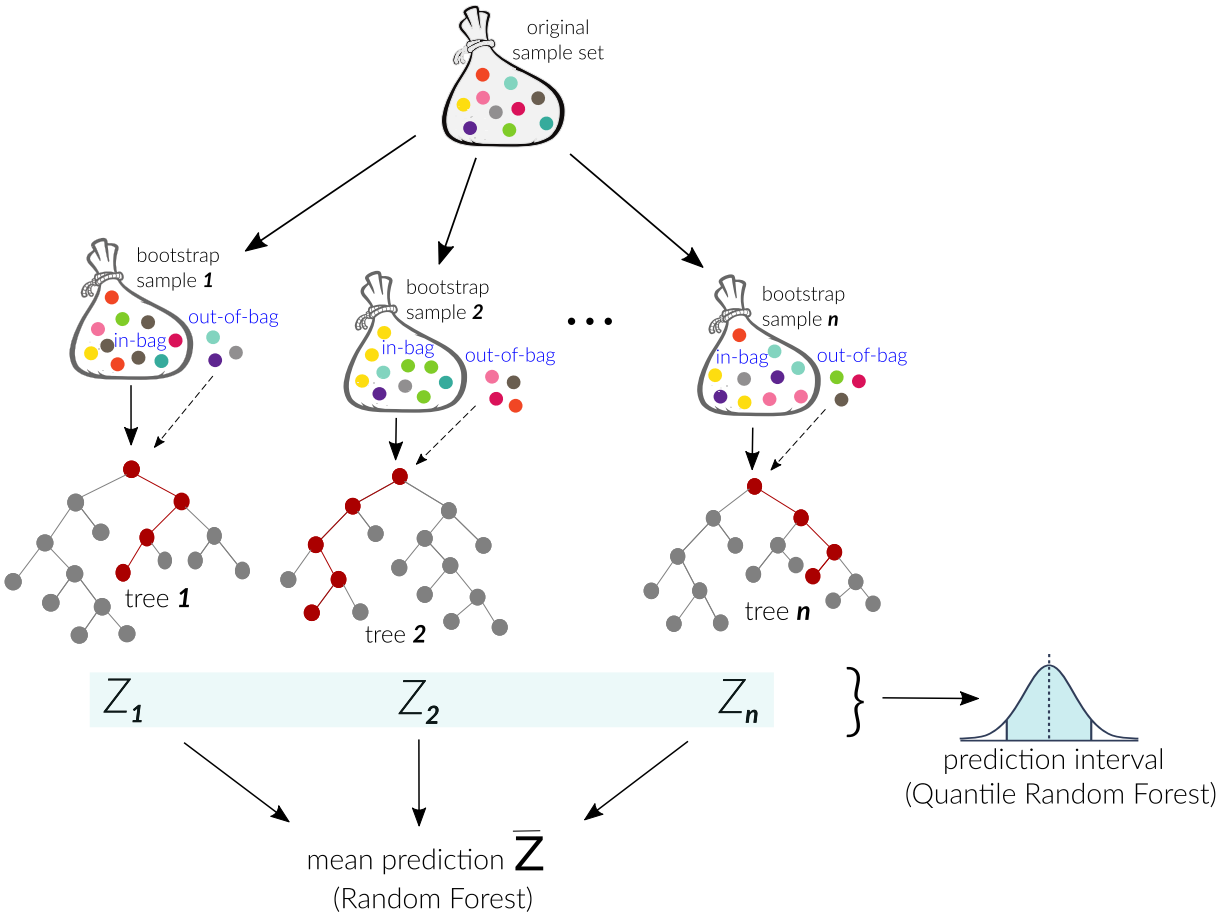

Fig. 5. Schematic diagram for Random Forest (RF). Regression trees are built based on a large number of bootstrap samples. Each tree is built by splitting the datasets at each node using randomly selected candidate variables and ends when the stopping criterion is reached. The average estimate from the regression trees is the final RF prediction. Using the full conditional distribution of estimates from all the trees, prediction interval between 2.5th and 97.5th percentiles quantify the uncertainties in the RF model.

where θ_j (in $m^3\ m^{-3}$) is the volumetric water content for measurement depth j (cm), Δz_j (cm) is the thickness of soil associated with the measurement depth, and z (cm) is the total averaging depth. Measurements from all 15 stations starting from April 2016 up to December 2018 (33 months) were used for the analysis. The daily mean values from the 15 min data were calculated in order to match the resolution of acquired meteorological datasets. Compared to the soil hydraulic parameters based on BOdemFysische EenhedenKaart (BOFEK2012, Wosten et al., 2013), which is a map of soil hydro-physical properties for the Netherlands, root zone soil moisture calculated using Eq. (1) did not reach the residual water content (θ_r) value averaged across the study sites. However, root zone soil moisture were found to reach and, for some stations, go beyond the average saturated water content θ_s based on BOFEK2012 (Fig. 4).

Table 1
Covariates used for training the Random Forest model.

Meteorological		Vegetation		Soil	
Symbol	Description	Symbol	Description	Symbol	Description
RG	Ave. Wind speed	LAI	Leaf Area Index	VWC5	Soil moisture at 5 cm
Q	Radiation	LAI_lag	1-day lag	VWC_lag	1-day lag
rd	Rainfall	Crop.grass		VWC_lag3	3-day lag
SQ	Sun Hours	Crop.corn		VWC_lag40	40-day lag
TN	Min. Temp	Crop.potato		BOFEK.305	
TX	Max. Temp	Crop.sugarbeet		BOFEK.304	
UG	Relative Humidity	Crop.wheat	Crop type (dummy)	BOFEK.311	BOFEK2012 codes (dummy)
EV24	Evapo- transpiration	Crop.onion		BOFEK.409	
RG_lag		Crop.fennel		BOFEK.317	
Q_lag		Crop.beans		BOFEK.309	
rd_lag		Crop.lettuce		BOFEK.312	
SQ_lag	1-day lag				
TN_lag					
TX_lag					
UG_lag					
E24_lag					
DOY	Day of year				

2.2. Random Forest regression

Random Forest (RF) is an ensemble-learning algorithm that combines the concepts of decision trees and bagging (Fig. 5, Breiman, 2001). Decision trees (DT), either for classification or regression, partition the variable space using a set of hierarchical rules such that the dataset are grouped recursively based on similar instances. A set of covariates (continuous or categorical variables) are used for recursively splitting the values of the variable of interest, which results in multiple parent and child nodes that resemble a tree-like structure. Splitting at each node involves random selection of candidate variables from the total number of covariates, referred to as the *mtry* parameter. The DT will evaluate each candidate to find the optimal split that maximizes the 'purity' which results in the largest decrease in the impurity at each child node. In this case, the estimated response variance for regression trees was used as a measure for impurity (Wright and Ziegler, 2015). RF creates diverse DTs to avoid highly correlated predictors by growing them from different subsets of the training data through a procedure called bagging. Often a large number of trees are created, and is referred to as the *ntree* parameter. Bagging, an abbreviation for 'bootstrap aggregation', is a technique for generating multiple training data by resampling with replacement of the original training set. This means that some data may be used more than once in the training, while others might never be used. For each bootstrap sample, a regression DT generates multiple parent and child nodes until the stopping criterion is reached. In this case, when the value for the minimum node size (*min. node size*) parameter is achieved. After all the trees are grown, the RF regression predictor is the mean from all the predictions from each individual tree. More detailed description of RF methods and parameters are given in Breiman (2001) and Hastie et al. (2009).

Estimation of RZSM using RF was implemented in two ways: 1) interpolation of randomly selected points within the whole time series data, and 2) extrapolation of 'future' RZSM based on 'past' values. For each method, a single RF model was built based on the combined measurements from all 15 soil moisture stations in the Raam. For RF interpolation, random samples were obtained from the daily time series data at each station. For RF extrapolation, the length of the time series at each station was first split based on the sampling proportion used. The first part of the time series was selected for training and constitutes the "past" data, while the remaining was used for model validation and constitutes the "future" data. Proportions of 50% up to 80% (with increments of 10%) of the daily time series measurements at each station were used to generate the samples from each station. These were then combined into one training set for building each RF model. The value of the *ntree* parameter was made proportional to the samples in each training set, and was set to a tenth of the amount of each training set. This corresponded to 600, 700, 900 and 1000 trees for interpolation and 600, 800, 900, 1100 trees extrapolation for each training set (50%, 60%, 70%, 80% of total measurements). Optimization of RF models were carried out by tuning *mtry* and *min. node size* parameters for each training set proportion tested.

2.2.1. Hyperparameter optimization

The RF model was tuned in order to select the combination of the hyperparameters *mtry* and *min. node size* that would yield the highest accuracy. Hyperparameters are parameters that need to be set prior to training a model and defines the configuration of the regression trees. Their values directly control the behaviour of the learning algorithm and have a significant effect on the performance of the model being trained. Other RF model parameters that are not tuned will simply 'learn' on their own during model training. The values for mtry will dictate splitting of RZSM values at the nodes of the regression tree while the minimum number of elements per node (*min. node size*) will serve as a stopping criterion in building the regression trees. We tested values of *mtry* from 1 to 25 and *min. node size* of 5,10, 20, and 30. A total of 100 combinations of hyperparameters (*mtry* and *min. node size*) were tested for each of the four training set proportions (50% to 80%) in the tuning phase. A 10-fold cross-validation (CV) scheme was applied to the training set in tuning the hyperparameters. This meant that the training set was split and for each *k*-fold, a 10th of the training set served as the test sample while the remaining is used for creating the regression trees. For each *k*-fold, the test sample contains randomly chosen points from the total training set that has not yet been included as a test sample before in previous folds. The mean root mean square error (RMSE) computed for each hyperparameter combination in the 10-fold CV scheme were compared to assess model performance (i.e. 100 RMSE's). Aside from having a separate validation set, a CV scheme is a preventive measure for model overfitting (Lever et al., 2016).

RMSE's were examined further to select the final RF model as the one with highest accuracy ('best model') might also be computationally expensive. Therefore, we compared the model with the best RMSE to another one that has a faster computation time but comparable RMSE as a 'tradeoff' to evaluate a simpler model without sacrificing accuracy. RMSE's were first ranked from lowest to highest and then a pairwise elimination process was applied by evaluating the improvement in RMSE. The final 'tradeoff' model was selected once a $< 1\%$ improvement in RMSE was found.

2.2.2. Random Forest covariates

Covariates or the set of predictor variables used to build RF regression trees include information on meteorological conditions, soil properties, land cover and vegetation characteristics at each site (Fig. 2 and

Table 1). Daily meteorological data from 36 KNMI (Royal Dutch Meteorological Institute) stations distributed within the entire Netherlands were interpolated to produce a 5 x 5 km gridded image in order to extract daily values at the location of each soil moisture station. Spatial estimates of the values at each KNMI station were obtained using a Thin Plate Splines interpolation (Sluiter, 2012). Temperature, wind speed, relative humidity, sun hours, potential evapotranspiration and radiation were selected among the total meteorological datasets available, as these were also the input variables in the process-based model applied in this study. They are, therefore, indicative of surface processes that influence the RZSM state. Gridded values of daily rainfall measurements with a 1 x 1 km pixel size were obtained directly from KNMI. Ordinary kriging was applied to around 300 measurement locations of rain gauges distributed across the Netherlands to produce the rainfall maps (Soenario et al., 2010). Leaf area index (LAI) from an 8-day MODIS composite with 500 m resolution was used to capture vegetation characteristics over the study sites. The values for days in between LAI measurements were linearly interpolated to obtain daily estimates. Both crop type and soil hydro-physical groups were also included as categorical covariates. The former is based on field observations while the latter was obtained from BOFEK2012. Fig. 2 indicates the BOFEK2012 codes for the soils within the Raam network only. Further description of each code is given in Wosten et al. (2013). These two categorical variables are re-coded into dummy or indicator variables for the RF regression. The categorical variables are transformed into a dichotomous (1 or 0) representation of its presence or absence for each data point. For example, the categorical variable "Crop" with a type "Corn", a value of 1 is assigned for measurements having the said crop, and 0 for measurements with another crop type.

The current soil moisture state is inevitably affected by its past values and past meteorological conditions. The so-called soil moisture memory (or persistence) has been widely investigated because of its importance in climate-related studies (e.g. Koster and Suarez, 2001). Therefore, lagged values were also calculated for meteorological and soil moisture datasets in order to incorporate past information in the RF model. This may be useful especially for forecasting where only past information is available. For surface soil moisture and meteorological datasets, values with a lag of 1 day were obtained. Additional lagged surface soil moisture values of 3 and 40 days were also calculated based on findings of soil moisture memory studies at global (McColl et al., 2017) and continental (European, Orth and Seneviratne, 2012) scale, respectively. A total of 39 covariates were used for the RF models (Table 1).

2.2.3. RF prediction intervals

Uncertainties in RF estimates are defined based on the 95% prediction interval (PI) obtained using quantile regression forest (qRF, Fig. 5). The idea behind qRF is that instead of recording the mean value of response variables from the trees, all responses for each tree are recorded (Meinshausen, 2006). This allows not only for the estimation of the conditional mean but also a good approximation of the full conditional distribution. PIs were defined using quantile regression based on the chosen quantiles (α's). For a given random variable, the conditional distribution function $F(y|X = x)$ is given by the probability that, for $X = x, Y$ is smaller than y. For a continuous distribution function, the α-quantile ($Q\alpha(x)$) specifies a value such that the probability of x being smaller than $Q\alpha(x)$ is, for a given random variable $X = x$, exactly equal to α. A 95% PI (I_{95}) for the RZSM estimates is based on 2.5% and 97.5% quantiles ($[Q_{.025}(x), Q_{.975}(x)]$).

2.2.4. Variable importance

Variable importance from the RF models determined using a permutation method (Wright and Ziegler, 2015). Rankings for covariates were based on the mean decrease in model accuracy after shuffling or randomly permuting the values of a predictor X_i, where $i = 1 \ldots n$ for each of the covariates used. By permuting the values of X_i, its association with the response variable Y (i.e. RZSM) is broken. Therefore, if the predictor X_i is associated with the response Y, a substantial decrease in accuracy is expected after prediction using the permuted and remaining non-permuted variables.

2.3. Process-based modelling with data assimilation

A soil water balance model was carried out to simulate one-dimensional daily RZSM at the study sites. In a water balance model, mass and energy fluxes over time and/or space are calculated to estimate soil moisture along the profile. We assumed that soil water movement would be restricted along the vertical dimension since the study sites are generally characterized by homogeneously textured soils and the terrain at the study sites is generally flat [e.g.][for modeling unsaturated flow in the Netherlands] (De Laat, 1980). The vertical water flow in unsaturated porous media is solved numerically using Richard's equation:

$$\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial Z}\left[K(h)\left(\frac{\partial h}{\partial Z} + 1\right)\right] - S \tag{2}$$

where t is the time (days), θ is the volumetric water content ($cm^3\ cm^{-3}$), h is the soil water pressure head (cm), Z is the spatial coordinate (cm) defined as positive upward, $K(h)$ is the unsaturated hydraulic conductivity function ($cm\ d^{-1}$) and S is a sink term representing water uptake by plant roots ($cm\ d^{-1}$). $K(h)$ is derived from a water retention curve, given by van Genuchten (1980):

$$\theta(h) = \frac{\theta_s - \theta_r}{[1 + (\alpha h)^n]^m}, h \leqslant 0$$

$$K(h) = K_s S_e^l \left(1 - \left(1 - S_e^{l/m}\right)^m\right)^2$$

$$m = 1 - \frac{1}{n} \tag{3}$$

$$S_e = \frac{\theta - \theta_r}{\theta_s - \theta_r}$$

where θ_r and θ_s denote residual and saturated volumetric water contents ($cm^3\ cm^{-3}$), respectively; α (cm^{-1}) and n (–) are fitting parameters of soil water characteristic curve; K_s is the saturated hydraulic conductivity ($cm\ d^{-1}$); l (–) is the pore connectivity parameter; and S_e (–) is the relative saturation.

2.3.1. Inverse modeling for parameter optimization

The soil water balance was carried in two parts using Hydrus-1D software (Simunek et al., 2005). The first part involved optimization of soil parameters describing the shape of the water retention curve (θ_s, θ_r, α, n) and hydraulic conductivity curve (K_s, l) using inverse modeling. We initially carried out simulations using soil hydraulic parameters available from (BOFEK2012, Wosten et al., 2013), but found the results to be unsatisfactory. Optimization of soil hydraulic parameters was based on Marquardt–Levenberg parameter estimation method (Marquardt, 1963) as implemented in Hydrus-1D, using soil water content measurements. The soil domain considered is 1 m to cover depths similar to the measurements stations. A variable atmospheric condition, based on rainfall and evaportranspiration, was set as the upper boundary conditions while a free drainage condition was set as the lower boundary conditions. Daily meteorological datasets from KNMI, as described in Section 2.2.2, were used for the upper boundary conditions. Initial conditions for the inverse modeling were set to the pressure head at field capacity with the assumption that the soil is close to saturation the start of a year when the simulations commenced. In addition, simulations from January until April, just before the start of the in situ measurements, were part of the spin-up period for the model. A single porosity van Genuchten – Mualem model without hysteresis was used for the simulation. The flow domain was subdivided based on the number of soil layers present in BOFEK2012 (Table S2). For instance,

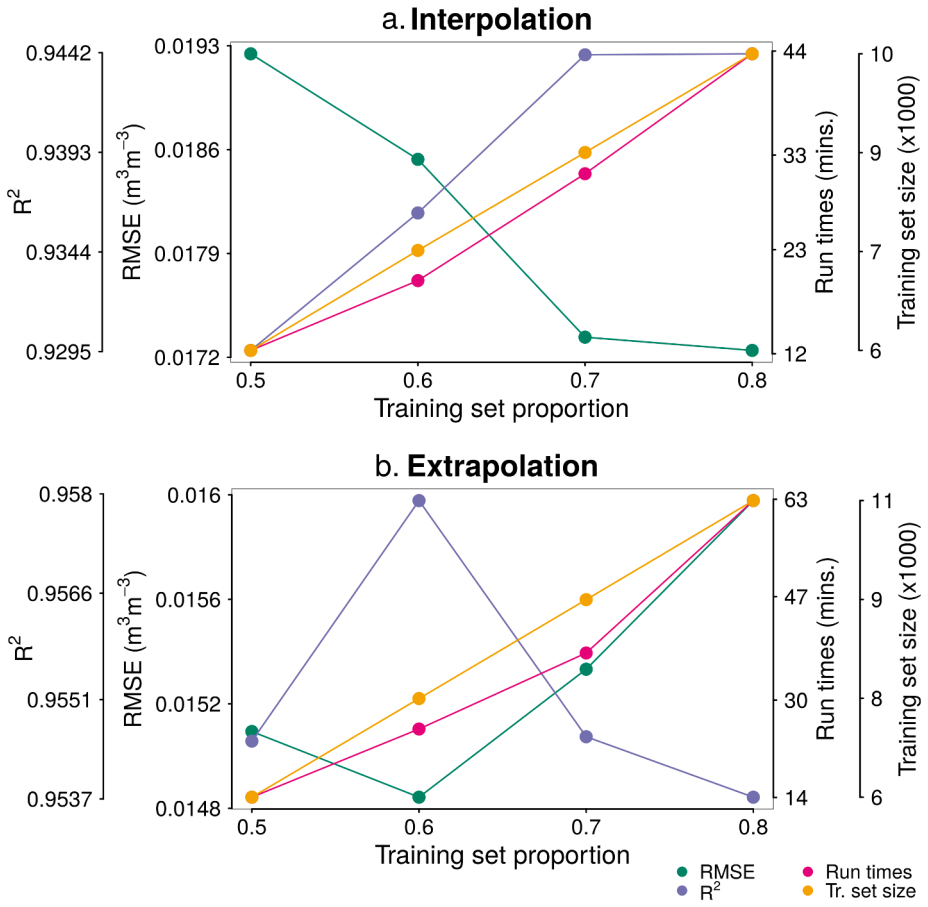

Fig. 6. Accuracy metrics and RF model specifications for different training sets tested.

station RM02 was subdivided into two layers while RM15 has only one layer in the flow domain. Simunek et al. (2005) provides more detailed information regarding the theory, methods and default parameters in H1D software. Subsequently, the second part was the forward modeling to estimate soil water content using the optimized set of soil hydraulic parameters. Observation points within the flow domain were selected at the same depths as the in situ measurement points. Furthermore, depth-averaged zone-weighted root zone soil moisture was calculated in a similar manner as the in situ values.

2.3.2. *Sequential data assimilation*

Data-assimilation is an often applied method to improve the accuracy of hydrological modelling using up-to-date measurements. The goal of data assimilation is to combine measurements and modelling efforts into an optimal state estimate of the variable of interest (Reichle, 2008). The difference between machine-learning and data-assimilation is that the latter depends on a dynamical model of the system, in this case H1D. To show the added-value of the machine-learning method, we show an application of data-assimilation with the modelling instrument used in this study. We recognize that data-assimilation should use information on uncertainties in both observations and modelling efforts, for which sequential methods such as the Ensemble Kalman Filter can be used (Evensen, 2009; Houtekamer and Mitchell, 1998; Pezij et al., 2019). However, in this study we only focus on a simple data-assimilation method, which is relatively easy implemented. Houser et al. (1998) and Heathman et al. (2003) showed the value of direction insertion for soil moisture modelling. Therefore, we applied a direct insertion data-assimilation method to update the soil moisture state.

We applied direction insertion by replacing the model state by the in situ measurements for every 20-day interval over the whole simulation period which covered two years. Measurements along the entire soil profile were used. At the end of each 20-day period, the model state was replaced by the soil profile provided by the in situ measurements. The model was subsequently run for the next 20-day period. The said assimilation interval was tested as it approximated the revisit times of some microwave satellites (e.g. Radarsat-2 or ALOS PALSAR-2) which have been assimilated into process-based models in the past. The days when data were assimilated were excluded for model evaluation.

3. Results and discussions

3.1. *Random Forest model tuning*

The RF models generated using different training sets indicate that the highest and lowest RMSEs are based on 50% and 80% of the total data for interpolation and 80% and 60% of the total data for extrapolation (Fig. 6a and b). However, a 50% training set performed relatively well in both cases based on in very minimal decrease in the RF model performance; RMSEs are only 0.002 $m^3\ m^{-3}$ and 0.0003 $m^3\ m^{-3}$ higher than those obtained using 80% and 60% training proportion for interpolation and extrapolation, respectively. Furthermore, the runtime is fastest with a 50% training set, decreasing the computation time of the best performing model by at least 44% (e.g. from 25 to 14 min for extrapolation). This aspect is of importance for machine learning techniques, especially as the volume of datasets become larger. Therefore, we selected the RF model from a 50% training set for further evaluation of the hyperparameter tuning results. Using the 50% training automatically meant that *ntree* = 600 was implemented in the RF models, for both interpolation and extrapolation.

A comparison of the results obtained from the tuning process using

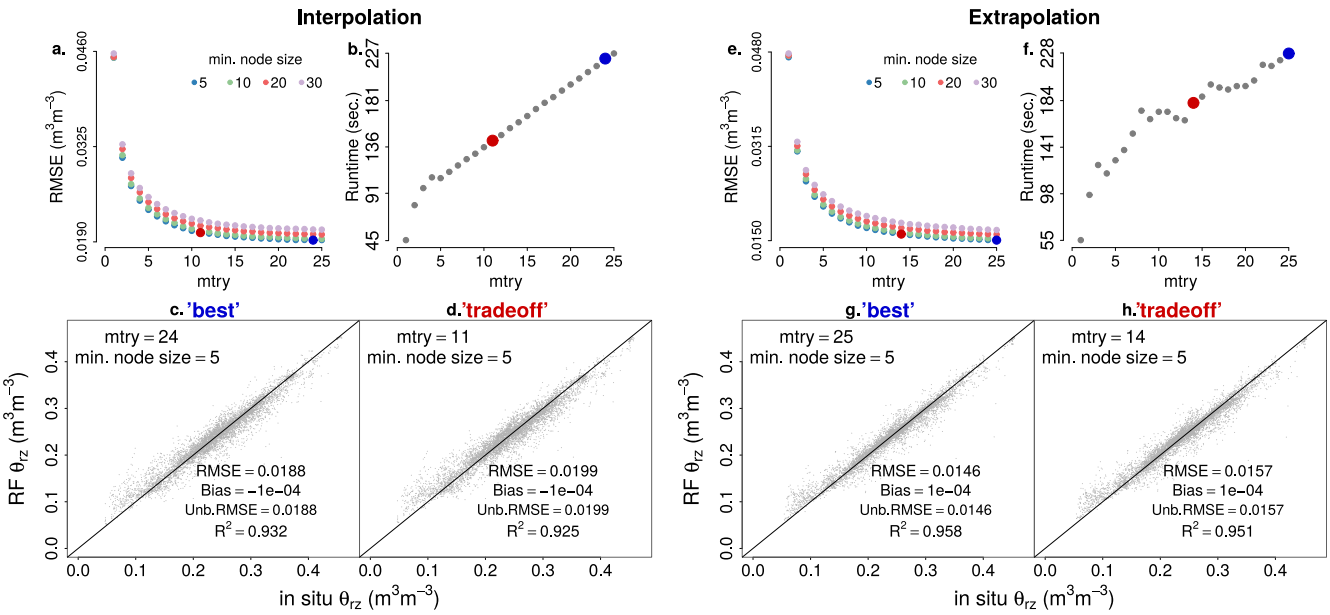

Fig. 7. Random Forest model tuning results using 50% training proportion. *a&c*: RMSE's for all hyperparameter combinations. *b&f*: model runtimes. The 'best' (red dot) and 'tradeoff' (blue dot) models are highlighted. Scatterplots (*c,d,g,h*) with corresponding accuracy metrics show the differences between the 'best' and 'tradeoff' models. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

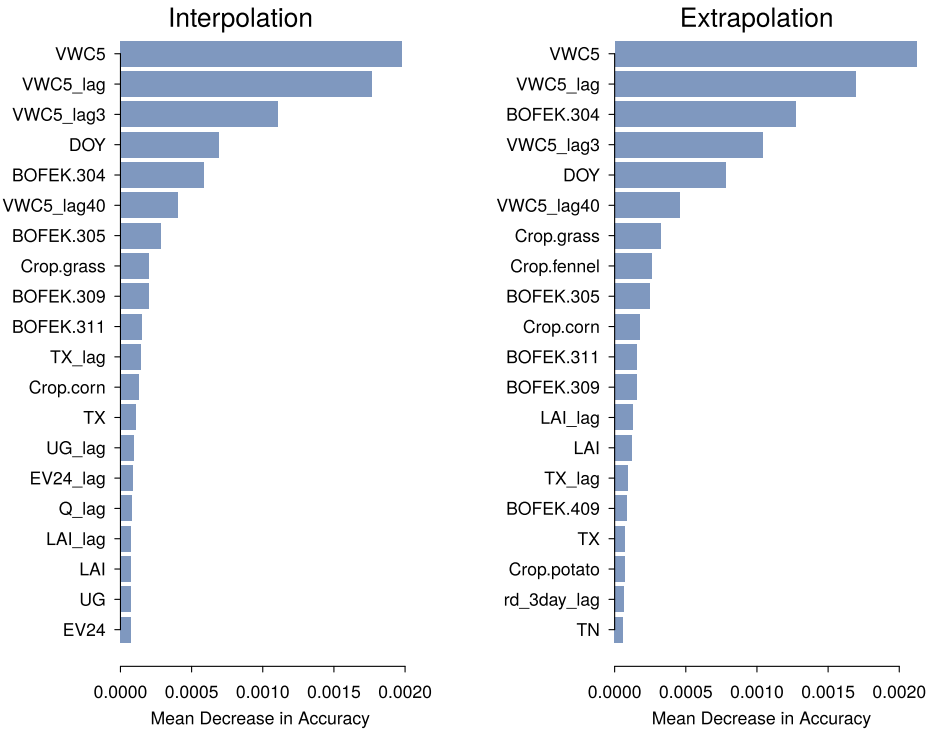

Fig. 8. Variable importance from the RF models.Only the top 20 VI's are listed for presentation purposes. The description of the variables are provided in Table 1.

50% training set is given in Fig. 7. The RMSEs are observed to exponentially decrease with increasing *mtry* values, which is combined with a consistent increase in accuracy with smaller *min. node size* (Fig. 7a and e). For both interpolation and extrapolation, a large *mtry* and a small *min. node size* resulted in the best RF model based on RMSEs (Fig. 7c and g). This is somewhat expected because the homogeneity of elements at each node is higher when the *min. node size* value is kept smaller. In addition, the largest *mtry* costs the most computing time, as expected. In contrast, the 'tradeoff' model with a smaller *mtry* value halves the computing time (from 25 to 10 s.) but only has a slightly lower RMSE (bottom panel). To balance accuracy and computation time, the hyperparameters from the 'tradeoff' model were used further for model evaluation.

3.1.1. *Variable importance*

Based on Fig. 8, surface soil moisture (SSM), soil properties and land cover types have larger impacts on RF model accuracy compared to meteorological variables. Lagged soil moisture values appear higher on the list of important variables (VI) in the RF model. For both interpolation and extrapolation, SSM with lags of up to 40 days are still highly

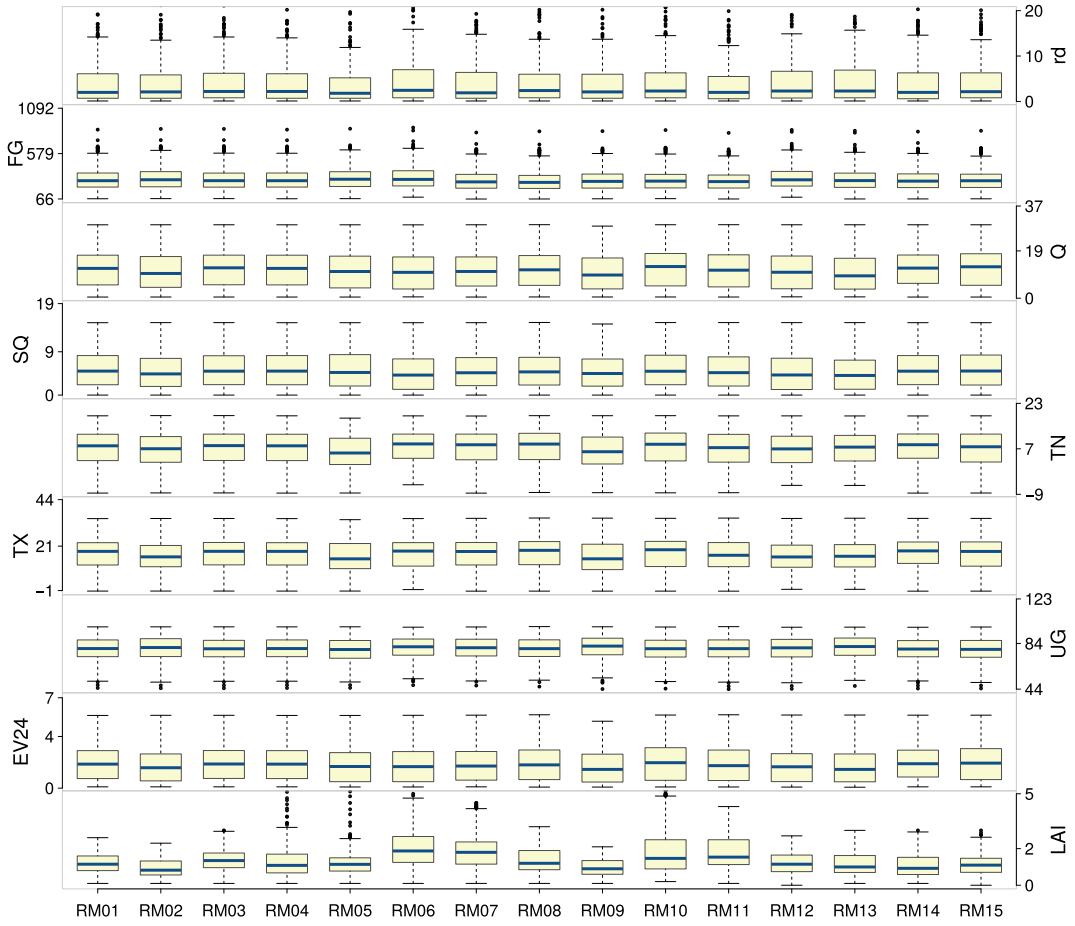

Fig. 9. Comparison of values of temporally varying covariates among the different stations. The description of the variables are given in Table 1.

relevant in estimating RZSM. Temperature appears to have the most important effects on RF model performance among all the meteorological covariates included. Since soil moisture is directly influenced by precipitation, it is surprising that current and antecedent rainfall did not rank higher in the VI list. Rather, the impact of precipitation on the RF model may be included within the SSM variables (VWC5 and VWC5_lag), which ranked highest in the VI list. Within the Raam catchment, meteorological conditions among the Raam stations were found to be similar due to its small areal coverage (Fig. 9). Since RF capitalizes differences or unique values in the covariates to separate RZSM into groups in building the regression trees, meteorological variables over the stations may have been less able to differentiate between RZSM among different stations based on the single RF model trained (i.e. one each for interpolation and extrapolation). We hypothesize that the influence of meteorological conditions may also be encapsulated within the DOY variable, which represents effects of seasonal changes on RZSM. The combination of DOY and temperature may have been adequate for the RF models to estimated RZSM in the Raam catchment. The results obtained, however, do not imply that meteorological variables are not important controls for RZSM. In this case, meteorological variability over the Raam catchment were secondary to variability in crop types and soil characteristics among the stations for estimating RZSM using the single RF model developed. Perhaps representing all the meteorological variables into one (or two) variables via dimensionality reduction methods (e.g. Principal Component Reddy et al., 2020), allows for their impact to be interpreted collectively and may potentially result in a different VI ranking for a consequent RF model.

3.2. Root zone soil moisture estimation in the Raam catchment

The RF model performance for stations with the best and worst accuracy obtained from RZSM interpolation and extrapolation are given in Fig. 10. Results for all stations are given in the supplementary (Fig. S1 & Fig. S2). RF interpolations (Fig. 10a and b) for RZSM have high accuracy in comparison to RF extrapolations (Fig. 10c and d). However, the soil moisture dynamics (i.e. an increase or decrease) are still captured in the extrapolated values, even though soil moisture state may be over- or underestimated. The accuracy of Hydrus-1D (H1D) simulations are generally lower than RF interpolations but are comparable with RF extrapolations (Table 2). For instance, the values from H1D simulations are closer to in situ values at the station with the worst performing RF extrapolations (RM02).

The results from RF generally have high R^2 (> 0.75) and low RMSEs (> 0.06 m^3 m^{-3}), indicating the capability of a data-driven method to accurately estimate RZSM. They are comparable, or may even be better than those from H1D simulations, which further adds weight to the utility of the RF model applied. Differences in accuracy between RF interpolation and extrapolation could be related to the impact of the training samples used to build each respective RF model. Higher accuracy for RF interpolations have resulted from inclusion of most, if not all, of the possible RZSM conditions within the Raam catchment using the randomly selected training set. This may not be the case for the RF extrapolation model trained, which consequently contributed to lower accuracy in the validation set. The 'past' data used to build the RF extrapolation model may exclude some of the meteorological or soil moisture conditions possible in the Raam catchment. Therefore, 'future' soil moisture conditions that are not represented in the training set are 'unseen' or 'foreign' values to the RF model, and are more likely to be

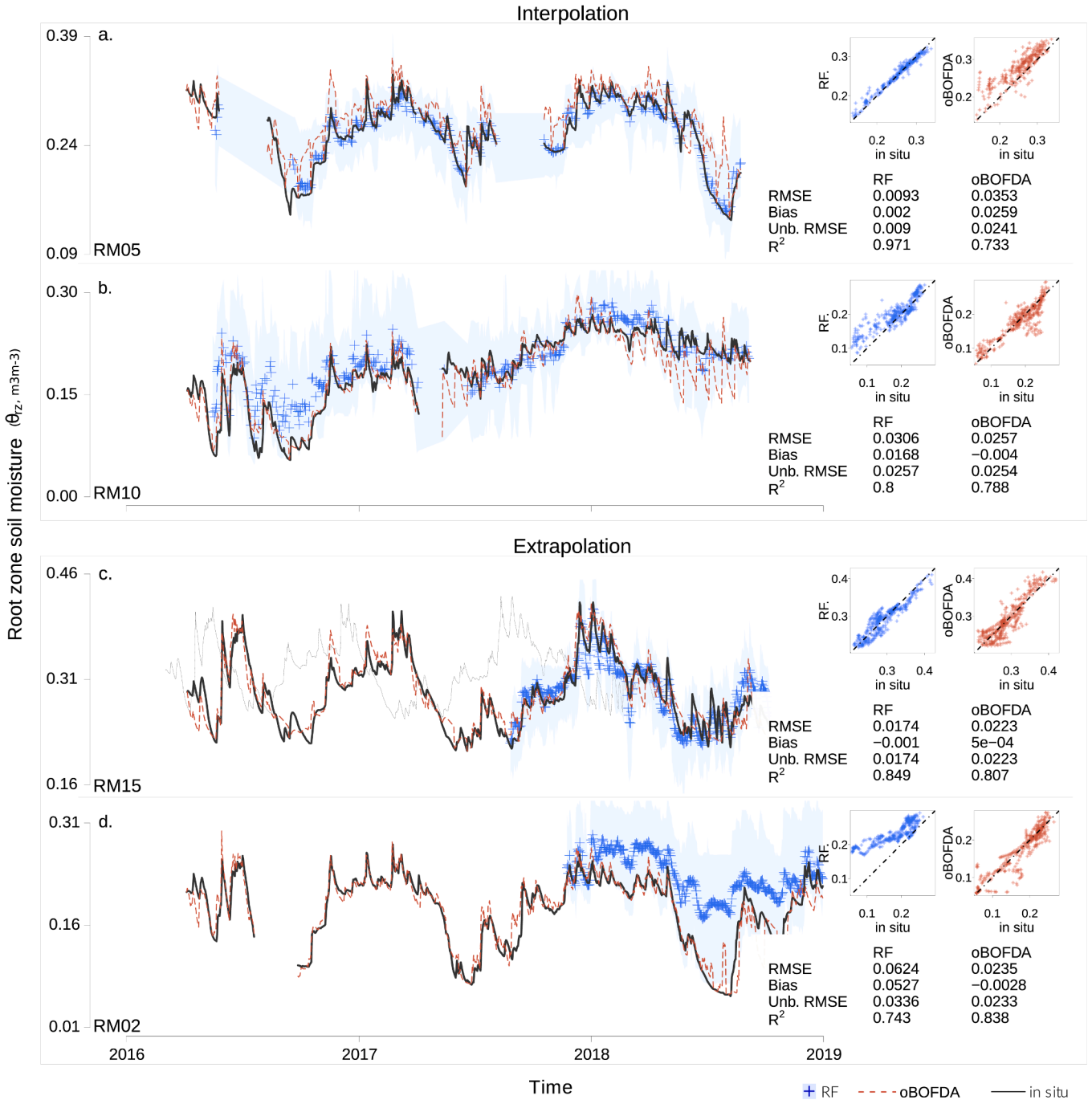

Fig. 10. Time series plots of Random Foredt (RF) estimates (blue +) for stations with the lowest (*a* and *c*) and highest (*b* and *d*) RMSEs, and corresponding prediction intervals (blue bands). H1D simulations with data assimilation are plotted as brown dotted lines. Scatteplots and accuracy metrics (right) compare model vs. in situ values. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Accuracy metrics for RF and Hydrus-1 (H1D). The range of values encountered from all 15 stations in the Raam network are reported. Except for the unitless R^2, the metrics are expressed in $m^3\ m^{-3}$.

	Interpolation		Extrapolation	
	RF	*H1D*	*RF*	*H1D*
RMSE	0.0097–0.0313	0.0185–0.0507	0.0168–0.0621	0.0201–0.0544
Bias	−0.0128–0.0178	−0.0204–0.0259	−0.0235–0.0526	−0.0232–0.0219
Unb. RMSE	0.0095–0.0263	0.0175–0.0506	0.0167–0.0422	0.0192–0.0542
R^2	0.7985–0.9730	0.6829–0.8652	0.6821–0.9611	0.4030–0.8443

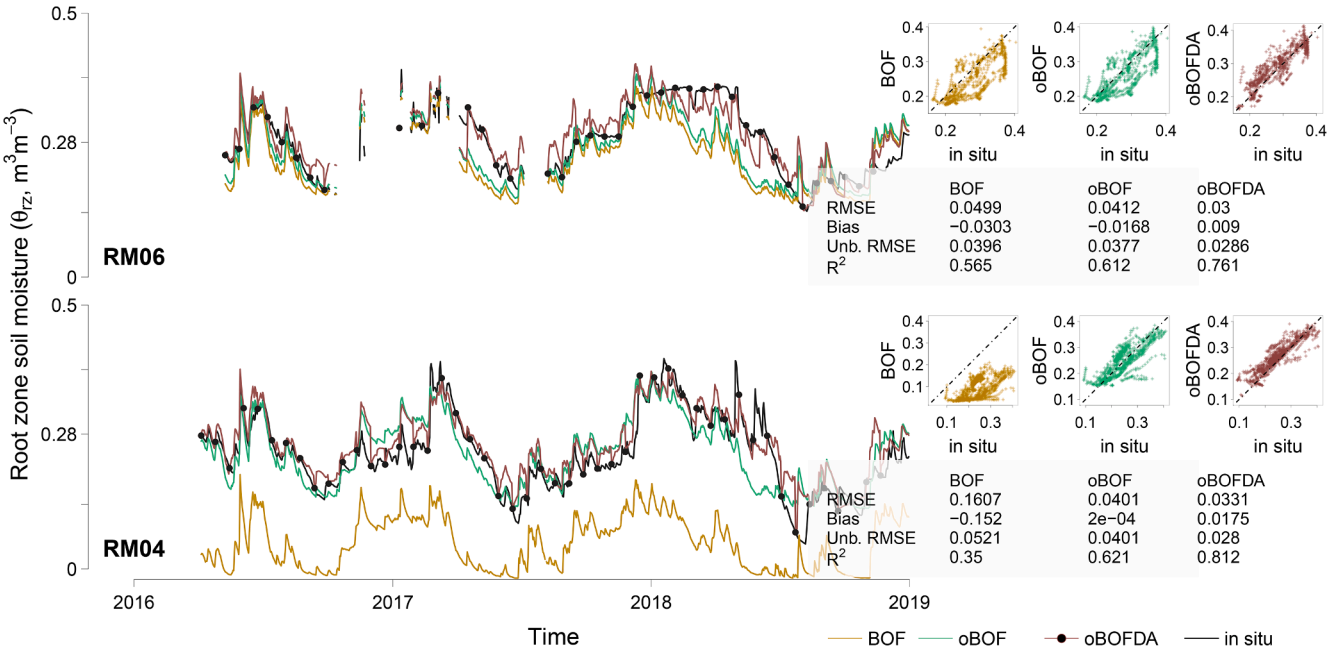

Fig. 11. Hydrus-1D simulations at the stations with the lowest (top) and highest (bottom) RMSE's based on soil hydraulic parameters from BOFEK2012 (yellow). Results using optimized soil hydraulic parameters (green) as well as those from data assimilation (DA) via direct insertion of in situ measurements (red) are plotted for comparison. Black dots represent data assimilated at every 20-day sampling interval. Scatterplots and accuracy metrics (right) compare model vs. in situ values. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

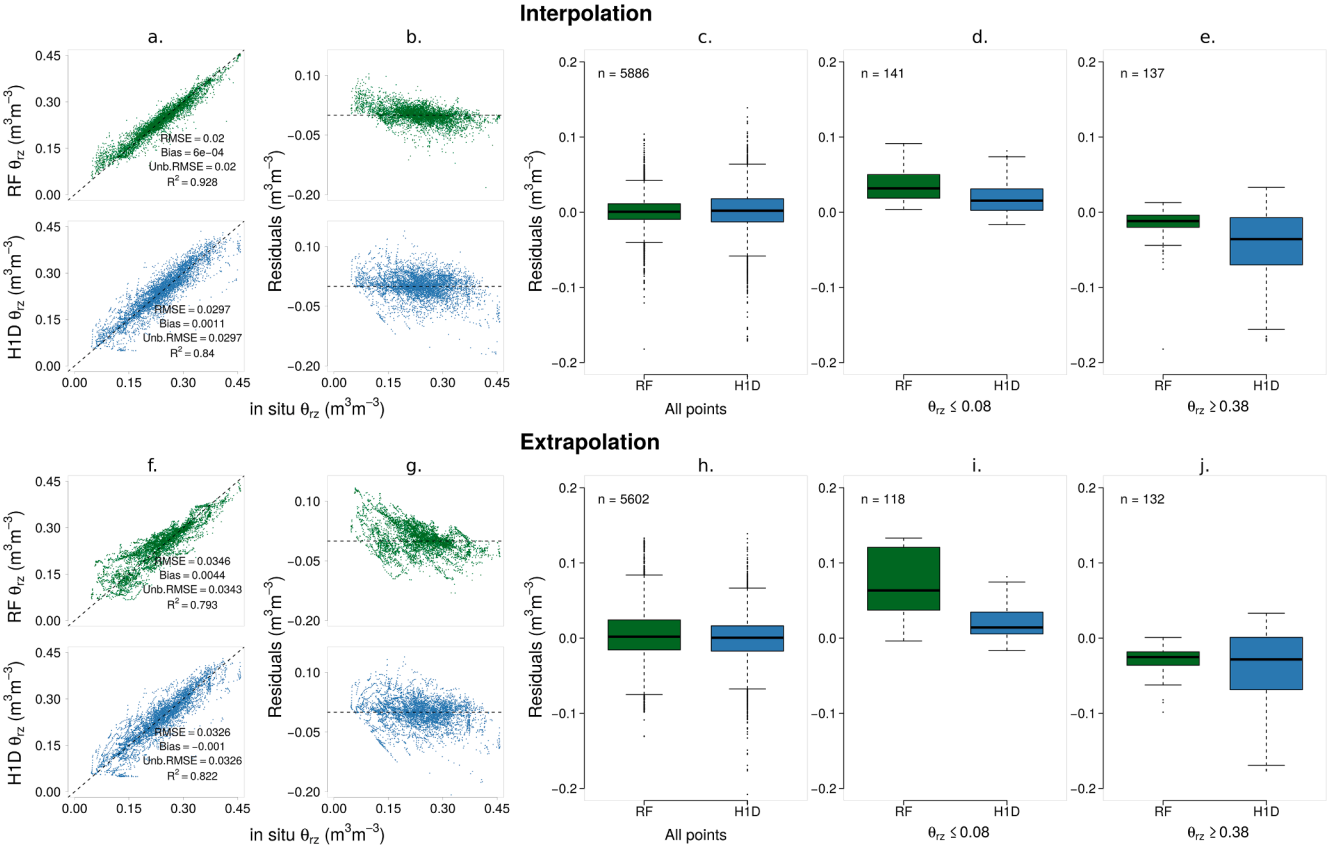

Fig. 12. Root zone soil moisture predictions from RF and H1D. *a&f*: Scatterplots of predicted vs. in situ root zone soil moisture. *b&g*: Residual scatter plots based on *a&f*. Boxplot showing the distribution of residuals for all soil moisture conditions (*c&h*), extreme dry (*d&i*), and extreme wet (*e&j*) conditions. Extreme conditions are based on the 2.5th and 97.5th percentiles (≤ 0.12 m^3 m^{-3} and $\geqslant 0.38$ m^3 m^{-3}) of the total dataset distribution, respectively.

poorly estimated. Poor extrapolation of values outside the training set is a known drawback of RF and other similar ML techniques (e.g. Hengl et al., 2018). This can be resolved by inclusion of the full suite of soil moisture conditions and corresponding covariates in the training dataset. However, this may not be always possible from in situ measurements since not all soil moisture conditions are encountered in the field within a short time span (<5 years). Remote sensing is seen as an additional source of SSM or RZSM information provided that the spatio-temporal resolutions from satellite images matches the intended scale of study. Another potential complementary dataset are those simulated by process-based models, especially for extreme meteorological conditions that are not encountered during field measurements. Inclusion of extreme meteorological conditions, modeled by process-based models, may potentially resolve the range of soil moisture values missing from in situ measurements alone. However, process-based model outputs should also demonstrate acceptable to high accuracy levels in order to be used as inputs.

For the H1D simulations, data assimilation (DA) by direct insertion (DI) of in situ measurements improved the root zone soil moisture estimates. H1D simulations at the stations with the highest and lowest RMSE's both show improvement after DA (Fig. 11). Simulations based on soil hydraulic parameters from BOFEK generally underestimated RZSM values for all the study sites. A substantial increase in model accuracy is obtained after optimizing the hydraulic parameters. Further improvement in the simulation accuracy were obtained after applying DA. Using the DI approach, model estimates are pushed towards the observations. However, for some stations, a large spike or drop in RZSM estimates are observed immediately after data was assimilated into the simulation. The model reverts back to the original state quite quickly, which could either imply a suboptimal DA sampling interval or that the model physics and/or parameters may not be completely adequate in explaining the measurements. Such effects can be mitigated by applying other types of DA (e.g. Ensemble Kalman Filter), which allow continuous estimation of model uncertainties. However, despite some limitations of the DI method selected, the primary goal was to demonstrate the improvement in model accuracy, and therefore, other (more complex) DA methods were not further pursued.

3.3. Model residuals for extreme soil moisture conditions

Residuals of the model estimates against in situ measurements from all stations in the Raam were further assessed to compare the model performance for extreme RZSM conditions. Accurate RZSM estimates of extremes conditions are vital in understanding the environmental impacts of climate or extreme climatology. In contrast to a single overall metric provided by RMSE or R^2, residuals allow investigation of specific RZSM values that are poorly estimated using the two methods applied. The results from the residual analysis generally reflect the accuracy obtained for RF (both inter- and extrapolation) and H1D, as described in the previous section. The range of residual values are smaller for RF interpolation and higher for RF extrapolation (Fig. 12c and h). However, based on the residuals, less accurate estimates are found towards drier and wetter soil moisture values from both RF and H1D.

The variability in the residuals for extreme conditions representing the 2.5th and 97.5th percentiles (equivalent to ≤ 0.08 m^3 m^{-3} and $\geqslant 0.38$ m^3 m^{-3}) from the total dataset distribution are given as boxplots in Fig. 12(*d,e,i,j*). For the two RF models, extreme dry conditions tend to be overestimated while extreme wet conditions tend to be underestimated, based on larger than zero residuals for the former and smaller than zero residuals in the latter. The degree of over- or underestimation is larger for RF extrapolation than RF interpolation. Furthermore, H1D simulations have smaller residuals than RF for extreme dry conditions but have worse estimates for extreme wet conditions.

Since extreme conditions represent only a small proportion of the total dataset, the probability of being excluded from the bootstrap samples used for building the regression trees is higher than other frequently encountered soil moisture values. This may have resulted to poor learning of the RF model, which is clearly demonstrated in the large residuals for RF extrapolations of extreme dry conditions. RF extrapolations, with a median of 0.05 m^3 m^{-3}, mostly overestimated extremely dry conditions and are worse than those from H1D simulations which had a median close to zero.

Aside from the impact of the frequency of extreme conditions to the bootstrap samples, large residuals obtained for RF extrapolations and H1D (Fig. 12*d,e,i,j*) may be related to the covariates used in the former and the type of flow model applied in the latter. Since only a pore-flow model was applied for simulation root zone soil moisture using H1D, the impact of preferential flow was excluded in the analysis. Preferential flow paths generated by biotic activity (plant roots or animals) are likely to be present at the study sites. However, additional model parameters for incorporating preferential flow are not readily available for the study sites and would require separate investigation. Arguably, pore-flow models are still most widely implemented for practical applications of process-based models. Migration to a framework that routinely incorporates preferential flow might be necessary for modelling at spatial–temporal scales where its impact are substantial. Similarly, covariates used in RF also dominantly reflect processes that are important for simulating pore-flow. As mentioned in Section 2.2.2, they were chosen based on the knowledge that they are inputs for the process-based model. Underestimation of the two RF models for extreme wet conditions indicates that the covariates selected may have been insufficient to achieve higher accuracy for those conditions. Addition of covariates that directly represent or indicate the likelihood of occurrence of preferential flow paths in the soil could potentially be beneficial for the RF model. However, deriving such covariates remain elusive since there are still theoretical and technological bottlenecks in understanding preferential flows in soils (Guo and Lin, 2018) that hamper accurate quantification and representation in spatio-temporal maps with resolutions suitable in this study.

3.4. Utility of data-driven methods for RZSM estimation

Comparison for the results from RF and H1D show that both methods are equally able to accurately estimate RZSM, although they operate differently. On the one hand, process-based models determine the rate of water movement along the soil profile which always require soil hydraulic properties. One the other hand, ML methods such as RF performs focus on patterns that allow hierarchical splitting of the dataset using suitable covariates. For both methods, techniques are available in order to optimize and improve naively implemented models that can elevate accuracy to acceptable levels. The question of utility for different scenarios or applications therefore arises. In other words, what are the advantages/disadvantages of one over the other, and how does this affect model selection for a certain application? For RF, one the of the advantages of a data-driven method is its ability to create a single model that will fit very large datasets without any assumption on the system dynamics. An RF approach may be attractive for areas with limited information on soil hydraulic properties because it can be applied using easily obtained meteorological and satellite-derived variables. Process-based models maybe applied over large areas in a spatially distributed manner but they need to explicitly account for heterogeneity in soil properties by modifying hydraulic parameters and/or the type of flow mechanisms expected for different parts of a study area. One common supplementary analysis for process-based models is to apply pedotransfer functions to estimate soil hydraulic parameters from commonly measured soil properties such as texture and organic matter content. RF, however, circumvents the need to carry out this intermediate and supplementary step by not requiring prior assumptions on the system dynamics, thus not anchoring its estimates on soil hydraulic properties. For this study, another difference between the two models applied is the use of SSM values in developing the RF model. Although theoretically, RF could be carried out excluding SSM, the results from

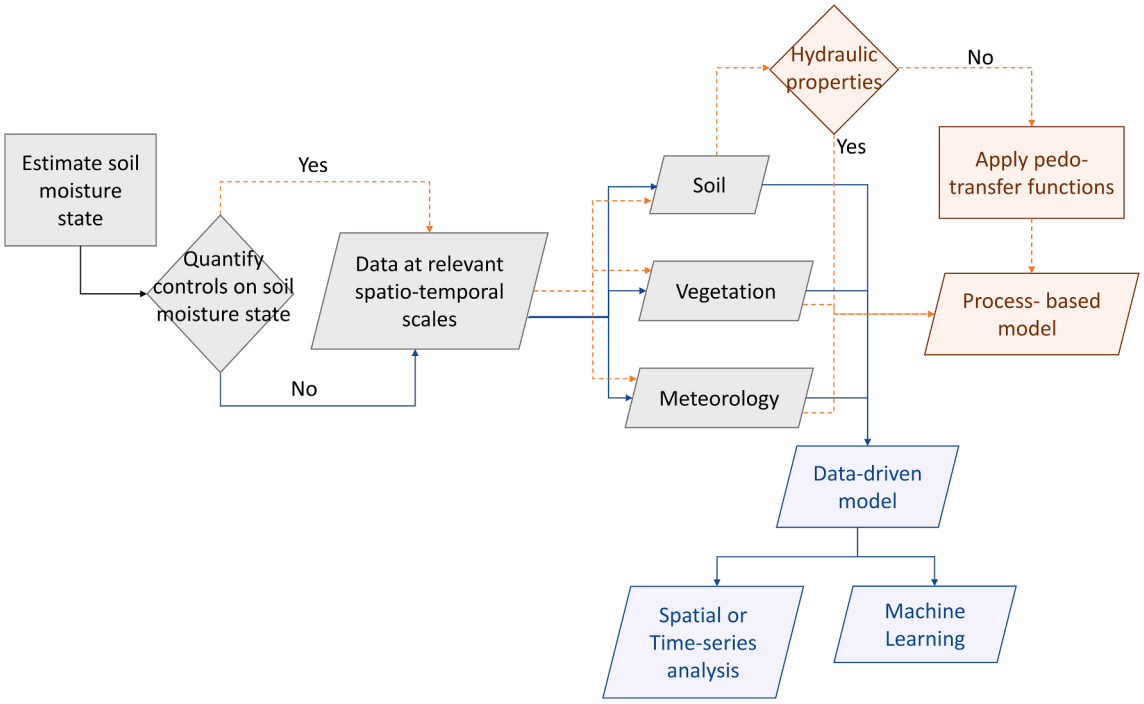

Fig. 13. A flowchart for model selection between data-driven or process based methods.

the variable importance list show that SSM is relevant for achieving good model performance. Satelite-derived SSM is a good alternative for the RF model in cases where in situ measurements are insufficient. Lastly, high RF accuracy for prediction of soil moisture values further opens opportunities for filling data gaps in highly non-linear time-series datasets.

The increasing amount of available soil moisture measurements globally could be a resource for expanding the application of data-driven methods in soil hydrology. Similar to what is carried out in this study, creation of a single model from numerous soil moisture networks could potentially allow for operational RZSM prediction or forecasting at different spatio-temporal scales. In situations where the primary goal is to determine the soil moisture state, RF is a good approach as it can be applied based on accessible surficial datasets. However, it is not capable of determining the dominant processes that control the soil moisture state, although a hint may be provided in the important variables list identified by the RF model. The impact of certain processes on soil moisture state may be better analyzed using a process-based model. The context in which each method may be the better option is summarized in Fig. 13.

4. Conclusions

In this study, we demonstrated the capabilities of a data-driven method using Random Forest for estimating root zone soil moisture with high accuracy, similar to process-based models. It may be advantageous to apply a Random forest framework for areas with limited information on soil hydraulic properties, and may circumvent the need to apply pedo-transfer functions. Increasing availability of soil moisture datasets,from in situ measurements worldwide and from satellites, provide opportunities in data-driven methods for large scale studies or operational (water) management. The results from the Random forest model does not explicitly elaborate on process controlling soil moisture state and may suffer from poor extrapolation results. It does, however, provide the important variables influencing the prediction accuracy which already hints at factors controlling soil moisture variability.

CRediT authorship contribution statement

Coleen Carranza: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing - original draft, Project administration, Visualization. **Corjan Nolet:** Methodology, Writing - review & editing, Visualization. **Michiel Pezij:** Methodology, Writing - review & editing. **Martine van der Ploeg:** Conceptualization, Supervision, Writing - review & editing, Project administration, Funding acquisition.

Declaration of Competing Interest

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

Acknowledgments

This study is part of the project entitled Operational Water Management using Sentinel-1 Satellites (OWAS1S) with project number 13871. The project is funded by Toegepaste and Technishe Wetenschappen (TTW) which is part of the Netherlands Organization for Scientific Research (NWO).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://doi.org/10.1016/j.jhydrol.2020.125840.

References

Ahmad, S., Kalra, A., Stephen, H., 2010. Estimating soil moisture using remote sensing data: A machine learning approach. Advances in Water Resources 33, 69–80.
Albergel, C., Rüdiger, C., Pellarin, T., Calvet, J.-C., Fritz, N., Froissard, F., et al., 2008. From near-surface to root-zone soil moisture using an exponential filter: an assessment of the method based on in-situ observations and model simulations.
Ali, M., Deo, R.C., Maraseni, T., Downs, N.J., 2019. Improving spi-derived drought forecasts incorporating synoptic-scale climate indices in multi-phase multivariate empirical mode decomposition model hybridized with simulated annealing and kernel ridge regression algorithms. Journal of hydrology 576, 164–184.

Araya, S.N., Ghezzehei, T.A., 2019. Using machine learning for prediction of saturated hydraulic conductivity and its sensitivity to soil structural perturbations. Water Resources Research 55, 5715–5737.
Benninga, H.-J.F., Carranza, C.D., Pezij, M., van Santen, P., van der Ploeg, M.J., Augustijn, D.C., et al., 2018. The raam regional soil moisture monitoring network in the netherlands. Earth System Science Data 10, 61.
Bolten, J.D., Crow, W.T., Zhan, X., Jackson, T.J., Reynolds, C.A., 2009. Evaluating the utility of remotely sensed soil moisture retrievals for operational agricultural drought monitoring. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 3, 57–66.
Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.
Coopersmith, E.J., Cosh, M.H., Bell, J.E., Boyles, R., 2016. Using machine learning to produce near surface soil moisture estimates from deeper in situ records at us climate reference network (uscrn) locations: Analysis and applications to amsr-e satellite validation. Advances in Water Resources 98, 122–131.
Cordova, J.R., Bras, R.L., 1981. Physically based probabilistic models of infiltration, soil moisture, and actual evapotranspiration. Water Resources Research 17, 93–106.
De Laat, P., 1980. Model for unsaturated flow above a shallow water-table, applied to a regional sub-surface flow problem (Pudoc, Centre for Agricultural Publishing and Documentation).
Dobriyal, P., Qureshi, A., Badola, R., Hussain, S.A., 2012. A review of the methods available for estimating soil moisture and its implications for water resource management. Journal of Hydrology 458, 110–117.
Dorigo, W., Van Oevelen, P., Wagner, W., Drusch, M., Mecklenburg, S., Robock, A., et al., 2011. A new international network for in situ soil moisture data. Eos, Transactions American Geophysical Union 92, 141–142.
Evensen, G., 2009. Data assimilation: the ensemble Kalman filter. (Springer Science & Business Media.
Feddes, R., Kabat, P., Van Bakel, P., Bronswijk, J., Halbertsma, J., 1988. Modelling soil water dynamics in the unsaturated zone-state of the art. Journal of Hydrology 100, 69–111.
Gao, X., Zhao, X., Brocca, L., Pan, D., Wu, P., 2019. Testing of observation operators designed to estimate profile soil moisture from surface measurements. Hydrological Processes 33, 575–584.
Guo, L., Lin, H., 2018. Addressing two bottlenecks to advance the understanding of preferential flow in soils. Advances in Agronomy (Elsevier) 147, 61–117.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.
Heathman, G.C., Starks, P.J., Ahuja, L.R., Jackson, T.J., 2003. Assimilation of surface soil moisture to estimate profile soil water content. Journal of Hydrology 279, 1–17.
Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6, e5518.
Houser, P.R., Shuttleworth, W.J., Famiglietti, J.S., Gupta, H.V., Syed, K.H., Goodrich, D. C., 1998. Integration of soil moisture remote sensing and hydrologic modeling using data assimilation. Water Resources Research 34, 3405–3420.
Houtekamer, P.L., Mitchell, H.L., 1998. Data assimilation using an ensemble kalman filter technique. Monthly Weather Review 126, 796–811.
Karandish, F., Šimnek, J., 2016. A comparison of numerical and machine-learning modeling of soil water content with limited input data. Journal of Hydrology 543, 892–909.
Kayastha, N., Solomatine, D., Lal Shrestha, D., 2014. Prediction of hydrological models' uncertainty by a committee of machine learning-models. In: 11th International Conference on Hydroinformatics.
Kornelsen, K.C., Coulibaly, P., 2014. Root-zone soil moisture estimation using data-driven methods. Water Resources Research 50, 2946–2962.
Koster, R.D., Suarez, M.J., 2001. Soil moisture memory in climate models. Journal of Hydrometeorology 2, 558–570.
Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A.K., Hochreiter, S., Nearing, G.S., 2019. Toward improved predictions in ungauged basins: Exploiting the power of machine learning. Water Resources Research 55, 11344–11354.
Kurc, S.A., Small, E.E., 2007. Soil moisture variations and ecosystem-scale fluxes of water and carbon in semiarid grassland and shrubland. Water Resources Research 43.
[Dataset] Lever, J., Krzywinski, M., Altman, N., 2016. Points of significance: model selection and overfitting.
Marquardt, D.W., 1963. An algorithm for least-squares estimation of nonlinear parameters. Journal of the Society for Industrial and Applied Mathematics 11, 431–441.
Matei, O., Rusu, T., Petrovan, A., Mihuţ, G., 2017. A data mining system for real time soil moisture prediction. Procedia Engineering 181, 837–844.
McColl, K.A., Alemohammad, S.H., Akbar, R., Konings, A.G., Yueh, S., Entekhabi, D., 2017. The global distribution and dynamics of surface soil moisture. Nature Geoscience 10, 100–104.
Meinshausen, N., 2006. Quantile regression forests. Journal of Machine Learning Research 7, 983–999.
Montavon, G., Samek, W., Müller, K.-R., 2018. Methods for interpreting and understanding deep neural networks. Digital Signal Processing 73, 1–15.
Orth, R., Seneviratne, S.I., 2012. Analysis of soil moisture memory from observations in Europe. Journal of Geophysical Research: Atmospheres 117.
Pezij, M., Augustijn, D.C., Hendriks, D.M., Weerts, A.H., Hummel, S., van der Velde, R., et al., 2019. State updating of root zone soil moisture estimates of an unsaturated zone metamodel for operational water resources management. Journal of Hydrology X 4, 100040.
Pezij, M., Augustijn, D.C., Hendriks, D.M., Hulscher, S.J., 2020. Applying transfer function-noise modelling to characterize soil moisture dynamics: A data-driven approach using remote sensing data. Environmental Modelling & Software 104756.
Porporato, A., Daly, E., Rodriguez-Iturbe, I., 2004. Soil water balance and ecosystem response to climate change. The American Naturalist 164, 625–632.
Prasad, R., Deo, R.C., Li, Y., Maraseni, T., 2018. Soil moisture forecasting by a hybrid machine learning technique: Elm integrated with ensemble empirical mode decomposition. Geoderma 330, 136–161.
Reddy, G.T., Reddy, M.P.K., Lakshmanna, K., Kaluri, R., Rajput, D.S., Srivastava, G., et al., 2020. Analysis of dimensionality reduction techniques on big data. IEEE Access 8, 54776–54788.
Reichle, R.H., 2008. Data assimilation methods in the earth sciences. Advances in Water Resources 31, 1411–1418.
Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., et al., 2019. Deep learning and process understanding for data-driven earth system science. Nature 566, 195–204.
Rigden, A.J., Mueller, N.D., Holbrook, N.M., Pillai, N., Huybers, P., 2020. Combined influence of soil moisture and atmospheric evaporative demand is important for accurately predicting us maize yields. Nature Food 1, 127–133. https://doi.org/10.1038/s43016-020-0028-7.
Ritter, A., Hupet, F., Muñoz-Carpena, R., Lambot, S., Vanclooster, M., 2003. Using inverse methods for estimating soil hydraulic properties from field data as an alternative to direct methods. Agricultural Water Management 59, 77–96.
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, 163–176.
Shiri, J., Keshavarzi, A., Kisi, O., Karimi, S., 2017. Using soil easily measured parameters for estimating soil water capacity: soft computing approaches. Computers and Electronics in Agriculture 141, 327–339.
Shiri, J., Karimi, B., Karimi, N., Kazemi, M.H., Karimi, S., 2020. Simulating wetting front dimensions of drip irrigation systems: Multi criteria assessment of soft computing models. Journal of Hydrology 124792.
Shrestha, D., Kayastha, N., Solomatine, D., 2009. A novel approach to parameter uncertainty analysis of hydrological models using neural networks. Hydrology and Earth System Sciences 13, 1235.
Simunek, J., Van Genuchten, M.T., Sejna, M., 2005. The HYDRUS-1D software package for simulating the one-dimensional movement of water, heat, and multiple solutes in variably-saturated media. Tech. rep.
Sluiter, R., 2012. Interpolation methods for the climate atlas. World Bank Policy Research Working Paper TR-335, Koninklijk Nederlands Meteorologisch Instituut (KNMI), De Bilt, the Netherlands.
Soenario, I., Plieger, M., Sluiter, R., 2010. Optimization of Rainfall Interpolation. Tech. rep., Koninklijk Nederlands Meteorologisch Instituut (KNMI)..
Spear, R.C., Cheng, Q., Wu, S.L., 2020. An example of augmenting regional sensitivity analysis using machine learning software, Water Resources Research 56, e2019WR026379. doi: 10.1029/2019WR026379.
Srivastava, P.K., Han, D., Ramirez, M.R., Islam, T., 2013. Machine learning techniques for downscaling smos satellite soil moisture using modis land surface temperature for hydrological application. Water Resources Management 27, 3127–3144.
Teweldebrhan, A.T., Schuler, T.V., Burkhart, J.F., Hjorth-Jensen, M., 2020. Coupled machine learning and the limits of acceptability approach applied in parameter identification for a distributed hydrological model. Hydrology and Earth System Sciences 24, 4641–4658. https://doi.org/10.5194/hess-24-4641-2020.
Tyralis, H., Papacharalampous, G., Langousis, A., 2019. A brief review of random forests for water scientists and practitioners and their recent history in water resources. Water 11, 910.
Ulaby, F.T., Dubois, P.C., Van Zyl, J., 1996. Radar mapping of surface soil moisture. Journal of Hydrology 184, 57–84.
van Genuchten, M.T., 1980. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils 1. Soil Science Society of America Journal 44, 892–898.
Van Looy, K., Bouma, J., Herbst, M., Koestel, J., Minasny, B., Mishra, U., et al., 2017. Pedotransfer functions in earth system science: challenges and perspectives. Reviews of Geophysics 55, 1199–1256.
Vereecken, H., Huisman, J., Bogena, H., Vanderborght, J., Vrugt, J., Hopmans, J., 2008. On the value of soil moisture measurements in vadose zone hydrology: A review, Water Resources Research 44.
Wagner, W., Lemoine, G., Rott, H., 1999. A method for estimating soil moisture from ers scatterometer and soil data. Remote Sensing of Environment 70, 191–207.
Wosten, J., de Vries, F., Hoogland, T., Massop, H., Veldhuizen, A., Vroon, H., et al., 2013. BOFEK2012, de nieuwe bodemfysische schematisatie van Nederland. Tech. rep., Alterra.
Wright, M.N., Ziegler, A., 2015. Ranger: A fast implementation of random forests for high dimensional data in c++ and r. arXiv preprint arXiv:1508.04409.
Zhang, D., Zhang, W., Huang, W., Hong, Z., Meng, L., 2017. Upscaling of surface soil moisture using a deep learning model with viirs rdr. ISPRS International Journal of Geo-Information 6, 130.
Zhuang, R., Zeng, Y., Manfreda, S., Su, Z., 2020. Quantifying long-term land surface and root zone soil moisture over tibetan plateau. Remote Sensing 12. https://doi.org/10.3390/rs12030509.