Evaluating the transferability of data-driven pedo-transfer functions for the wflow_sbm parameter KsatHorFrac in central and Western Europe

MSc Thesis

H.B. van der Gaast (Rik)

13-08-2024

Deltares UNIVERSITEIT TWENTE.

Colophon

Title:	Evaluating the transferability of data-driven pedo-transfer functions for the wflow_sbm parameter KsatHorFrac in central and Western Europe
Version:	Final
Date:	13-08-2024
Author:	H.B. van der Gaast (Rik)
Author e-mail:	h.b.vandergaast@student.utwente.nl
Student number:	s1980289
Programme:	Civil engineering and management
Profile:	Integrated water management
Supervised by:	Dr. ir. M.J. Booij (Martijn) (University of Twente) Dr. ir. A. Bomers (Anouk) (University of Twente) Prof. Dr. A. Weerts (Albrecht) (Deltares) Dr. R. Imhoff (Ruben) (Deltares)

Preface

The master thesis you are about to read is the final part which completes the master's program Civil Engineering and management at the University of Twente. This research, titled: 'Evaluating the transferability of data-driven pedo-transfer functions for the wflow_sbm parameter KsatHorFrac in central and Western Europe' was done in collaboration with Deltares.

During the master's track I discovered that I enjoy working with hydrological models. Hence, it did not take long to before this assignment caught my eyes. However, I did not have any experience with machine learning at all, but I was in for a challenge. I would like to thank my supervisors at Deltares, Albrecht Weerts and Ruben Imhoff for this opportunity in the first place. Secondly, I want to thank them for sharing their knowledge and ideas during our weekly update meetings. Besides the weekly meetings, you always had time to answer questions that popped up or have an additional meeting. Additionally, I would also like to thank Martijn Booij and Anouk Bomers, which were my supervisors from the University of Twente. I really appreciate your feedback throughout the whole process. I especially struggled during the proposal stage, but because of our meetings I think we created a solid ground to build this thesis on.

Finally, I want to thank my parents for always supporting me during the years I studied in Enschede. Emma, we worked on our theses simultaneously and I really think we pulled each other along in the process, bringing out the best in both of us. Finally, I would like to thank my friends in Enschede which made the past seven years a time to never forget.

I hope you will enjoy reading this thesis.

Rik van der Gaast

Enschede, August 2024

Summary

This thesis report provides insight into the transferability of data-driven Pedo-Transfer Functions (PTF) for the parameter KsatHorFrac in the distributed hydrological model wflow_sbm. KsatHorFrac is necessary to determine the horizontal saturated hydraulic conductivity by multiplying it with the vertical saturated hydraulic conductivity. PTFs are used to predict the value of a parameter based on widely available data, such as soil characteristics. Having a PTF for all parameters in a hydrological model is preferable, since good functioning PTFs bypass the need for calibrating each individual parameter. As of current, KsatHorFrac does not have a validated PTF in the wflow_sbm model. However, Ali et al. (2023) developed a PTF for KsatHorFrac based on seven soil characteristics in Great Brittain (GB) using the Machine Learning (ML) technique Random Forest (RF). In this study, the transferability of the RF PTF developed by Ali et al. (2023) (original PTF) is evaluated in subbasins in central and Western Europe. Additionally, the PTF is updated with topographic predictors (updated PTF) in order to evaluate how this impacts the KsatHorFrac prediction and resulting model performance on discharge simulations in wflow_sbm.

First, suitable validation subbasins were selected in the Seine, Loire, Rhone, Rhine, Po and upper section of the Danube basins. These subbasins were selected on the level of conformity of the SoilGrids variables with the training subbasins in GB: The level of conformity was translated into scoring categories 0 to 7, where 7 means all variables are within the range of the GB training subbasins and 0 indicates no variables are within the range of the GB training subbasins. At first, this selection consisted of 102 subbasins. However, after analysing the hydrographs of the discharge simulation, an additional 26 subbasins were discarded as the simulation results were unrealistic (mostly overestimating the measured discharge). Resulting in a selection of 76 subbasins to evaluate the transferability of both PTFs. Subsequently, an optimised KsatHorFrac value for each subbasin was calculated using a sensitivity analysis. This value and resulting model performance of wflow_sbm in terms of np KGE was defined as benchmark, to which the predicted KsatHorFrac and resulting model performance were compared. Using a Wilcoxon test it was found that the predicted KsatHorFrac in the scoring category 0-1 and 2 were significantly lower compared to the optimised value. Furthermore, the np KGE of all scoring categories was significantly lower compared to the optimised np KGE, indicating that the original PTF does not function well in areas which are different compared to the GB training subbasins.

To improve the original PTF, the model was trained on more characteristics than only the SoilGrids variables as used in Ali et al. (2023). Updating the PTF with the topographic variables elevation and slope resulted in overall higher predicted KsatHorFrac in the GB training subbasins. The KsatHorFrac prediction and resulting model performance in wflow_sbm in the validation subbasins for updated and original PTF were compared. It was found that the predicted KsatHorFrac was significantly higher for the scoring categories 0-1, 2 and 5 when the updated PTF was implemented. This resulted in a significant increase in wflow_sbm performance in the scoring categories 0-1. For the remaining categories, the difference was statistically insignificant.

The results indicate that the original PTF of Ali et al. (2023) predicts the KsatHorFrac accurately for subbasins with a score between 5-7, which have similar soil characteristics compared to the training subbasins in GB. This is not the case for subbasins with a low score (0-2), where the predicted KsatHorFrac was significantly lower compared to the optimised value. Updating the PTF with the topographic predictors: elevation and slope resulted in an overall improvement of KsatHorFrac prediction in subbasins with a low score 0-1, since the predicted KsatHorFrac was more similar compared to the optimised KsatHorFrac. However, for both PTFs, the np KGE resulting from the predicted KsatHorFrac was significantly lower compared to the optimised np KGE. This indicates that the predictive capability of a data-driven PTF for KsatHorFrac can still be improved.

1. Introduction

Hydrological modelling is used to simulate runoff in catchments, which is getting more and more relevant as climate change can intensify both droughts and extreme rainfall events (Singh, 2018). Extreme rainfall events can lead to an increased runoff, which can cause floodings. In order to mitigate the potential damage, accurate forecasts are necessary. A hydrological model simulates hydrological fluxes and states based on forcing data, such as rainfall and evapotranspiration, and catchment properties such as soil characteristics, topography and vegetation (Devia et al., 2015). However, before such a model can create reliable simulations for a specific study area, the model should be calibrated for that particular area. For fully distributed models, sensitive parameters for all the grid cells need to be calibrated. This results in a high-dimensional problem which makes calibrating distributed hydrological models not feasible. As a first step to simplify this calibration problem, Samaniego et al. (2010) proposed a multiscale parameter regionalization (MPR) technique. Here, relations between basin predictors and global parameters were found by calibrating predefined functions. Most often these functions were simple transfer functions, as they were hand-picked and only included a few predictors.

As a continuation on this calibration problem, pedo-transfer functions (PTF) were implemented into distributed models. A PTF is a function which predicts a soil (hydraulic) related parameter based on other soil related predictor variables which are most often widely available in terms of data (Abdelbaki, 2021). Transfer functions are created empirically within one particular area. Therefore, PTFs are more reliable for areas with similar variable values compared to the area where they were developed. Parametrisation of the soil using PTFs is also common practice in Land Surface Models (LSM) (Chou et al., 2022). LMS's are mainly focussed on vertical energy fluxes such as heat and water (Scanlon et al., 2018), whereas hydrological models also incorporate horizontal flows. However, both models are used to assess the availability of water resources (Schellekens et al., 2017). Besides the benefits of implementing PTFs, Blyth et al. (2021) states that the PTFs regarding surface-subsurface exchange are currently limited by the scares number of sets applicable for global modelling. Van Looy et al. (2017) concludes that methodological advances are necessary to allow for global modelling, and highlights the availability of high resolution soil data to improve the parameter prediction of PTFs.

1.1. Context

1.1.1. KsatHorFrac

Wflow_sbm is a fully distributed hydrological model which makes use of PTFs for parameter estimation. Because of the PTFs, wflow_sbm can be used for any catchment (Van Verseveld et al., 2024). One of the more sensitive parameters of the model is KsatHorFrac (Imhoff et al., 2020), which is the ratio between the vertical and horizontal saturated hydraulic conductivity (Eq.1). KsatHorFrac can be found in the soil module of wflow_sbm (elaborated in Section 2.1). The vertical flow of the water is amongst others dictated by the vertical saturated hydraulic conductivity (K_{SatVer}), for which the PTF of Brakensiek et al. (1984) is used. Besides flowing downward into the ground, the ground water also flows downstream in the saturated zone. This is dictated by the horizontal saturated hydraulic conductivity (K_{SatHor}). K_{SatHor} plays a role in the subsurface runoff though lateral flows. When KsatHorFrac is equal to one, the K_{SatHor} is equal to the K_{SatVer}. When the KsatHorFrac increases, K_{SatHor} becomes larger, resulting in a faster subsurface flow and thus a higher baseflow. As of currently, no PTF for KsatHorFrac is implemented in wflow_sbm. Hence, the parameter is set to an uncalibrated default value of 100. In order to make reliable simulations, this parameter has to be calibrated manually.

$$K_{satHor} = K_{satVer} * KsatHorFrac \quad \text{Eq. 1}$$

The effect that KsatHorFrac has on the discharge simulation in wflow_sbm has been investigated by multiple studies. Wannasin et al. (2021) manually calibrated KsatHorFrac for different subbasins in the upper region of the Greater Chao Phraya River (GCPR) basin in Thailand. In this study different KsatHorFrac values between 100 and 800 were considered, which were uniformly applied to the subbasins. The authors discussed that this small range of considered KsatHorFrac values affected the uncertainty of the discharge series. However, the uncertainty can clearly be seen in the troughs of the hydrograph in Figure 1.

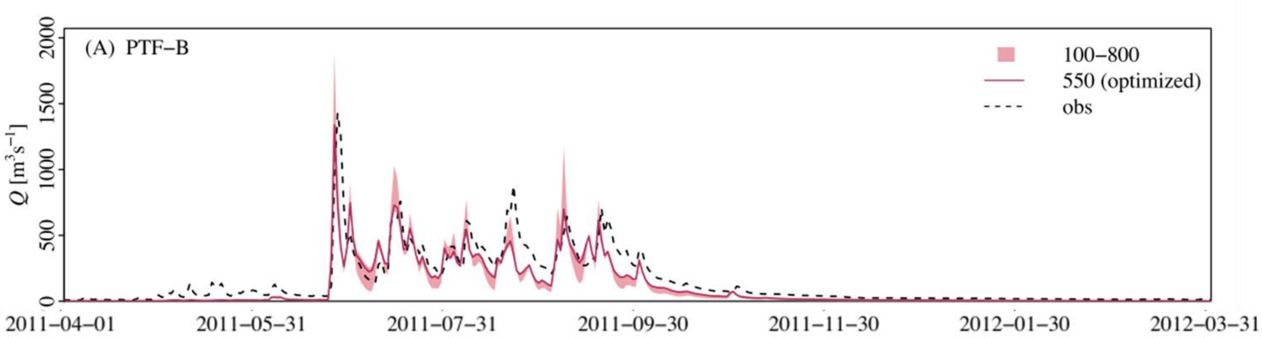


Figure 1: Hydrograph of the Nan_natural subbasin in 2011 (section of calibration period). Observed discharge and optimised discharge are indicated by a dotted and solid line respectively. The uncertainty band caused by the range of KsatHorFrac values (100-800) is *shown light coloured (Wannasin et al., 2021)*

Van Verseveld et al. (2024) illustrated the effect of different values for KsatHorFrac (1, 20, 50 and 100) on the simulated discharge in the Whanganui River basin (New Zealand). Figure 2 shows the simulated discharge which resulted from the four KsatHorFrac values. The figure shows that a KsatHorFrac of 1 has relatively low baseflow and high peaks. On the opposite, a KsatHorFrac of 100 has higher baseflow and lower peaks. The authors expect that changing KsatHorFrac causes a shift in the contribution of overland flow and lateral subsurface flow to simulated discharge.

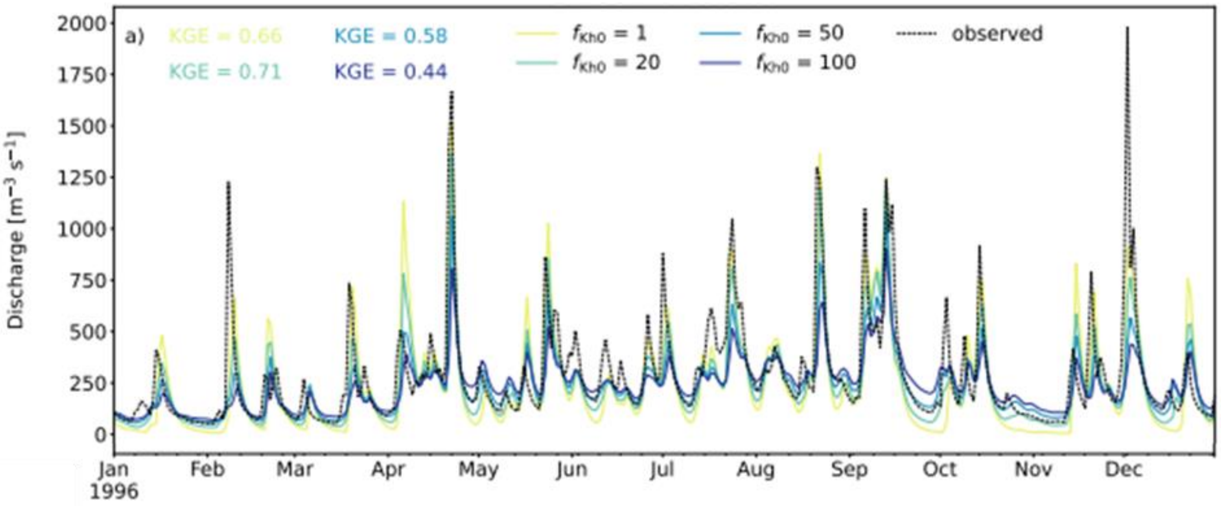


Figure 2: Hydrographs of the Whanganui River basin in 1996. Observed and simulated discharge are indicated by black and coloured lines respectively (Van Verseveld et al., 2024)

Both studies discuss that the model performance of wflow_sbm can be improved by implementing a PTF for the sensitive KsatHorFrac parameter. To add to this, Wannasin et al. (2021) suggest to calibrate the KsatHorFrac based on soil types instead of a uniform value across the basin.

1.2. State of the art

Ali et al. (2023) investigated the development of a PTF for KsatHorFrac, using Machine Learning (ML). The authors applied two different ML algorithms: Random Forest (RF) (explained in Section 2.2) and

Boosted Regression Trees (BRT). The algorithms were trained and tested in 459 subbasins in Great Brittain (GB). From the total number of subbasins, 344 subbasins were used to train the algorithms using seven depth and subbasins averaged SoilGrids variables (elaborated in Section 2.3) and a subbasin averaged optimised KsatHorFrac value as target variable. The considered range during the KsatHorFrac optimisation was between 1 and 10,000 (Weerts, 2024). This gave satisfactory results since no significant difference between the optimized KsatHorFrac value and the KsatHorFrac predicted by the algorithms could be observed in the remaining 115 test subbasins. Subsequently, the two ML PTFs were validated in the Loire basin (France). On average, the wflow_sbm model performed better when using the KsatHorFrac value predicted by the ML PTFs when compared to the default KsatHorFrac of 100. However, no significant difference could be observed between the predicted and default parameter values. Furthermore, it was found that the performance was especially low in the South-East of the basin. The authors conclude that the study shows the potential of ML to create a PTF which relates soil characteristics to soil hydraulic properties, such as KsatHorFrac.

1.3. Knowledge gap

As KsatHorFrac it is the ratio between the vertical and lateral saturated hydraulic conductivity it is difficult to create a dataset of measured values. This is because these are difficult to measure and thus creating an extensive dataset of only one of the two components is costly and time-consuming (Abdelbaki, 2021). This highlights the need for a PTF for KsatHorFrac which formed the base of the research of Ali et al. (2023). Even though a PTF was created and applied in other subbasin as where the PTF was developed for, large scale validation has not been done. Hence, a conclusion about the transferability of the PTF cannot be made.

Furthermore, it was chosen to train the ML algorithm with a limited number of predictors which are all related to the soil. Topographic predictors were not included in the research even though these could prove to be relevant according to literature (S. Gupta et al., 2021). Since research into this area has not been conducted before, it is unclear whether including topographic predictors in the development of a PTF for KsatHorFrac will improve the parameter prediction and the resulting model performance of wflow_sbm.

1.4. Research aim and research questions

1.4.1. Research aim

The aim of this research consists of two parts: First, the research aims to get more insight into the transferability of the Random Forest (RF) KsatHorFrac PTF of Ali et al. (2023) (original PTF). It was chosen to focus on this algorithm since it outperformed the BRT algorithm in both parameter estimation and resulting wflow_sbm performance. Subsequently, the aim is to evaluate how parameter estimation and resulting model performance of wflow_sbm changes in the validation subbasins when the RF PTF of Ali et al. (2023) is updated by including topographical predictors in the development phase of the PTF (updated PTF).

1.4.2. Research questions

In order to achieve the research aims, the following questions need to be answered.

The first question aims at identifying subbasins for which both the original and updated PTFs can be validated. These catchments should be outside of GB, as both PTFs will be trained and tested there.

1. *Which catchments should be considered for validating the original and updated PTF?*

The second question focusses on the validation of the original PTF in subbasins identified in research question 1. The results of this step provide insight in the transferability of the PTF as this proves to be problematic for empirically developed PTFs (Zuo & He, 2021).

2. *How does the KsatHorFrac prediction of the original PTF and resulting model performance of wflow_sbm change in comparison to the optimised scenario, when the original PTF is implemented for subbasins for which it was not developed?*

In the third question, the training set will be updated with topographic predictors. In order to train and test the updated PTF, optimised KsatHorFrac values from the Ali et al. (2023) are used. These can be used since the same catchments in GB are considered and the optimised values are calibrated.

3. *How does the KsatHorFrac prediction of the PTFs change in the GB training subbasins when topographic predictors are implemented in the RF training phase?*

The final question will focus on the validation of the updated PTF. It will be investigated whether the prediction of the updated PTF and resulting model performance of wflow_sbm will improve with respect to the original PTF. This is done by running wflow_sbm with the updated PTF in the subbasins which were identified in question one.

4. *How does the KsatHorFrac prediction of the updated PTF and resulting performance of wflow_sbm change, with respect to the original PTF, when the updated PTF is implemented for subbasins for which it was not developed?*

1.5. Report outline

The outline of the remainder of this report are as follows: First Chapter 2 describes the models which are considered, together with the necessary data. The methods which were used to answer the research questions can be found in Chapter 3. Chapter 4 discusses the results of the study in the same format as Chapter 3. Finally, the limitations and interpretations of the results and methods and an overall conclusion and recommendations of the study can be found in chapters 5 and 6 respectively.

2. Models and data

This chapter describes the models and materials which are used in the study. First wflow is described, which consists of the hydrological model wflow_sbm and the model builder HydroMT. Subsequently, the Random Forest algorithm is described together with the research method which was used by Ali et al. (2023) to develop the original PTF. Finally, the relevant data are summarised.

2.1. Wflow

2.1.1. Wflow_sbm

Wflow_sbm is a distributed hydrological model, which means that the model is divided into different grid cells. Each grid cell has its own set of parameters, which creates a high resolution model. Additional advantages are that distributed models can use the available data to full effect. Not only can forcing data be applied to the gridded model, data captured by satellites can function as input as well. As a result, physical relations between land cover and soil (hydraulic) properties can more easily be made. Instead of calibrating all sensitive parameters for each grid cell individually, PTFs are used for sensitive to estimate the optimal parameter value based on environmental data (soil, vegetation, etc.). Because of the PTFs, wflow_sbm can be used for any catchment (Van Verseveld et al., 2024). The vertical part of soil module of Wflow_sbm is mostly based on the topog_sbm model (Vertessy & Elsenbeer, 1999). This includes gravity based infiltration as well as capillary rise. Lateral flows such as channel, overland and subsurface flow is based on the following models: TOPKAPI (Benning et al., 1995), G2G (Bell et al., 2007), 1K-DHM (Tanaka & Tachikawa, 2015) and again, Topog_SBM (Vertessy & Elsenbeer, 1999). An overview of the processes of the wflow_sbm model can be seen in Figure 3. KsatHorFrac can be found in the saturated store of the soil module.

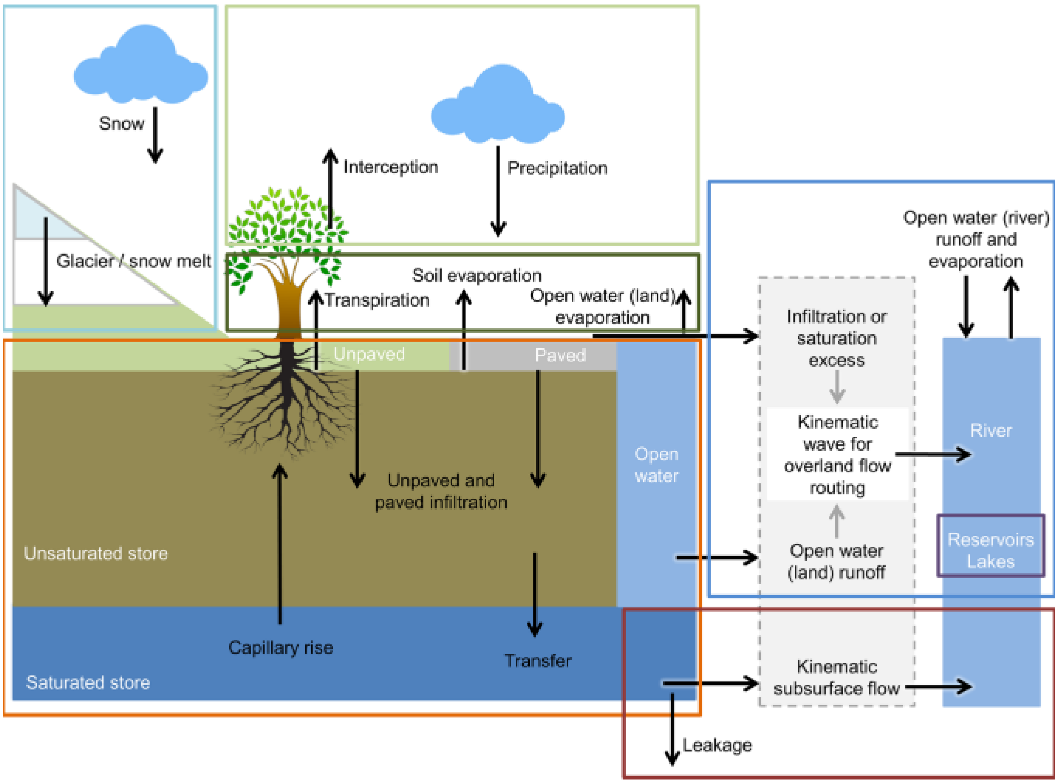


Figure 3: Overview of processes and fluxes in the wflow_sbm model. Routines: snow and glaciers (blue), interception (green), soil module (orange), lateral subsurface flow (brown), surface routing (dark blue) (Van Verseveld et al., 2024)

2.1.2. Subsurface processes

Before addressing the location of KsatHorFrac within wflow_sbm, the key elements of the soil module are described. Before water can flow to the surface water via subsurface flow, water needs to enter the saturated store. Prior to the saturated store, the water travels from the surface to the unsaturated zone through infiltration. A feature of wflow_sbm is that the unsaturated zone can be divided into multiple layers. This becomes relevant for the transfer of water from the unsaturated zone to the saturated store. For an unsaturated store layer which is divided into multiple layers, the transfer from an unsaturated store to the next is:

$$Q^{t}_{transfer,pot,n} = K_{satver,n}\left(\frac{\theta - \theta_r}{\theta_s - \theta_r}\right)^{c_n} \qquad \text{Eq. 2}$$

$$c_n = \frac{2 + 3\lambda_n}{\lambda_n} \qquad \text{Eq. 3}$$

Where:

- $Q^{t}_{transfer,pot,n}$ = Potential transfer of water from an unsaturated layer n [mm t^{-1}]
- K_{satver} = Vertical saturated hydraulic conductivity at layer n [mm t^{-1}]
- θ_s = Saturated soil water content [mm mm^{-1}]
- θ_r = Residual soil water content [mm mm^{-1}]
- c_n = Brooks-Corey power coefficient at layer n [-]
- λ_n = Pore size distribution index at layer n [-]

Wflow_sbm considers a K_{satver} which is variable with depth. The relation between these is negative which means that Ksat reduces for an increasing depth (Eq. 4). Which is in accordance with multiple studies (Ameli et al., 2016; Águila et al., 2023) . As a result of the variability along the vertical axis, the term vertical saturated hydraulic conductivity is applied (K_{satver}).

$$K_{satver,n} = f_{Kv,n}K_0 e^{-f_{Kv}z_{bottom,n}} \qquad \text{Eq. 4}$$

Where:

- K_0 = Vertical saturated conductivity at surface [mm t^{-1}]
- $f_{Kv,n}$ = Optional multiplication factor (default = 1.0) [-]
- f_{Kv} = Scaling parameter [mm^{-1}]
- $z_{bottom,n}$ = Soil depth at bottom of layer n [mm]

Now it is known how the vertical inflow into the saturated store functions. However, the saturated store also interacts with neighbouring cells in the horizontal plane. Hence, lateral water transfers are also necessary. The route of the lateral subsurface flow is modelled using the kinematic wave approach. The water in the saturated store flows towards lower laying grids (downhill). Here, the subsurface flow is determined by Eq. 5. An important parameter is K_{sathor}(Eq. 6), which relates to the K_{satver}.

$$Q_{subsurface} = \frac{K_{h0}c_{landslope}}{f_{ssf,Kv}} * \left(e^{-f_{ssf,Kv}z_{ssf,watertable}} - e^{-f_{ssf,Kv}z_{ssf,soil}}\right)w \qquad \text{Eq. 5}$$

$$K_{h0} = 0.001K_{v0}f_{Kh0}\frac{\Delta t_b}{\Delta t} \qquad \text{Eq. 6}$$

Where:

- $Q_{subsurface}$ = subsurface flow [m^3 d^{-1}]
- K_{h0} = Horizontal saturated hydraulic conductivity at surface [m d^{-1}]

- $c_{landslope}$ = Land slope [m m^{-1}]
- K_{v0} = Vertical saturated hydraulic conductivity at surface [mm t^{-1}]
- f_{Kh0} = KsatHorFrac [-]
- $z_{ssf,watertable}$ = Depth of water table after unit conversion [m]
- $z_{ssf,soil}$ = Depth of soil after unit conversion [m]
- Δt = Time step [s]
- Δt_b = Base time step of wflow_sbm [s]

2.1.3. Hydromt

In this research, all validation subbasins were simulated separately. Hence, all subbasins had an independent wflow_sbm model. These models were build using the HydroMT open-source Python package (Eilander et al., 2023). HydroMT is used to automate the process of model building and making it possible to reproduce results. HydroMT can be divided into three sections: input data, methods and workflows, and models. An overview of theses sections can be seen in Figure 4. HydroMT supports four different kinds of data: RasterDataset such as forcing data, DataFrame such as conversions between parameters, GeoDataFrame such as river location and GeoDataset such as observation gauges. The data which was used in this research is described in Section 2.3. All calculations and model building is done in the blue area in the figure below. Here the data are converted into model layers and parameter maps. In addition to generating the hydrological model, HydroMT creates the model components which can be visualised in Qgis.

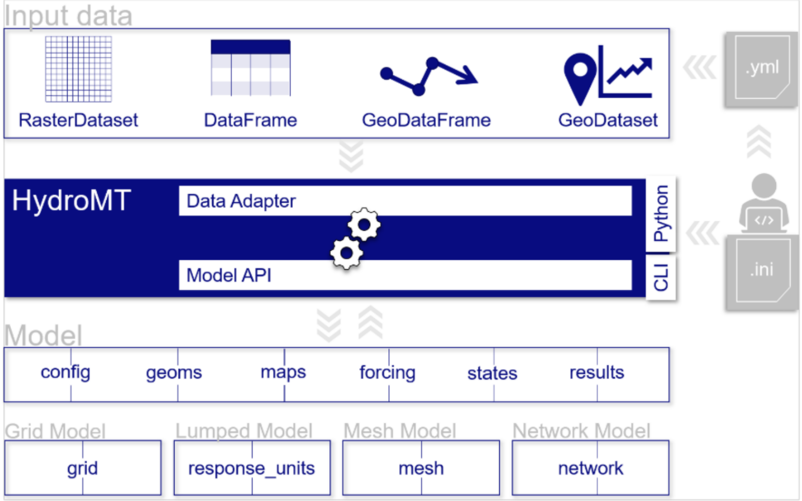


Figure 4: Schematic overview of HydroMT (Eilander et al., 2023)

2.2. Random Forest

2.2.1. Algorithm

Random Forest (RF) is a Machine Learning (ML) method which can be used for classification and regression (Breiman, 2001). As the name suggests, the RF is built up from a large number of individual decision trees (representation in Figure 5). Each decision tree is created using a bootstrapped sample from the training data. At each split, the decision is made by a random variable from the bootstrapped sample. This step needs to be done multiple times in order to create a variety of individual decision trees and thus a robust RF. Not all entries in the bootstrapped dataset are used to create the decision trees, these samples are called ‘out of bag’

The end result is determined by taking the average of all decision trees. Because of this, the method is robust against outliers in the data and reduces the chance of overfitting. However, a drawback of the

RF algorithm is the poor extrapolation capability (Hateffard et al., 2024). The performance of the algorithm is evaluated using the R^2 (Eq. 7) and Root mean Squared Log-transformed Error (Eq. 8). It was chosen to use the log-transformed error because of the large range of possible KsatHorFrac values (1-10,000). Multiple RFs, with a different number of trees, or different variables at the splits can be made. Subsequently, each tree is fed with the 'out of bag' samples. The algorithm with the best performance indicators is chosen to be the optimal Random Forest.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left[\log_{10}\left(\widehat{f_{Kh0}}\right)_i - \log_{10}(f_{Kh0})_i\right]^2}{\sum_{i=1}^{N}\left[\log_{10}\left(\overline{f_{Kh0}}\right)_i - \log_{10}(f_{Kh0})_i\right]^2} \qquad \text{Eq. 7}$$

$$RSMLE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\log_{10}\left(\widehat{f_{Kh0}}\right)_i - \log_{10}(f_{Kh0})_i\right]^2} \qquad \text{Eq. 8}$$

Where:

- R^2 = r-squared
- $RSMLE$ = Root Mean Squared Log-transformed Error
- N = Number of observations
- $\widehat{f_{Kh0}}$ = Predicted KsatHorFrac value from ML algorithms
- $\overline{f_{Kh0}}$ = Mean of optimised KsatHorFrac values
- f_{Kh0}= Optimised KsatHorFrac value

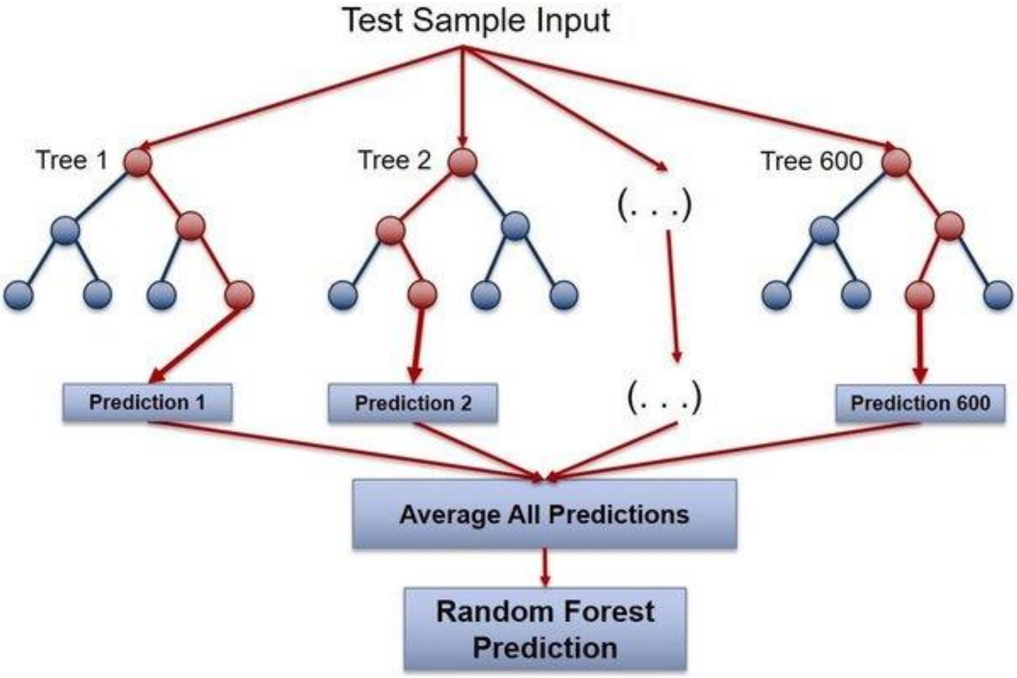


Figure 5: Random Forest model (Blakely et al., 2018)

2.2.2. Original PTF

Ali et al. (2023) used the research method in Figure 6 to develop the PTF for KsatHorFrac using a RF and BRT algorithm. However, in this research only the RF algorithm is considered. The algorithm has two main inputs: The depth and subbasin averaged SoilGrids variables (Table 2) for the GB subbasins and optimised KsatHorFrac values for the corresponding subbasins. The optimisation of the KsatHorFrac will be elaborated upon in Section 3.2. After randomly splitting the subbasins into a training (75%) and test (25%) set, the SoilGrids variables are standardised. This is because the large difference in magnitude between the variables. Multiple RFs are created using different number of trees and selection of variables at decision splits.

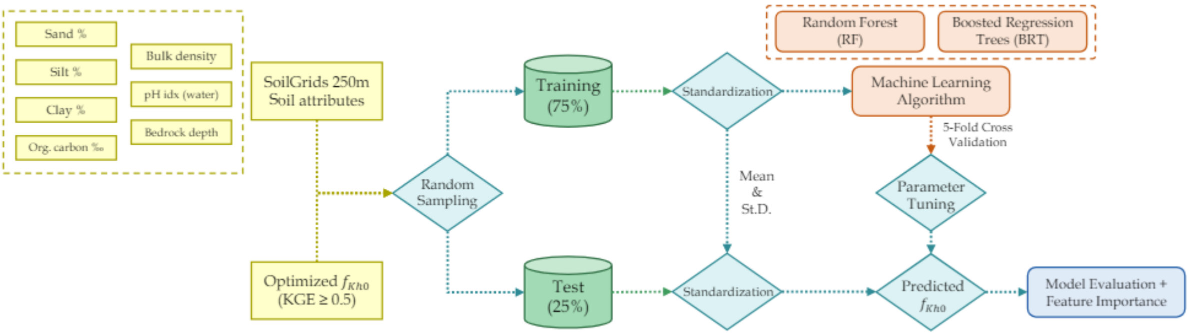


Figure 6: Flowchart of the method to develop a PTF for KsatHorFrac for a Random Forest and Boosted Regression (Ali et al., 2023)

The RF with the best performance can be seen in blue in Figure 7. This is the PTF which will be validated in this research and will be referred to as 'original PTF'. Section 3.2 elaborates on how to implement the PTF to the validation subbasins.

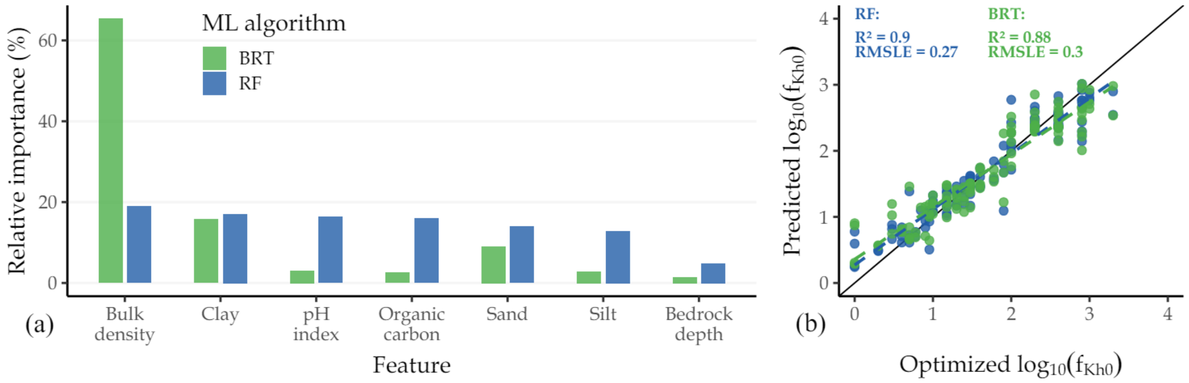


Figure 7: Results of the BRT and RF algorithms: (a) relative importance of the predictors, (b) scatterplot of the predicted and optimised KsatHorFrac values and corresponding performance indices (Ali et al., 2023)

2.3. Data

An overview of the data which is used in the wflow_sbm models can be seen in Table 1. The discharge time series of the Global Runoff Data Centre (GRDC) are used. Amongst this data is data of the European Water Archive (EWA), which contains long-term daily flow data (European Environment Agency, 2011). The forcing data which was used as input for the simulations is the E-OBS dataset version 25.0.e. The data consist of measured precipitation data which is retrieved from the weather stations of European National Meteorological and Hydrological Services (NMHSs) as well as other data institutions. For the temperature and potential evapotranspiration (PET), ERA5 forcing was used. For PET, the De Bruin method was used (De Bruin et al., 2016). The topography of the MERIT hydro database consists of: flow direction, elevation, slope, stream order, uparea and river width. The SoilGrids parameters (Table 2) are used in wflow_sbm as well as the training of the RF Algorithm which was used to develop the original PTF for KsatHorFrac.

Table 1: Overview of the datasets which were considered for wflow_sbm and the Random Forest algorithm

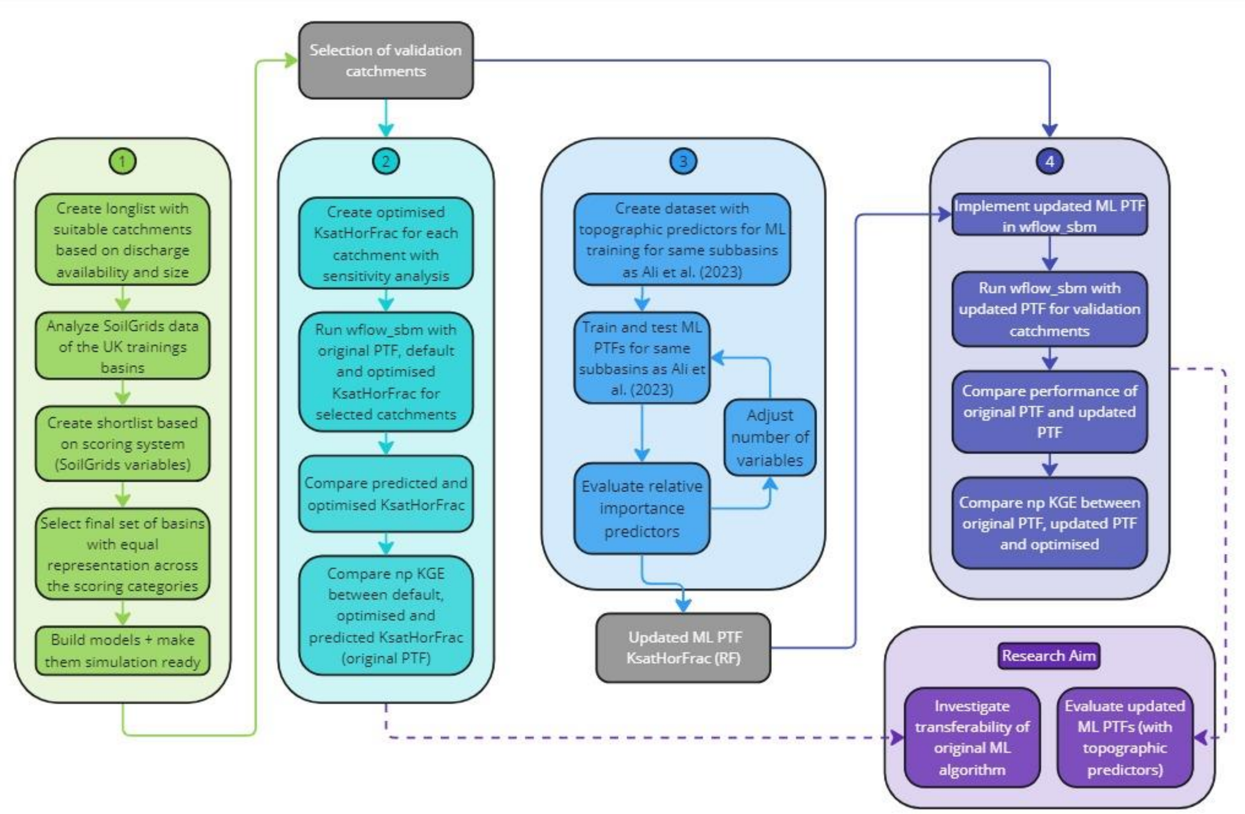
Variable	Dataset	Resolution	Period	Source
Dynamic data				
Discharge	GRDC, EWA	Daily	1958 to 2017	(GRDC, 2024)
Precipitation	E-OBS v25.0e	Daily	01-01-1950 to 31-12-2021	(Cornes, Schrier, Besselaar, & Jones, 2018)
Temperature	ERA5	Hourly	1950 to present	(Hersbach, et al., 2020)
Potential evapotranspiration	ERA5	Hourly	1950 to present	(Hersbach, et al., 2020)
Static data				
Topography	MERIT Hydro	3-arc sec (90 m at equator)	-	(Yamazaki, et al., 2019)
Soil maps	*SoilGrids v1.0 (* Table 2)	250 m	-	(Hengl, et al., 2017)
Land use	Vito	100 m	-	(Buchhorn, et al., 2020)

Table 2: SoilGrids v1.0 parameters (Hengl et al., 2017) which are predictors for the original PTF (Ali et al., 2023)

Variables	Abbreviation	Unit
Soil organic carbon	ORCDRC	g kg^{-1}
pH index (H_2O solution)	PHIHOX	10^{-1}
Sand content	SNDPPT	kg kg^{-1}
Silt content	SLTPPT	kg kg^{-1}
Clay content	CLYPPT	kg kg^{-1}
Bulk density	BLDFIE	kg m^{-3}
Depth to bedrock	BDTICM	cm

3. Research methods

This chapter describes the steps which were taken to answer the research questions. A flowchart of the methods is shown in Figure 8. Here, the different colours represent the research question which corresponds to the number at the top.



Selection of validation catchments
1
Create longlist with suitable catchments based on discharge availability and size
Analyze SoilGrids data of the UK trainings basins
Create shortlist based on scoring system (SoilGrids variables)
Select final set of basins with equal representation across the scoring categories
Build models + make them simulation ready
2
Create optimised KsatHorFrac for each catchment with sensitivity analysis
Run wflow_sbm with original PTF, default and optimised KsatHorFrac for selected catchments
Compare predicted and optimised KsatHorFrac
Compare np KGE between default, optimised and predicted KsatHorFrac (original PTF)
3
Create dataset with topographic predictors for ML training for same subbasins as Ali et al. (2023)
Train and test ML PTFs for same subbasins as Ali et al. (2023)
Adjust number of variables
Evaluate relative importance predictors
Updated ML PTF KsatHorFrac (RF)
4
Implement updated ML PTF in wflow_sbm
Run wflow_sbm with updated PTF for validation catchments
Compare performance of original PTF and updated PTF
Compare np KGE between original PTF, updated PTF and optimised
Research Aim
Investigate transferability of original ML algorithm
Evaluate updated ML PTFs (with topographic predictors)

Figure 8: Flowchart of research methods. The numbers 1-4 indicate the research questions. Purple arrows indicating that RQ 1 and 2 correspond to the first part of the research aim, while RQ 3 and 4 correspond to the second part of the research.

3.1. Subbasin selection

The first part of the research aims at identifying subbasins which are used for the validation of both the original PTF and the updated PTF. First a longlist will be created with the help of a set of criteria. Subsequently, this longlist will be reduced to a shortlist by implementing a scoring system.

3.1.1. Longlist

The focus area of the study is Central and Western Europe. This is because the longlist is partly based on the already available wflow_sbm models of the Rhine, Po and Loire. In order to make the study area continuous it was chosen to add the Seine and Rhone basins. Finally, the upper section of the Danube was added since an additional discharge dataset was available. However, within this study area a large number of subbasin can be found. In order to reduce this number to a longlist the following set of criteria was used:

- Potential subbasins should be gauged and historical daily discharge data should be available. Since other studies which use wflow_sbm have a calibration period between 2-4 years (Imhoff et al., 2020;Wannasin et al., 2021) it was decided that a minimum of 4 years should be available.
- Secondly, the subbasins should be headwaters.

The first criteria is the most important. Without historical discharge data no evaluation can be done. As the quality of the GRDC data is controlled, and thus reliable (GRDC, 2024) and wflow_sbm model can easily import GRDC data such as gauge locations using HydroMT, it was chosen to use the GRDC discharge data. Regarding the second criterium: The specific focus on headwaters is important as optimising the KsatHorFrac values is faster when the basins do not have an inflow from adjacent basins. Additionally, it is less likely that human interventions such as dams or reservoirs are present in the headwaters. The subbasin size is also important as it is not only related to computational time, but also to the homogeneity of soil. Smaller subbasins result in more homogeneous SoilGrids variables and resulting model parameter values when averaged. This will be elaborated upon in Section 3.1.2. Both arguments reduce the uncertainty of the discharge measurements. The selection of the headwaters within these basins was done manually in the Qgis (desktop 3.34.0) software. During this procedure the subbasins without inflow from adjacent subbasins were selected. This resulted in a longlist of 967 subbasins (Figure 9). As can be seen from this figure, the distribution of headwaters is not equal across the basins. In the Po basin, few headwaters can be found. This is caused by a limited number of subbasins within the available Po model and GRDC discharge gauges.

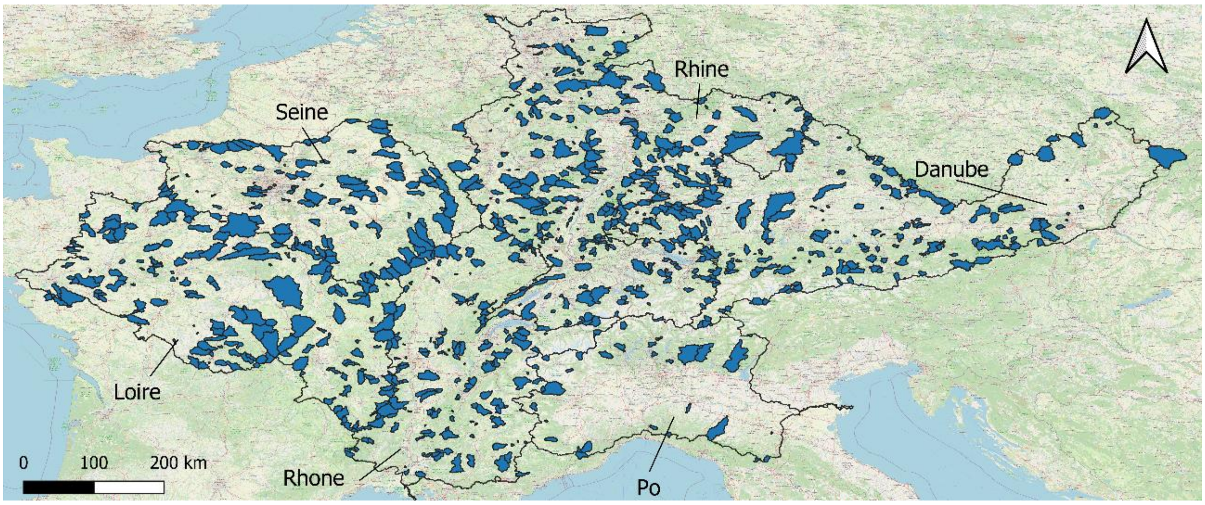


Figure 9: Subbasins from the longlist. Covering: Loire, Seine, Rhone, Rhine, Po and upper Danube basin

3.1.2. Shortlist

After the first selection round, the goal is to reduce the longlist to a shortlist since not all subbasins are suitable for the research. In order to evaluate the transferability of the original and updated PTF, an equal distribution of subbasins should be similar and different, in terms of soil characteristics to the GB training subbasins. The soil characteristics are represented by the predictors of the original PTF (SoilGrids v1.0). Subbasins with both similar and different soil characteristics are needed in order to test the following hypothesis: *"Subbasins with different soil characteristics compared to the GB training subbasins will have a lower model performance compared to subbasins with similar soil characteristics compared to the GB training subbasins when using the original PTF".*

Reducing the longlist to a shortlist requires multiple actions. Firstly, as the seven SoilGrids variables consist of seven depth layers, data processing is necessary in order to compare the potential subbasins with the training subbasins in GB. First, the parameters are averaged in depth according to the trapezoidal rule (Eq. 9) as stated in Hengl et al. (2017). Subsequently, the parameter values should be averaged for each individual subbasin. Which results in a depth and basin-average value for the seven SoilGrids variables for every subbasin in Figure 9.

$$\frac{1}{b-a}\int_a^b f(x)dx \approx \frac{1}{b-a}\frac{1}{2}\sum_{k=1}^{N-1}(x_{k+1}-x_k)(f(x_{k+1})-f(x_k)) \qquad \text{Eq. 9}$$

Where:

- a = Depth of top of soil layer [cm]
- b = Depth of bottom of soil layer [cm]
- N = Number of depths [-]
- x_k = k-th depth [cm]
- $f(x_k)$ = Value of SoilGrids variable at depth x_k [Table 2]

In order to determine whether a subbasin is similar or different compared to the SoilGrids variables of the trainings basins, a scoring system is implemented. The depth and subbasin averaged SoilGrids variable values of the subbasins in the longlist are compared to the boundaries (explained in following paragraph) of each depth and subbasin averaged SoilGrids variables of the GB trainings subbasins. If the averaged variable value of a validation subbasin is between the boundaries, one point is awarded. Hence, each basin from the longlist receives a score between 0 and 7. Where 0 means that all soil variable values are outside of the range of the GB training subbasins, and 7 means all parameter values are within that range. However, deciding which basins will form the shortlist depends on the final result of the scores: A large contrast between subbasins is preferred to test the hypothesis. Hence, if a sufficient number of subbasins remains in the ultimate scoring categories (0 or 7), only the subbasins within the ultimate scoring categories are selected. If the number of subbasins in these ultimate scores is not found to be sufficient, scoring categories are grouped together (e.g. 0 and 1, 6 and 7).

A large variety of subbasins is used when training the original PTF. Because of this outliers can be included in the training data. However, these outliers will probably not affect the output of the algorithm since RF is robust against outliers in the training data (Breiman, 2001). Hence, the outliers should not be considered when applying the scoring system. Therefore it was chosen to look at different ranges of the SoilGrids variables from the GB training subbasins. Besides the complete dataset per parameter, the following subsets are considered: data between the whiskers of a boxplot and the 95% and 90% confidence intervals. Here the whiskers of the boxplot are determined by looking at 1.5 times the inter quartile range (Q3 +1.5*(Q3-Q1) and Q1 -1.5*(Q3-Q1).

3.1.3. Final selection

Instead of randomly selecting subbasins, it was chosen to base the decision on subbasins size. To ensure that the sizes of the basins among scoring category are not significantly different, the Kruskal-Wallis-test with a significance level of 0.02 is used (Kruskal & Wallis, 1952). The Kruskal-Wallis-test is used to test a hypothesis for more than 2 samples. Additionally, as this test is non-parametric, the size of the basins does not have to be normally distributed. The Null hypothesis states that the size of the basins from each sample come from the same population. Whereas the alternative hypothesis states that at least one sample comes from a different population. The final selection of subbasins does not only depend on the result of research question 1. Subbasins can be discarded if problems are encountered during the model building in HydroMT. The subbasins were selected from basin size models. The river network in these models is based on a river geometry dataset (Lin et al., 2020), which only covers rivers wider than 30 meters. Since several headwaters do not have rivers which are that wide, they are not covered by the dataset. because of this, the rivers are build using another method which estimates the geometry based on a power law and the discharge . In order to prevent inconsistent model results, all subbasins with an area which deviates from the upstream area according to the GRDC gauges with more than 15% (used in State of Global Water Resources report 2022 (WMO, 2023)) are discarded from the final selection. Additionally, if a problem is encountered during the simulations there is a possibility to discard a subbasin from the validation set. Even though subbasins can be discarded, it should be ensured that every scoring category has a similar number of subbasins.

3.2. Transferability of the original PTF

In order to successfully evaluate the transferability of the original PTF, both the predicted KsatHorFrac and resulting performance of Wflow_sbm are compared to the default and a benchmark. However, as stated in the introduction, there is no extensive dataset of KsatHorFrac available. An alternative method to create a benchmark value is to create an optimised KsatHorFrac value for each subbasin. During the sensitivity analysis only KsatHorFrac is varied. This can be done, since the other sensitive wflow_sbm parameters (Appendix A.) are predicted using validated PTFs.

In the sensitivity analysis, KsatHorFrac will be varied between 1 to 10,000 in logarithmic intervals (Weerts, 2024), which is uniformly applied in each subbasin. The simulated discharge will be compared to the observed discharge and evaluated using the non-parametric Kling Gupta Efficiency ($npKGE$) as shown in Eq. 10. This adjustment to the conventional KGE is sensitive to the baseflow of the hydrographs, instead of peak flows (Pool et al., 2018). As KsatHorFrac dictates the subsurface flow, it is logical to evaluate the baseflow instead of looking at the complete hydrograph. $npKGE$ defined as follows:

$$npKGE = 1 - \sqrt{(\beta - 1)^2 + (\alpha_{NP} - 1)^2 + (r_s - 1)^2}, \qquad \text{Eq. 10}$$

$$\beta = \frac{\mu_{sim}}{\mu_{obs}}, \qquad \text{Eq. 11}$$

$$\alpha_{NP} = 1 - \frac{1}{2}\sum_{k=1}^{n}\left|\frac{Q_{sim}(I(k))}{n\overline{Q_{sim}}} - \frac{Q_{obs}(J(k))}{n\overline{Q_{obs}}}\right|. \qquad \text{Eq. 12}$$

Where:

- $npKGE$ = Non-parametric KGE [-]
- r_s = Spearman correlation coefficient [-]
- α_{NP} = Non-parametric form of discharge variability [-]
- β = Bias between simulated and observed mean discharge [-]
- μ = Mean discharge [m^3/s]
- Q = Discharge [m^3/s]
- n = Length of simulated discharge [s]

The simulation duration for the sensitivity analysis is dependent on the availability of discharge data of each individual subbasin. The duration should be at least four years. As this method of creating an optimised parameter is time consuming it was chosen to limit the simulation duration to a maximum of twenty-two years (including a warm up period of two years). In order to keep a reliable comparison, the considered simulation period for each subbasin will be the same throughout the study. The KsatHorFrac value with the highest $npKGE$ is considered as optimised value and acts as the benchmark value for each individual catchment. The predicted KsatHorFrac form the PTF and resulting $npKGE$ from the simulation are compared to this benchmark. After the sensitivity analyses has been conducted, the original PTF will be implemented in the wflow_sbm models. The simulation duration for the models is the same as during the sensitivity analysis in order to make a logical comparison.

The optimised KsatHorFrac is a lumped value. Whereas the predicted KsatHorFrac is distributed. In order to compare the values, the predicted KsatHorFrac is averaged for each subbasin. The difference between the optimised KsatHorFrac and lumped predicted KsatHorFrac is evaluated in order to evaluate the spatial distribution of the results. Subsequently, the difference between $npKGE$ resulting from the optimised, predicted and default (=100) KsatHorFrac is evaluated. Both results are combined to identify subbasins which stand out in terms of KsatHorFrac and/or $npKGE$ difference between the optimised and predicted values. The hydrographs of these subbasins are further analysed in order to evaluate why the subbasins stand out. These subbasins are also addressed in RQ 4.

The optimised and predicted KsatHorFrac and resulting $npKGE$ are compared by conducting statistical tests on the difference between the two scenarios. First a Wilcoxon test is used to identify whether there is a statistically significant difference between the optimised model performance ($npKGE_{opt}$) and model performance resulting from the original PTF ($npKGE_{org}$). Subsequently, as the subbasins are divided into scoring categories (RQ 1), a Kruskal-Wallis test is conducted in order to identify whether there is a statistically significant difference between the different groups.

The PTF is successful if $npKGE_{org}$ is similar to $npKGE_{opt}$, and higher than the $npKGE$ using the default KsatHorFrac (=100) ($npKGE_{def}$). As the scoring categories (Section 3.1.2) are related to the similarities of the SoilGrids parameters, the following hypothesis is set up: *'The difference between the optimised model performance and when the original KsatHorFrac PTF is used is smaller for subbasins with a high score (4-7) compared to subbasins with a low score (0-3)'.* The motivation behind this hypothesis is that the original PTF is trained in GB. Hence, it is expected that the PTF should function well on subbasins which have similar SoilGrids variable values as those in GB.

3.3. Creating an updated PTF

3.3.1. New predictors

The original PTF considers the seven SoilGrids variables which were summarised in Table 2. As other research into KsatHorFrac PTFs is lacking, research into K_{SatVer} formed the basis to consider other variables for the updated PTF. Multiple studies concluded that topographical variables are suitable for predicting K_{SatVer} through PTFs (Gupta et al., 2021; Ayele et al., 2020). Two predictors which were used in multiple studies are slope [-] and elevation [m]. A final predictor which was considered is drainage density [km/km^2]. This predictor is defined as the total river length per unit area (Horton, 1932). Collins & Bras (2010) state that drainage density impacts runoff. Additionally, they compared different researches which stated drainage density is related to environmental factors. Environmental factors, according to Gupta et al. (2021) impact K_{SatVer}. Prior to creating the updated PTF, the three predictors have to be added to the original database which was used within the algorithm. This is done by calculating the subbasin average value of the three predictors for all GB training subbasin considered by Ali et al. (2023). These three values are then added to the seven SoilGrids predictors in the original database, creating a database of ten predictors in total.

3.3.2. Random Forest training and testing

Creating the updated PTF is done in the same subbasins in GB as the original PTF. The database of the ten predictors, together with the optimised KsatHorFrac values for the GB subbasins retrieved from Ali et al. (2023) form the input of the algorithm. Table 3 shows values of important parameters which were considered in the training and testing of the algorithm.

Table 3: Parameter values considered for the RF training and testing in the subbasins of GB, as selected by Ali et al. (2023)

Variable	Value	Variable	Value
Training subbasins	344	Portion of input	1, 0.9, 0.8, 0.6, 0.5, 0.3
Testing subbasins	115	Number of folds	5
Number of trees	10, 50, 100, 150, 200, 300, 400, 500	Random seed	1000
Max depth	None		

To get more insight into the behaviour of the RF algorithm, the relative importance of all considered predictors will be created and analysed. From these results, it can be seen whether the topographic predictors have a large impact on KsatHorFrac or the original predictors are dominant. If a predictor has a small relative importance, it can be chosen to discard this predictor from the algorithm. Due to the random nature of the RF algorithm, the relative importance of the variables is not the same each time the algorithm is run. In order to investigate the sensitivity of this randomness, the algorithm is run ten times. After each run, a PTF file is created. If the outputs of the algorithm show different results, the ten runs should be averaged in order to get one single input into wflow_sbm. However, these PTF files cannot be averaged to one single file. The resulting distributed KsatHorFrac maps can be averaged. Which is described in the following section.

3.3.3. Updated KsatHorFrac map

Each of the ten PTF files is used to create a distributed map of KsatHorFrac in the GB subbasins and validation subbasins from RQ1. Each map is created by combining the 10 input predictors with each PTF file, resulting in ten different KsatHorFrac maps. These ten KsatHorFrac maps are averaged to get one single map which can be put into wflow_sbm. Prior to creating the KsatHorFrac maps, the input maps of the topographic predictors were upscaled from 90 meters to 250 meters to match the resolution of the SoilGrids predictors. This is done in Qgis by using the mean value upscaling technique from the Saga extension (Conrad et al., 2015). In order to get insight into how the updated PTF will behave in the validation subbasins (determined in RQ 1), the predicted KsatHorFrac maps of the updated and original PTF are compared for the GB training subbasins.

3.4. Transferability of the updated PTF

The map with the KsatHorFrac values which result from the updated PTF is implemented in the wflow_sbm model for the validation subbasins from RQ1 in the same way as the original KsatHorFrac map. Subsequently, the models of the subbasins are run for the same time periods as considered in RQ2. The statistical evaluation follows the same method as applied in RQ2. However, here the updated and original predicted KsatHorFrac values, and resulting model performance in wflow_sbm ($npKGE_{upd}$) are compared to the optimised benchmark.

Finally, it is investigated how the hydrographs of the subbasins highlighted in RQ2 have changed as a result of the updated PTF. Additionally, the $npKGE$ statistics (r, alpha, beta) and the KsatHorFrac are compared. From these statistics, it can be seen whether the updated PTF has improved the KsatHorFrac prediction and the model performance of wflow_sbm in the individual subbasins.

4. Results

The results of this research are described in this chapter. First the selection of the validation subbasins is shown in section 4.1. Secondly, the transferability of the original PTF is evaluated in the selected validation subbasins in section 4.2. Thereafter, section 4.3 describes how an updated version of the original PTF has been created, which is compared to the original PTF in the final section of this chapter.

4.1. Subbasin selection

4.1.1. Shortlist

A dataset covering all considered subbasins in GB (provided by Ali et al., 2023)) with depth-averaged, basin-averaged values for the seven SoilGrids variables was analysed, in order to determine the four different ranges: min-max, whiskers, 95th percent, 90th percent. After selecting only the 344 training subbasins from the dataset, the boxplots in Figure 10 were created. From these boxplots, the min-max range and the range of the whiskers can be determined. As can be seen in the boxplots, only bedrock depth, organic carbon content and pH index have clear outliers as indicated by the dots beyond the whiskers.

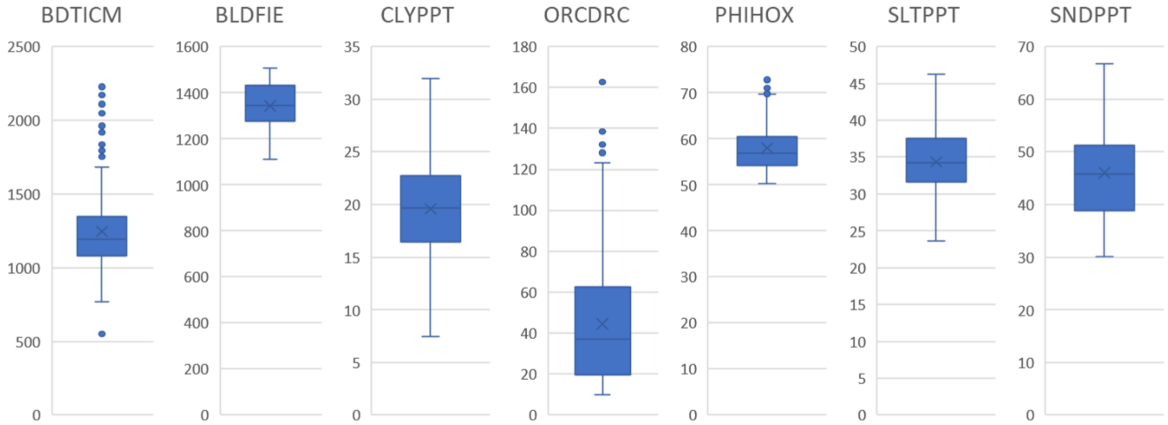


Figure 10: Boxplots of the trainings subbasins for the seven SoilGrids parameters which were used to train the ML algorithms of Ali et al. (2023). Here the whiskers of the boxplot are 1.5 times the inter quartile range (Q3 +1.5(Q3-Q1) and Q1 -1.5*(Q3-Q1).*

Even though the remaining variables do not have outliers in the data, it is still interesting to investigate how the range changes if extremes are removed from the data. Hence, looking at the middle 95 and 90 percent of the dataset. The boundaries of the four ranges are stated in Table 11 in Appendix B.1. All 967 subbasins from the longlist (Figure 9 in Section 3.1) receive a score based on the four ranges of the SoilGrids variables as described in Section 3.1.2. The distribution of the scoring categories for the four ranges is shown in Table 4. The distribution of the scores is skewed towards the higher half of the scores for all four ranges. Especially for the complete range and the whiskers a small number of subbasins can be found, with zero subbasins scoring 0 points when the complete range of all SoilGrids variables are considered.

Table 4: Number of subbasins within a scoring category for the four ranges of the SoilGrids variables of the GB training basins

Scores	Min-max	Whiskers	95th	90th
0	0	2	11	24
1	3	26	25	40
2	20	32	42	55
3	54	45	124	134
4	85	144	111	102
5	181	139	181	185
6	157	209	173	181
7	467	370	300	246

In order to investigate the relatively small number of subbasins with a low score (0-2), the SoilGrids variables are analysed in more detail. The scoring system assumes that all variables are independent from each other. This is not the case for the percentages of clay (CLYPPT), silt (SLTPPYT) and sand (SNDPPT) as these form a soil type together. The percentages of the dominant soil types per basin are shown in Figure 11. Most of the subbasins share the same soil types with GB: sandy loam, loam and clay loam. The Seine has a different distribution with respect to the other basins: with more silt loam (22.52%), more silty clay loam (8.11%) and no sandy loam at all. It can be seen that if the complete dataset of the training basins in GB is considered in the analysis, no subbasins score 0 points, even though these dominant soil types are different compared to GB. Overall, the analysis of the soil types did not reveal progression towards a larger selection of subbasins with a low score for the complete dataset. However, since the dominant soil type in the Seine is completely different compared to GB, it is expected that when the extremes are removed from the dataset, the subbasins with a low score are located in the Seine basin.

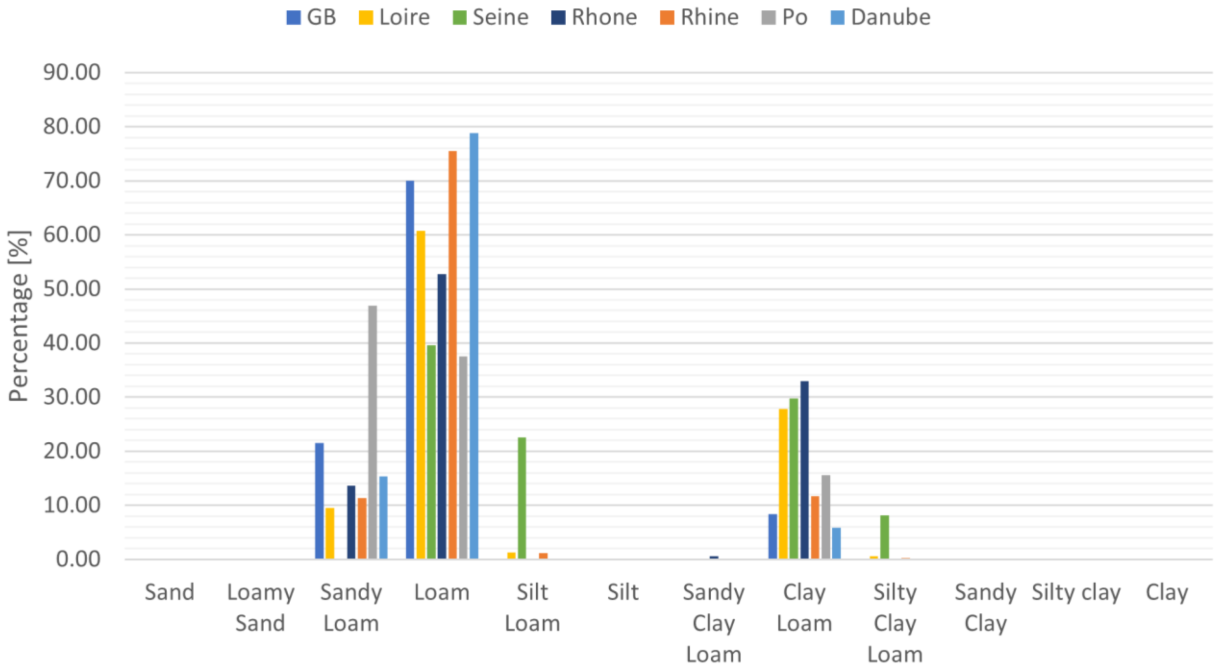


Figure 11: Distribution of dominant soil types per basin. The percentages are determined by the number of subbasins within a soil category as part of the complete basin.

If the 90th percentile of the data is considered, the number of subbasins with a score of 0 increases to 24. This is only 2.5 percent of the subbasins in the longlist. Because of this, the categories with a score of 0-2 as well as 5-7 are combined, reducing the shortlist to 731 subbasins.

4.1.2. Final selection

The final selection was made based on the middle 90% of the GB training basins dataset since there the most subbasins within the scoring category 0 was found (Table 4). The group with a score of 0 is the limiting factor since it includes the least number of subbasins. Hence, from the groups with a score of 1, 2, 5, 6 and 7 an equal amount of subbasins is selected (24). The subbasins are selected based on the size of the subbasins from the group with a score of 0. Even though the equal distribution is size was the main focus, the location was taken into account in order to ensure an equal spread across the basins. Using a Kruskal-Walis test is was made sure that the size of the subbasins are not significantly different amongst the scoring categories (H-value(3.576) < chi squared (13.338), with Df = 5). This resulted in 144 potential validation subbasins.

After building the models using HydroMT, it was found that the for several subbasins the size between the model domain and reality differed more than 15%. As deviations of this magnitude could result in unreliable model results, these subbasins were discarded. This decreased the amount of subbasins to 107. However, the distribution among the score categories was not equal and ranged between 17 and 19. Five additional subbasins were discarded to keep the amount of subbasins in all score categories equal at 17. This reduced the amount of validation basins from 144 to 102. The subbasins which were discarded were selected based on size such that the size of the subbasins among the scoring categories remained equal (Kruskal-Wallis test: H-value (12.578) < chi squared (13.338), with Df = 5). The locations of the subbasins in the final selection can be seen in Figure 12. Here the colour indicates the score. From Figure 12, it can be concluded that the distribution of subbasins within a scoring category is not even across the six basins. The subbasins with low scores are primarily concentrated in the Seine basin. Which could be attributed to the different soil types in the Seine as shown in Figure 11. Subbasins with a high score can be found throughout all basins are spread more evenly.

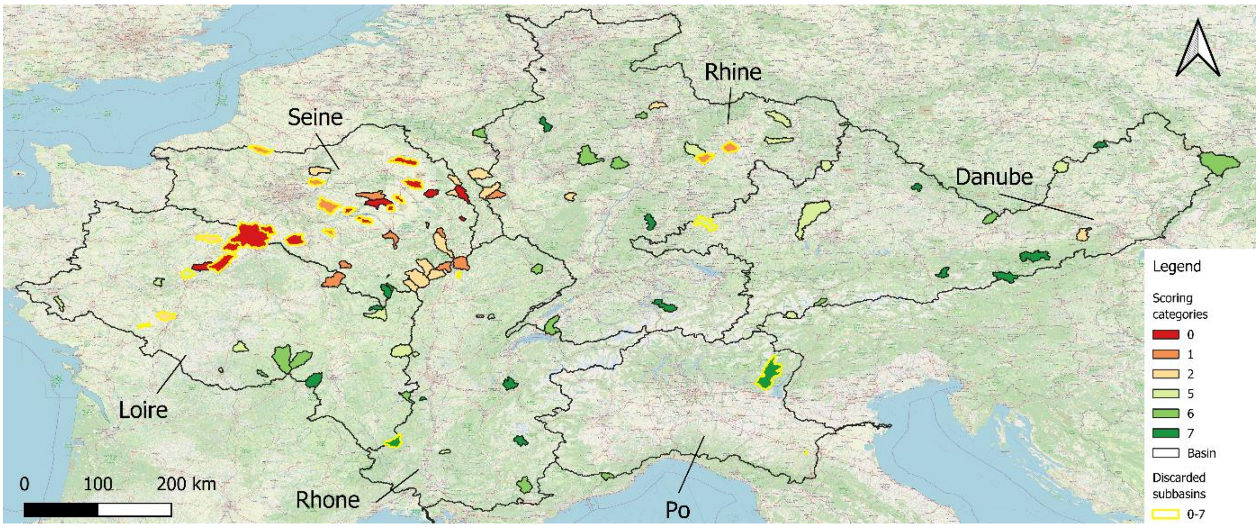


Figure 12: Final selection of validation basins, yellow highlighted subbasins will be explained in the Section 4.2.1

4.2. Evaluating the transferability of the original PTF

This Chapter includes the results of Research Question 2. First, the optimised and predicted KsatHorFrac values are presented and discussed. Secondly, the Non-Parametric Kling Gupta Efficiency (np KGE) for the default, optimised and RF simulations are described. After the description of the results for all validation subbasins, a number of subbasins will be evaluated and described in more detail to highlight causes for observed differences in the both KsatHorFrac value and resulting model performance in wflow_sbm.

4.2.1. Unsuitable validation subbasins

After assessing the hydrographs which resulted from the optimised KsatHorFrac in the sensitivity analysis, it was found that 26 subbasins provided unreliable simulations. This was based on the np KGE, corresponding β, sensitivity curve and visual inspection. In most of the discarded subbasins the simulated discharge overestimated the measured discharge. These subbasins were mainly located in the Seine subbasin. Furthermore, it was found that several hydrographs displayed recognisable patterns which suggested the presence of hydraulic structures which regulated river discharge. The subbasins which are eliminated are highlighted in yellow in Figure 12. An overview of the changes per scoring category can be seen in Table 5. Here it can be seen that a large number of subbasins is removed in the scoring categories 0 and 1. Therefore, it was decided to group these two categories together when statistical tests are conducted. After removing the 26 subbasins, the size of the subbasins amongst the scoring categories was not statistically significant (Kruskal-Wallis test: H-value (11.42) < chi squared (11.67), with Df = 4).

Table 5: Overview of the number of subbasins after RQ1 and at the start of RQ2 per scoring category

	0	1	2	5	6	7	Sum
RQ1	17	17	17	17	17	17	102
Removed	12	7	3	2	0	2	26
RQ2	15		14	15	17	15	76

4.2.2. Comparing optimised and predicted KsatHorFrac from the original PTF in the validation subbasins

The optimised KsatHorFrac can be seen in Figure 13 A. This figure shows that the spread of optimised KsatHorFrac values covers the complete input range (10^0-10^4). Generally the optimised values are of a similar magnitude in the same area, which is as expected. Mainly high values ($10^{3.5}$-10^4) can be found in the eastern part of the Seine basin. However, some subbasins in the Danube, Rhone and Po stand out where high values can be found in between lower values (KsatHorFrac < $10^{2.5}$). The subbasin averaged KsatHorFrac values resulting from the original PTF are shown in Figure 13 B. From this figure it becomes clear that the highest subbasin averaged predicted KsatHorFrac is a power of magnitude smaller compared to the highest optimised value: the spread of the predicted KsatHorFrac is between 10 and 1000, with the majority of the values between 100 and 1000. This upper limit can also be observed in the results of Ali et al. (2023). In the GB testing subbasins the predicted KsatHorFrac was under-estimated in comparison to the optimised KsatHorFrac for values around 10^3 and higher (Figure 7). Because of this, no clear outliers can be identified in Figure 13 B as generally subbasins in close proximity of each other have similar parameter values. The difference between the optimised and predicted KsatHorFrac values can be seen in Figure 13 C. Blue indicates the subbasins where the predicted KsatHorFrac has a higher averaged value. The optimised KsatHorFrac is higher in the red subbasins. A large set of red subbasins can be identified in the east of the Seine basin. For these subbasins, the absolute difference between the predicted and optimised KsatHorFrac is in the range of 1,000 to 10,000.

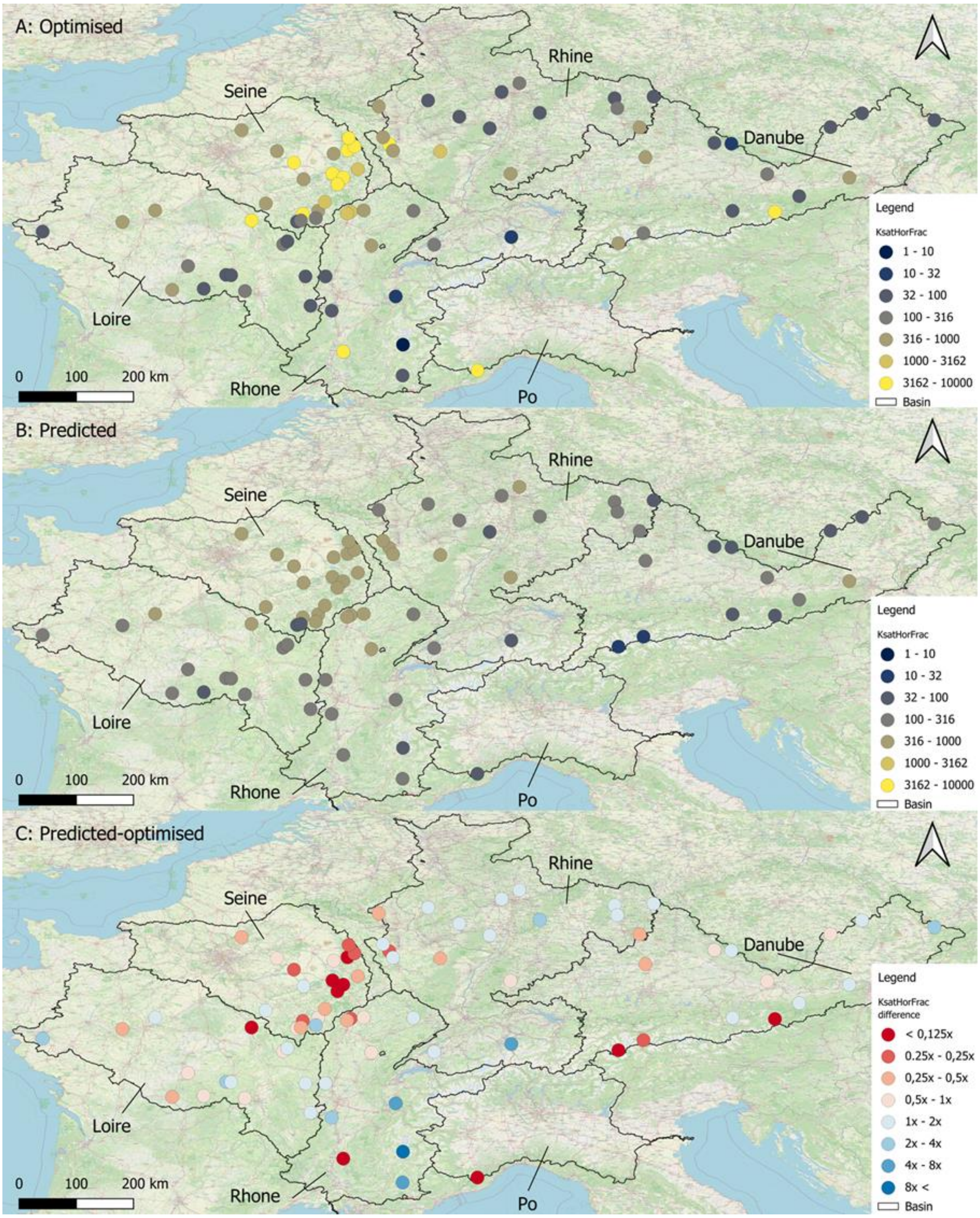


Figure 13: Subbasin averaged (A) optimised and (B) predicted KsatHorFrac values. (C) Difference between predicted and optimised KsatHorFrac values, with values below one indicating lower KsatHorFrac values for predicted than for the optimised values and vice versa.

The optimised and predicted KsatHorFrac values from Figure 13 A and B are shown in Figure 14 A. From this figure it can be concluded that the predicted KsatHorFrac in the subbasins with a high score (5-7) is lower compared to subbasins with a low score (0-2). It also appears that the subbasins with a high score follow the 1:1 line better compared to the subbasins with a low score. This suggests that the KsatHorFrac prediction is lacking in subbasins with different soil characteristics compared to the GB training subbasins. Furthermore, it can be seen that an optimised KsatHorFrac of 10,000 (10^4) is found in several subbasins, but this is not the case for the predicted KsatHorFrac, where the maximum value is only 699 ($10^{2.84}$). Additionally, this highest value only applies to subbasins with a low score. The difference between the predicted and optimised KsatHorFrac per scoring category is shown in Figure 14 B. Using a Wilcoxon test it was found that the differences of groups 5, 6 and 7 are insignificant. The differences of groups 0-1 and 2 are statistically significant with a p-value of 0.0043 and 0.025 respectively. This means that the predicted KsatHorFrac is significantly lower compared to the optimised KsatHorFrac for subbasins with low score. This difference can be caused by the fact that the optimisation in some subbasins with a low score (0-2) is not working to the fullest potential as these subbasins tend to have a 'poor' model performance in general (described in Discussion Section 5.2.1). This is caused since the optimised KsatHorFrac value is dependent on the np KGE. Hence, the value of KsatHorFrac resulting from the sensitivity analysis could be different from the actual optimised KsatHorFrac value for subbasins with a 'poor' model performance.

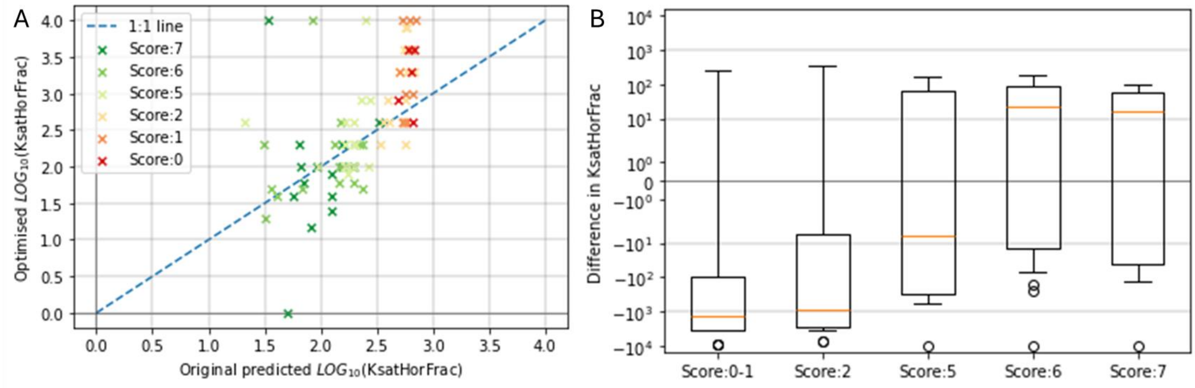


Figure 14: (A) Scatterplot of, and (B) difference in boxplot between, the predicted KsatHorFrac and the optimised KsatHorFrac per scoring category

4.2.3. Comparing the model performance of wflow_sbm resulting from different KsatHorFrac values

$npKGE_{def}$, $npKGE_{opt}$, and $npKGE_{org}$ are shown in Figure 16 A, B and C respectively. Based on a visual inspection, the three scenarios show similar results in terms of magnitude and spatial distribution. The majority of the subbasins has a $npKGE$ of at least 0.5, with 78%, 79%, and 75% for default, optimised and predicted respectively (Figure 15 A). This implies the majority of the basins perform intermediate (0.5-0.75) to good (>0.75) (Rogelis et al., 2016). However, the worst performing basins in terms of np KGE (0 - 0.4) can be found in and around the Seine, with the exception of the east of the Seine basin where the KsatHorFrac is higher compared to the rest of the basin.

The most prominent improvements of the $npKGE_{opt}$ in relation to $npKGE_{def}$ can be found in the southeast of the Seine basin where several subbasins have crossed the threshold of 0.8. the improvement ranges between 0.01 and 0.21 A similar trend can be observed in the east of the Rhine basin. For the predicted KsatHorFrac, this improvement is less evident as the $npKGE$ values are mostly in the same intervals. The differences in $npKGE$ can also be seen in the boxplots in Figure 15 A. From this figure, it can be concluded that that the median $npKGE_{opt}$ is the highest, followed by $npKGE_{org}$ and $npKGE_{def}$ respectively.

However, in order to validate this, the changes in $npKGE$ between the default, predicted and optimised are evaluated. The comparison between optimised and default, predicted and default, and predicted and optimised are shown in Figure 16 D, E and F, respectively. As shown in Figure 16 D, nearly all basins have increased in performance when the optimised and default scenarios are compared. From the 76 subbasins, 14 (18.4%) have an equal KsatHorFrac value of 100 and thus an equal performance. The subbasins with the largest improvement can be seen in the south east of the Seine basin, which is in line with the increased predicted KsatHorFrac values showed in Figure 13 B.

The difference between $npKGE_{org}$ and $npKGE_{def}$(Figure 16 E) is concentrated around zero: 63 subbasins are in the range of -0.06 and 0.06 (82.9%). In total, the predicted KsatHorFrac outperforms the default value in 49 of the 76 subbasins (64.5%). However, in one subbasins the $npKGE$ of the default KsatHorFrac outperforms the predicted value by 0.18. This subbasin can be found in the upstream part of the Danube basin (orange red) and is described in more detail in Section 4.2.4.

As shown in Figure 16 F, $npKGE_{opt}$ outperforms $npKGE_{org}$ in a large number of subbasins. This result is in line with the expectation, as the optimised KsatHorFrac was derived using observed discharge during the sensitivity analysis. However, the predicted KsatHorFrac is not outperformed by the optimised value in six subbasins. From these six subbasins the largest increase in terms of np KGE is small with only 0.003. However, this difference is negligible and are thus considered equal. In one subbasin the optimised KsatHorFrac outperforms the predicted KsatHorFrac by 0.18 in terms of np KGE. This is the same subbasin as in Figure 16 E. All previous observations regarding the difference in np KGE between two scenarios are summarised in Figure 15 B.

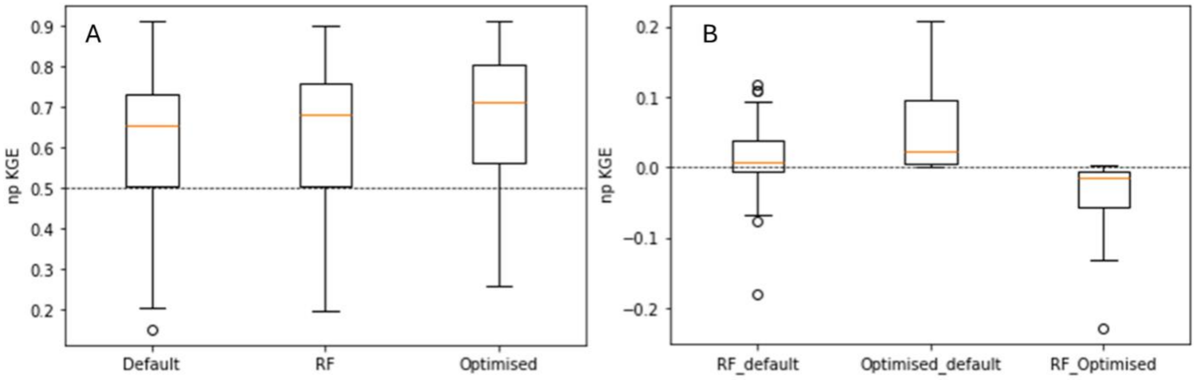


*Figure 15: (A) Boxplots of np KGE for the default, RF and optimised simulations. Whiskers represent Q1-1.5*IQR, RQ3+1.5*IQR. Dotted line at 0.5 represents the threshold between poor (np KGE <0.5) and intermediate(0.5< np KGE) performance. (B) Boxplots of the difference in np KGE between the default, Predicted and optimised simulations. Dotted line indicates the threshold where the np KGE between the scenarios is equal.*

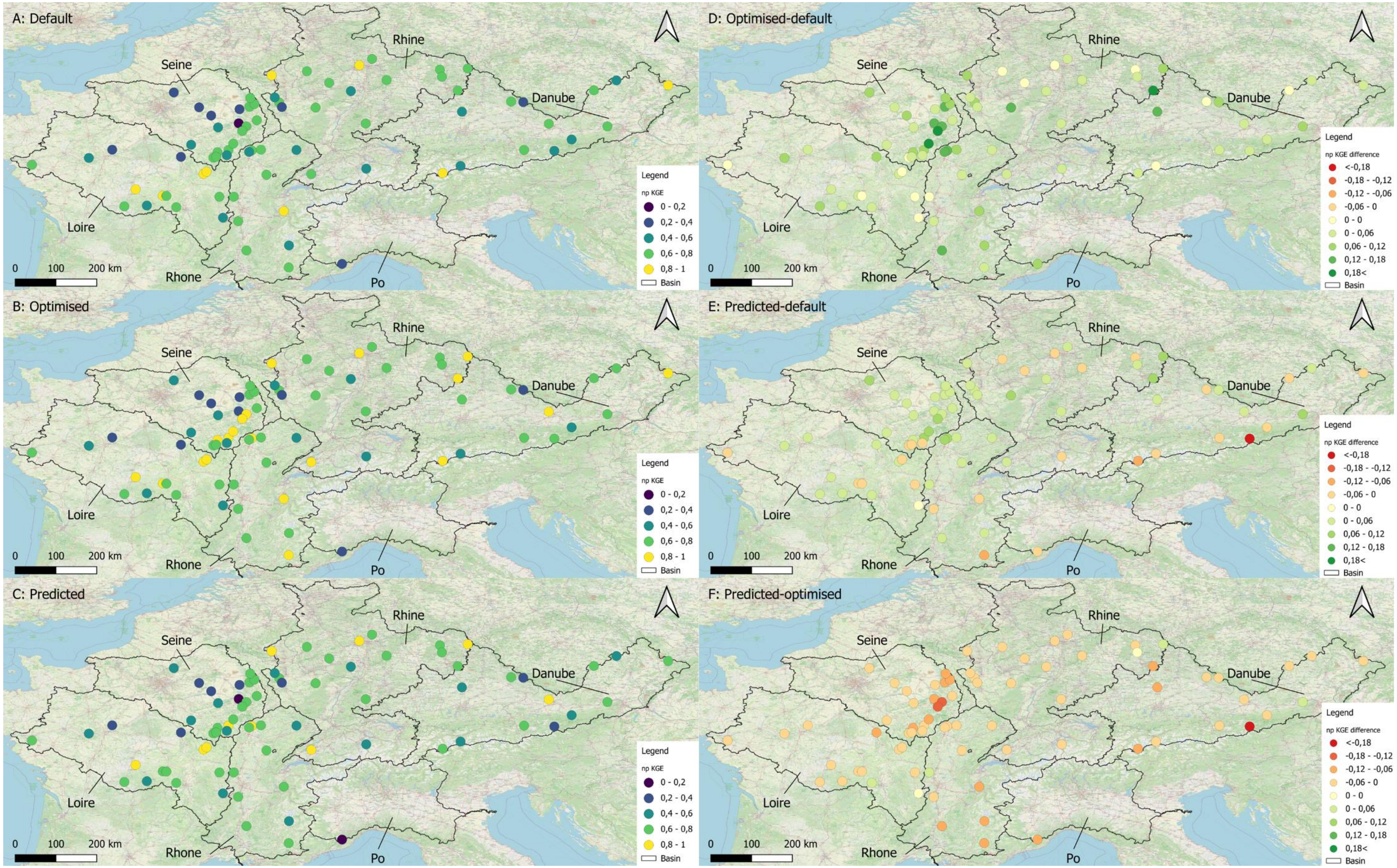


Figure 16: Left: Non-Parametric KGE for the (A) default, (B) Optimised, and (C) predicted KsatHorFrac values. Right: Difference in Non-Parametric KGE between (D) optimised and default, (E) predicted and optimised and (F) predicted and default KsatHorFrac values

$npKGE_{org}$ and $npKGE_{opt}$ are plotted against each other in Figure 17 A for the different scoring categories. When focussing on the optimised KsatHorFrac, it becomes evident that the subbasins with a poor performance ($npKGE$ <0.5) are mainly subbasins with a low score (0-2). 16 subbasins have a performance below this threshold. From the subbasins with a low score (0-2), 37,93% also have a poor performance. For subbasins with a high score (5-7) this is only 10.62%. This suggests that subbasins with a low score generally perform worse compared to subbasins with a high score when using wflow_sbm. For the predicted KsatHorFrac, 19 subbasins have a poor performance. Here the split between low (0-2) and high (5-7) scoring categories is 41.38 % and 14.90 % respectively, indicating that basins with a poor performance generally fall in scoring category 0 - 2.

The difference between $npKGE_{org}$ and $npKGE_{opt}$ from Figure 16 F, combined with the scoring categories is shown in Figure 17 B. Using a Wilcoxon test, it was found that the differences between the $npKGE$ are significant in all scoring categories (p ranges from 6.1e-4 to 6.1e-5), indicating that the $npKGE$ of the simulations with the predicted KsatHorFrac are significantly lower than those of the optimized KsatHorFrac. Additionally, a Kruskal-Wallis test was used to prove that the difference between the scoring categories are not statistically significant (p = 0.78). This rejects the hypothesis that stated that the difference between the optimised model performance and when the original PTF is used is smaller for subbasins with a high score compared to subbasins with a low score.

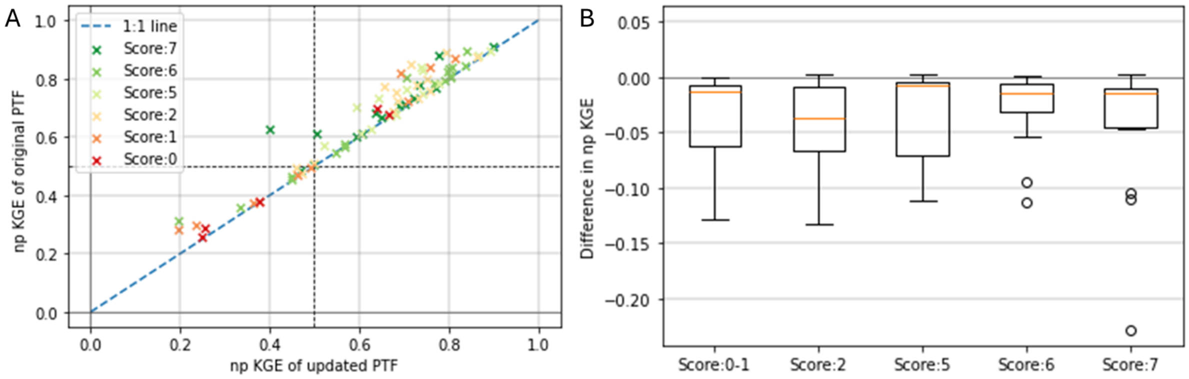


Figure 17: (A) Scatterplot of, and (B) difference in boxplots between, the predicted np KGE against the optimised np KGE per scoring category

4.2.4. Evaluating stand-out subbasins

Selzthal subbasin

The most prominent outlier in terms of both $npKGE$ and KsatHorFrac difference between the predicted and optimised scenario is subbasin Selzthal in the Danube basin. The subbasin has a score of 7 and an optimised and predicted KsatHorFrac value of 10,000 and 34.17 respectively. The subbasin experienced a decrease $npKGE$ when applying the predicted KsatHorFrac (0.40) in comparison to the default (0.58) and optimised (0.63) KsatHorFrac. The hydrographs resulting from the optimised and predicted KsatHorFrac and a magnified section can be seen in Figure 18 B and Figure 19 B respectively. It stands out that the baseflow resulting from the predicted KsatHorFrac is higher compared to the optimised hydrograph. This is not logical since a higher KsatHorFrac should result in a higher baseflow. Furthermore, it can be concluded that both the optimised and predicted KsatHorFrac simulation underestimate the peak discharge. This can be seen in the statistics of the $npKGE$: The Beta values of the optimised and predicted are 0.71 and 0.70 respectively, supporting the underestimation of both simulations. The optimised simulation has a correlation (r) of 0.77, which is 0.49 for the predicted simulation. This difference can clearly be seen in the magnified sections of both hydrographs. The optimised simulation in Figure 18 B follow the same trajectory and timing compared to the measured discharge. However, the peaks are unrealistically smooth compared to the measured discharge. The smoothness of the predicted discharge in Figure 19 B is more logical besides the period between April and June, where the simulated discharge misses the higher discharges completely. It is possible that this could be attributed to snowmelt which appears not to be captured by the predicted simulation.

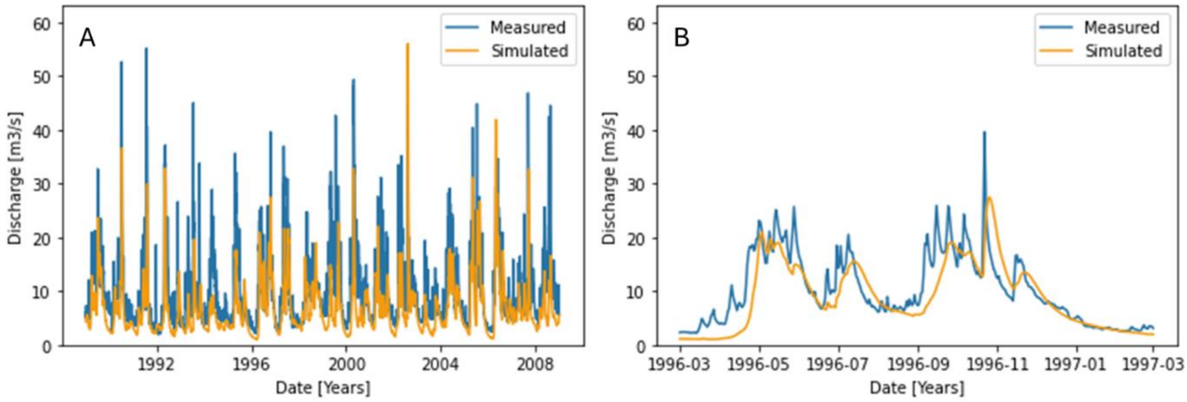


Figure 18: (A) Hydrograph of subbasin Selzthal in the Danube basin using the optimised KsatHorFrac (10,000) and (B) magnified section between 03-1996 and 03- 1997

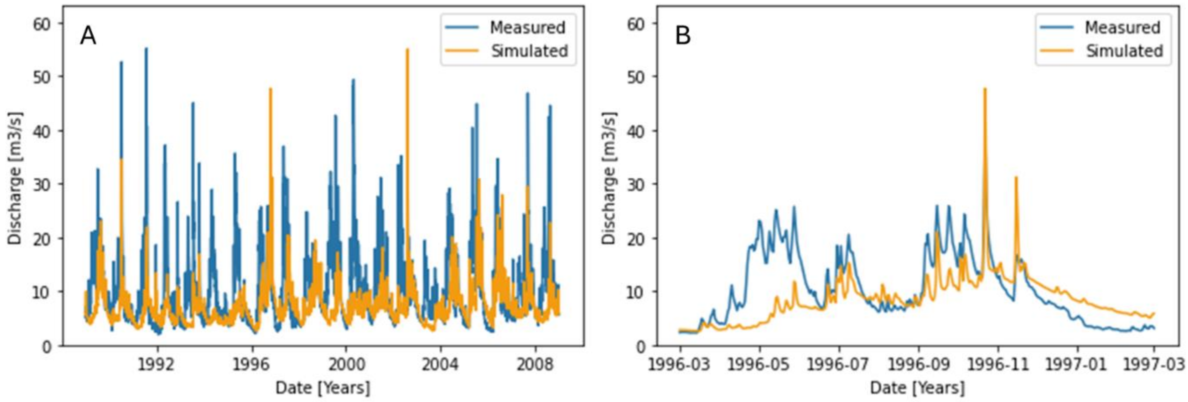


Figure 19: (A) Hydrograph of subbasin Selzthal in the Danube basin using the predicted KsatHorFrac (avg. 34.17) and (B) magnified section between 03-1996 and 03-1997

Blaise subbasin

Another outlier which had one of the largest differences in $npKGE$ and KsatHorFrac values between the predicted and optimised scenario is the Blaise subbasin, located in the Seine basin with a score of 0. The $npKGE_{org}$ is 0.69 whereas the $npKGE_{opt}$ is 0.82. When comparing both hydrographs in Figure 20 A and Figure 21 A it can be seen that the predicted KsatHorFrac simulation produces higher peaks. Additionally, the baseflow appears to be lower compared to the optimised simulation. This can also be seen in the $npKGE$ statistics, as the optimised simulation has a higher correlation, alpha and beta. Especially the alpha has a lower value, which indicates the variability of the discharge is larger in the predicted KsatHorFrac simulation. These changes become more clear when magnifying a section of the hydrographs. Figure 20 B shows that optimised simulation follows both the peaks and baseflow of the measured discharge well. This is not the case for the predicted KsatHorFrac simulation in Figure 21 B where almost every peak is overestimated. The predicted KsatHorFrac is 539.22 whereas the optimised KsatHorFrac is 10,000. This causes the baseflow to be underestimated and the peaks over estimated. Additionally, sharp drops can be seen after peaks. This is caused by the relatively small KsatHorFrac value: the horizontal saturated hydraulic conductivity is small, which results in less baseflow and an increase in fast runoff in the form of overland flow.

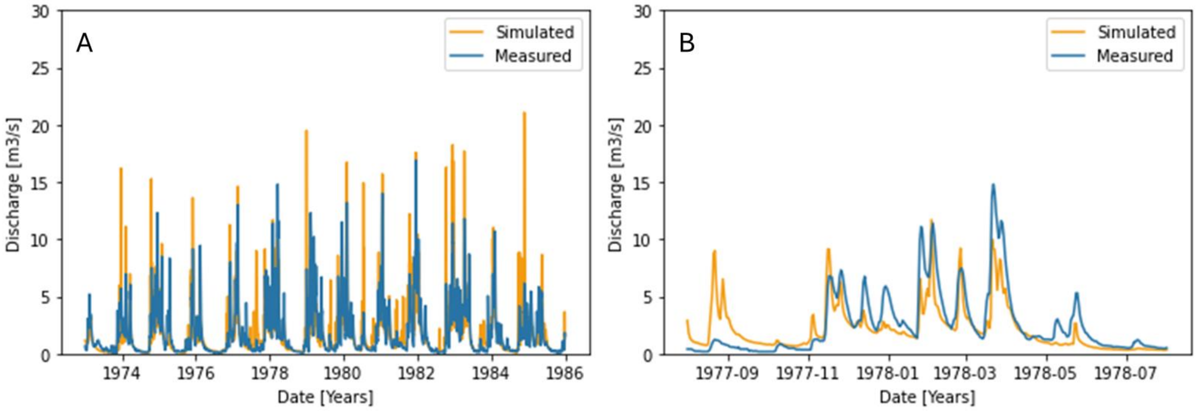


Figure 20: (A) Hydrograph of subbasin Blaise in the Seine basin using the optimised KsatHorFrac (10,000) and (B) magnified section between 08-1997 and 08-1998

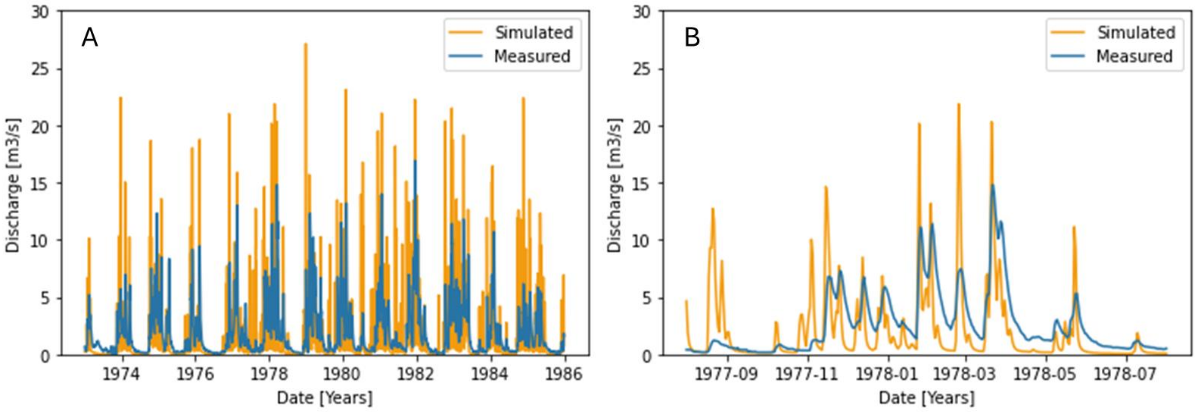


Figure 21: (A) Hydrograph of subbasin Blaise in the Seine basin using the predicted KsatHorFrac (avg. 539.22) and (B) magnified section between 08-1997 and 08-1998

4.3. Results from an updated PTF that takes into account topographic predictors

4.3.1. New topographic predictors

The subbasin-average values of the elevation, slope and drainage density are added to the input dataset of subbasin-average SoilGrids variables. The average relative importance of the ten runs is shown in Figure 22. The average performance indicators across the ten runs are: RMSLE = 0.2608 and r^2 = 0.9039. This performance in nearly identical to the performance of the original PTF (Figure 7). The details of the ten individual runs are shown in Figure 36 in Appendix C.1. Overall, the SoilGrids variables are more important compared to the topographic variables which are added. Even though RF algorithms are robust against overfitting (Breiman, 2001), it was decided to discard drainage density and absolute depth to bedrock (BDTICM) from the algorithm since their relative importance is only 2.39% and 2.60% respectively.

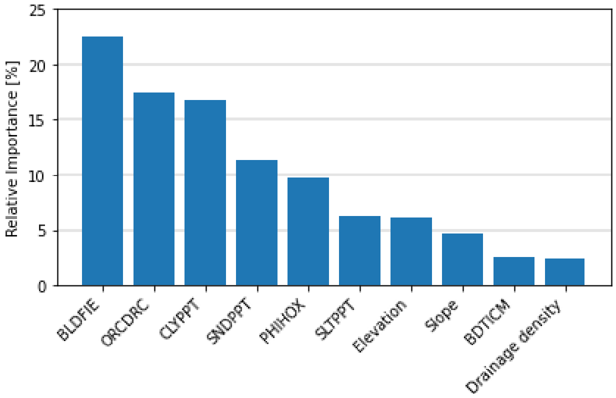


Figure 22: Ten run average of the relative importance when elevation, slope and drainage density are added to the SoilGrids v1.0 variables selected by Ali et al. (2023)

Running the RF algorithm with the updated selection of variables ten times resulted in the relative importance in blue in Figure 23 (comparison with 100 runs is described in the discussion, Section 5.1.2). Removing the drainage density and depth to bedrock has reduced the spread in relative importance. The average performance indicators across the ten runs are: RMSLE = 0.2657 and r^2 = 0.9005. The details of the ten individual runs are shown in Figure 37 in Appendix C.2. The importance of bulk density (BLDFIE) has reduced from 22.58 % to 18.01% whereas slope increased from 4.70% to 7.71% in relation to the relative importance from Figure 22. However, this change could be attributed to the removal of the least important predictor: Bedrock depth. Removing the drainage density and bedrock depth has also altered the order of importance as elevation has exceeded SLTPPT. Although the difference is only fractional: the importance of elevation went from 0.16% below to 0.29% above SLTPPT. Therefore, this change could also be attributed to the algorithm only being run 10 times. The relative importance of the original PTF can also be seen in Figure 23 (orange). All SoilGrids parameters reduced in importance when the updated predictors are implemented, except for ORCDRC. It can also be seen that the less important predictors PHIHOX and SLTPPT from the original PTF show a larger reduction compared to the more important predictors.

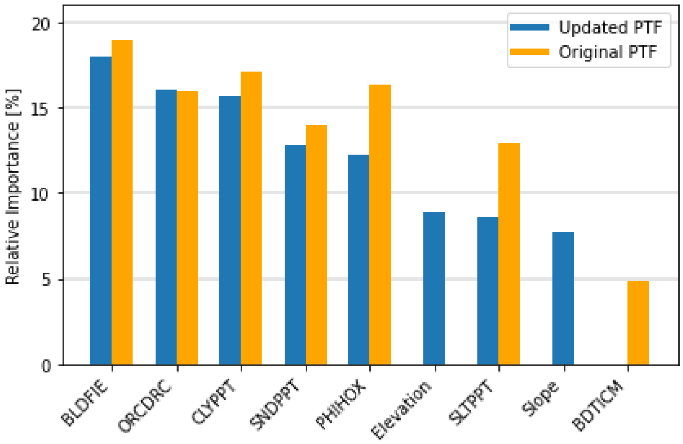


Figure 23: Ten run average of the relative importance of the predictors for the updated PTF (blue) and relative importance of the original PTF

4.3.2. Spatial differences between the original and updated PTF

Besides the shift in relative importance of the predictors after ten runs, the spatial differences are also investigated in the GB subbasins. The average difference between the lowest and highest predicted KsatHorFrac value between the ten distributed maps was found to be 33% (map in Appendix C.3.). Because of this difference between extremes, it is logical to take the average value of the ten maps.

Figure 24 shows the ratio between the predicted KsatHorFrac values from the updated (average from ten maps) and original PTF in the training subbasins. As can be seen, most of the area is green, indicating the predicted KsatHorFrac is higher for the updated PTF compared to the original PTF. The predicted KsatHorFrac of the original PTF exceeds that of the updated PTF only locally. In the extreme cases, the updated PTF predicts 5.4 times higher values compared to the original PTF. The original PTF on the other hand predicts at maximum 4.3 times higher values compared to the updated PTF. On average, the updated PTF predicted 1.27 times higher KsatHorFrac values compared to the original PTF. From the figure, it can clearly be seen that overall the updated PTF predicts higher KsatHorFrac values compared to the original PTF (validated by evaluating the spread of the histogram (Appendix C.4.)). This is a positive change since the original PTF under-estimates the KsatHorFrac in the GB testing subbasins for values around 10^3 and higher. Since the predicted KsatHorFrac values in the GB training subbasins are higher when applying the updated PTF, it is expected that the predicted KsatHorFrac will be higher in the validation subbasins as well. Which again is positive as the predicted KsatHorFrac in subbasins with a low score (0-2) was underestimated with respect to the optimised KsatHorFrac.

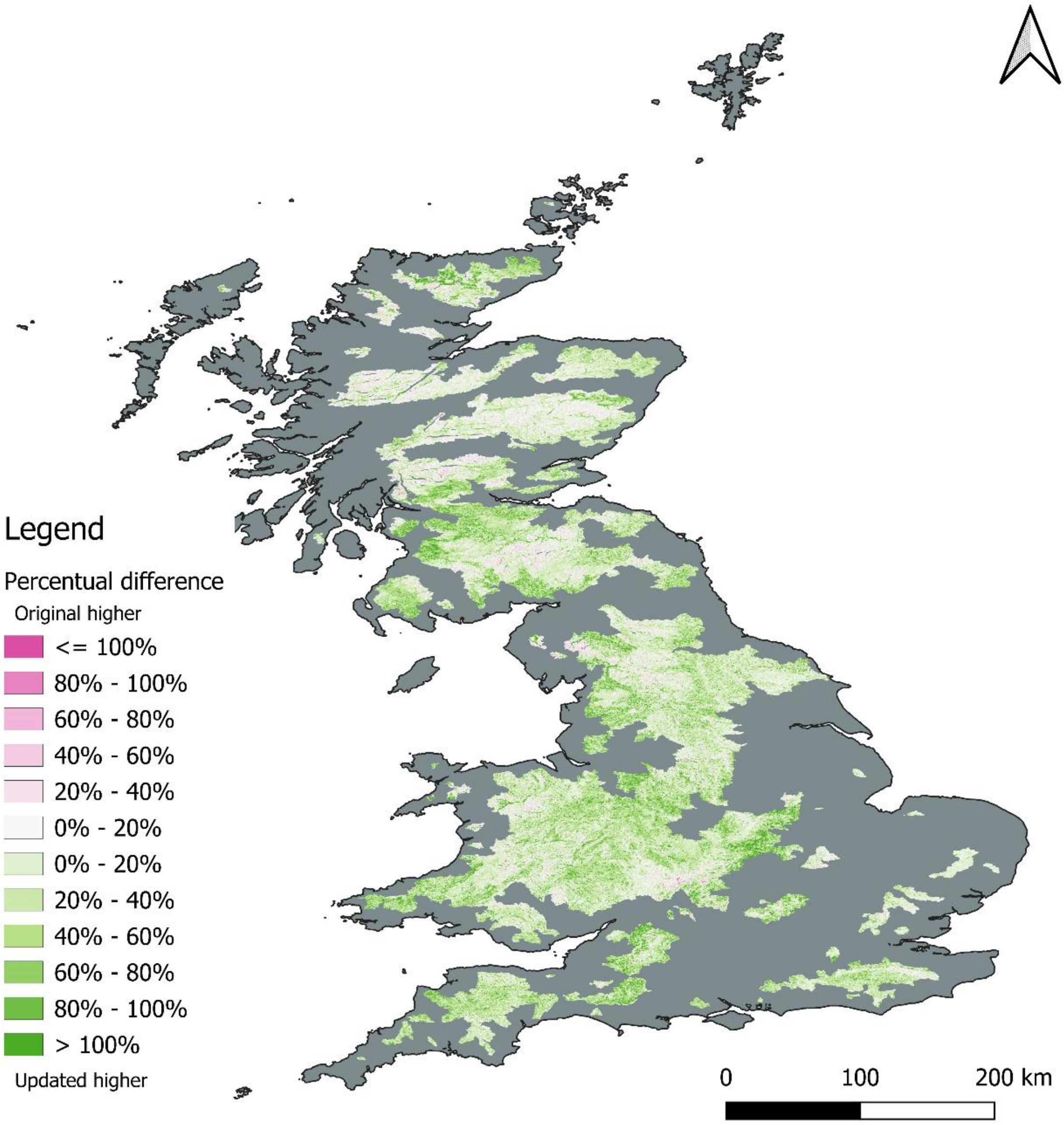


Figure 24: Ratio of predicted KsatHorFrac values between the updated and original PTFs in the GB training subbasins. Green indicates the updated PTF predicts a higher parameter value. Pink indicates the original PTF predicts a higher parameter value

4.4. Evaluating the transferability of the updated PTF

4.4.1. Relating the change in predicted KsatHorFrac to the optimised benchmark

The difference in KsatHorFrac between the updated PTF and the original PTF per scoring category is shown in Figure 25 A. Values above zero indicate the updated PTF predicts higher values compared to the original PTF. Using a Wilcoxon test it was determined that the updated PTF predicts significantly higher values in subbasins with scoring categories: 0-1, 2 and 5. This is in line with expectations based on the results from Section 4.3.2. Subbasins with a score of 7 receive a significantly lower KsatHorFrac value. Even though the updated PTF predicts lower values in subbasins with a score of 6 compared to the original PTF, the difference is not statistically significant. The corresponding p-values can be found in the top row of Table 6. These results follow the trend which can be seen in Figure 24 as the updated PTF generally predicts higher values in the GB training subbasins compared to the original PTF.

Figure 25 B shows the difference in KsatHorFrac between the updated PTF and optimised parameter values. Values above zero indicate the updated PTF predicts higher values compared to the optimised parameter values. The difference of the median between Figure 14 B and Figure 25 B have only marginally changed. However, the increase in predicted KsatHorFrac values in scoring categories 0-1 and 2 from Figure 25 A did increase the spread of the IQR of those scoring categories. Using a Wilcoxon test is determined that the differences between predicted and optimised KsatHorFrac are not statistically significant, for all scoring categories (p-values in second row of Table 6). This is a change in comparison to the difference in KsatHorFrac value between the original PTF and the optimised parameter value as described in Section 4.2.2 (Figure 14 B). There, the KsatHorFrac in scoring categories 0-1 and 2 were significantly lower for the predicted value compared to the optimised KsatHorFrac (p-values in bottom row of Table 6). This suggests the predicted KsatHorFrac values from the updated PTF are closer to the optimised KsatHorFrac values.

Based on these results, it is expected that the model performance of wflow_sbm improves for subbasins with scoring categories 0-1, 2 and 5 while it should slightly decrease for scoring category 7 when considering the updated and original KsatHorFrac. Furthermore, based on the differences between Figure 14 B and Figure 25 B it is hypothesised that the most prominent changes in np KGE resulting from the updated and optimised can be observed in scoring categories 0-1 and 2.

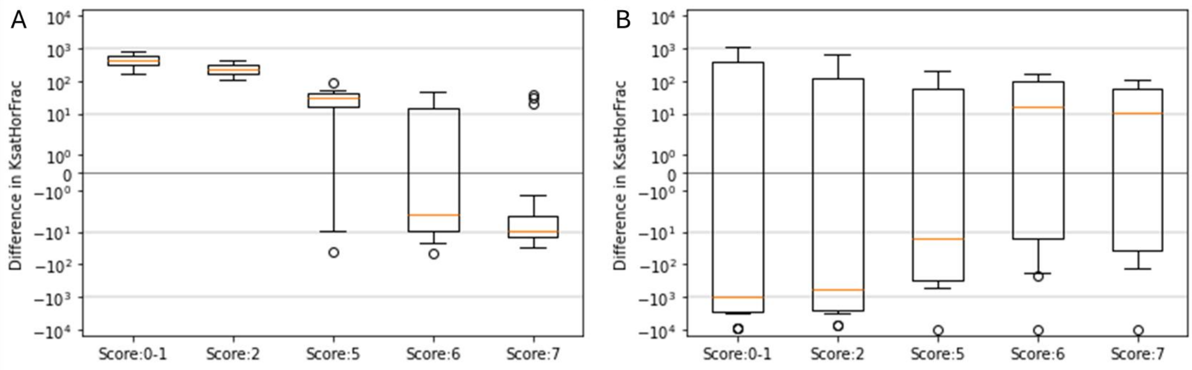


Figure 25: Boxplot of difference in KsatHorFrac between the updated PTF and (A) original PTF and (B) optimised KsatHorFrac, per scoring category

Table 6: P-values resulting from a Wilcoxon test for the difference in KsatHorFrac value, bold numbers indicate a statistically significant difference

	0-1	2	5	6	7
Updated - original	**6.10e-5**	**1.22e-4**	**0.048**	0.31	**0.026**
Updated - optimised	0.15	0.15	0.64	0.51	0.45
Original – optimised	**4.27e-3**	**0.025**	0.25	0.40	0.28

4.4.2. Comparing the change in model performance to the optimised benchmark

The difference between $npKGE_{upd}$ and $npKGE_{org}$ per scoring category is shown in Figure 26 A. Using a Wilcoxon test, it was determined that there is a statistically significant difference in $npKGE$ for subbasins with a score of 0-1. Even though the median value shows that the updated PTF outperforms the original PTF in scoring category 2, the difference is found to be insignificant. The difference of the remaining scoring categories are closer to zero, also resulting in a statistically insignificant difference (p-values in top row of Table 7). The increased $npKGE$ in scoring categories 0-1 and 2 are in line with the hypothesis from the previous section even though the increase in scoring category 2 is not statistically significant. For scoring categories 5 and 7 the results are not in line with the hypothesis. Here it was expected that wflow_sbm performed better when using the updated PTF in comparison to the original PTF for scoring category 5. The opposite was expected for scoring category 7 as the updated PTF predicted lower KsatHorFrac values.

Figure 26 B shows the difference between $npKGE_{upd}$ and $npKGE_{opt}$. Using a Wilcoxon test is was determined that there is a statistically significant difference in all scoring categories (p-values in second row of Table 7). Even though $npKGE_{upd}$ was significantly higher compared to $npKGE_{org}$ for the scoring category 0-1, the comparison between the updated PTF and optimised values yield nearly the same result as in Section 4.2.3 (Figure 17 B). There, the $npKGE_{org}$ was significantly lower compared to $npKGE_{opt}$ for all scoring categories (p-values in bottom row of Table 7). Even though the differences are significant for both the original and updated PTF, the spread of $npKGE_{upd}$ is smaller. The bottom whiskers in Figure 26 B are closer to 0 compared to Figure 17 B in all scoring categories besides category 7. There, two of the three outliers got within the bottom whisker, which caused it to extend from -0.05 to approximately -0.1.

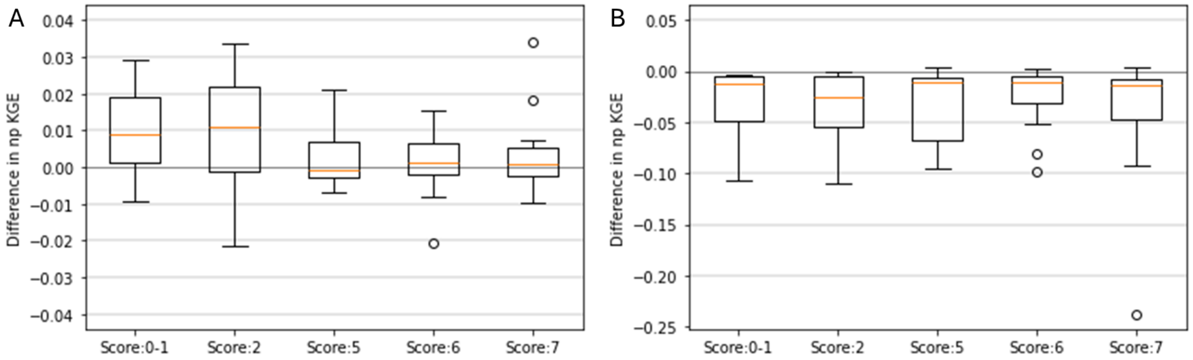


Figure 26: Boxplots of difference in np KGE between the updated PTF and (A) original PTF and (B) optimised, per scoring category

Table 7: P-values resulting from a Wilcoxon test for the difference in np KGE, bold numbers indicate a statistically significant difference

	0-1	2	5	6	7
Updated - original	**0.015**	0.068	0.98	0.52	0.41
Updated - optimised	**6.10e-5**	**1.47e-3**	**3.05e-4**	**4.58e-5**	**1.22e-4**
Original – optimised	**6.10e-5**	**6.10e-4**	**4.27e-4**	**7.63e-5**	**1.22e-4**

4.4.3. Overall applicability of updated PTF

When considering the performance of wflow_sbm it can be concluded that advances are made when using the updated PTF compared to the original PTF, although the increase in performance is only small: The updated PTF only results in a significant increase of $npKGE$ in the scoring category 0-1 with respect to the original PTF. Additionally, no significant decrease in wflow_sbm performance can be observed in the validation subbasins when applying the updated PTF. A more promising result can be seen in the difference in prediction of KsatHorFrac between the original and updated PTF. It was found that the difference in KsatHorFrac value between the original and updated PTF was significantly higher in the scoring categories 0-1, 2 and 5. No significant difference was observed in scoring category 6. A significant decrease was found in scoring category 7. These results are in line with the results from Figure 24, where the updated PTF predicted a higher KsatHorFrac in the GB training subbasins compared to the original PTF. The increase in predicted KsatHorFrac between the original and updated PTF is reflected by the R^2 and orientation of the trendline in Figure 27. Here it can be seen that the trendline of the updated PTF in Figure 27 B follows the 1:1 line more closely compared to the trendline from the original PTF in Figure 27 A. This indicates that the KsatHorFrac values predicted by the updated PTF are more similar across the complete range compared to the original PTF, where a clear deviation from the 1:1 line is observed for higher KsatHorFrac values.

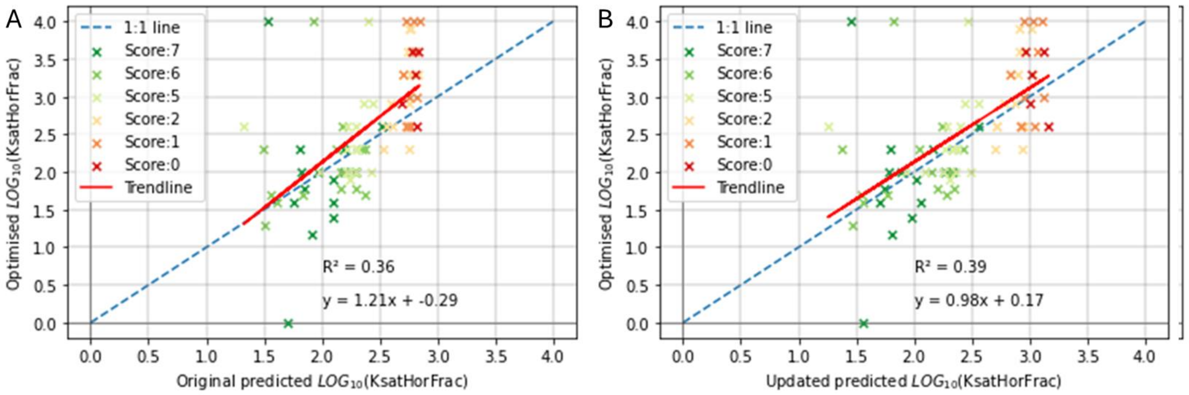


Figure 27: Scatter plot of predicted KsatHorFrac values by the (A) original PTF and (B) Updated PTF compared to the optimised KsatHorFrac values. The scoring categories are included, as well as a trendline with corresponding performance indicator R^2

4.4.4. Re-evaluating stand-out subbasins

Selzthal subbasin

From the results in the previous section it can be presumed that the updated PTF did not significantly impact the results of the KsatHorFrac and np KGE since the Selzthal subbasin has a score of 7. Slight changes can be seen in the zoomed in section of the hydrograph in Figure 28 B. The peaks are higher and the baseflow lower compared to the original PTF in Figure 19 B. This can be attributed to the decrease in KsatHorFrac. The statistics of the $npKGE$ are shown in Table 8. Here it can be seen that the np KGE is slightly lower compared to the original PTF. However, as stated in Table 7, this difference is not statistically significant.

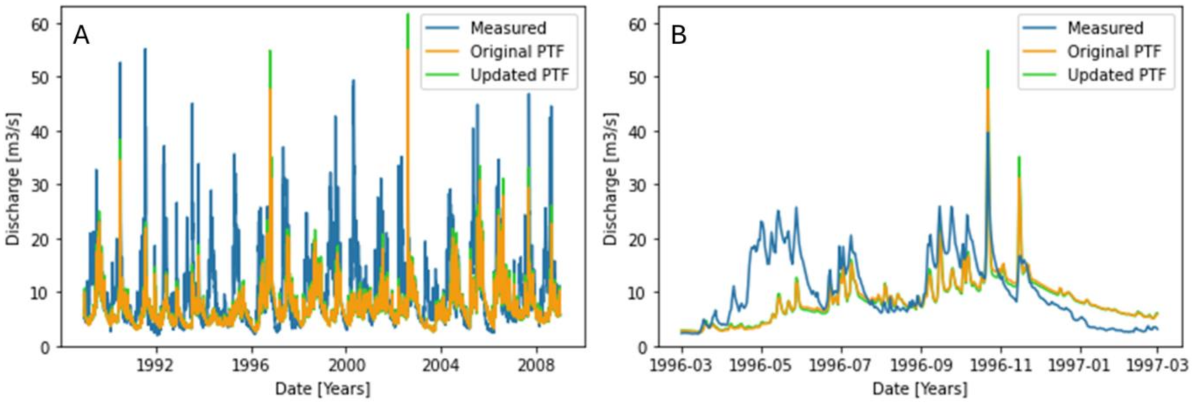


Figure 28: (A) Hydrograph of subbasin Selzthal in the Danube basin using the updated PTF and (B) magnified section between 1996-03 and 1997-03

Table 8: np KGE statistics (np KGE, r, alpha, beta) and KsatHorFrac values of the optimised scenario, original PTF and updated PTF for the Selzthal subbasin

	Np KGE	R	Alpha	Beta	KsatHorFrac
Optimised	0.63	0.77	0.97	0.71	10,000
Original PTF	0.40	0.49	0.91	0.70	34.17
Updated PTF	0.39	0.48	0.91	0.70	28.48

Blaise subbasin

The hydrograph with the updated PTF in Figure 29 A has changed slightly compared to the hydrograph with the original PTF (Figure 21 A). The peaks have slightly decreased and the baseflow has slightly increased. This in line with expectation as the updated KsatHorFrac has increased significantly w.r.t. the original KsatHorFrac. The increase of the baseflow has also increased the $npKGE$. Even though the increase is only small, it is significant according to the Wilcoxon test (Table 7)

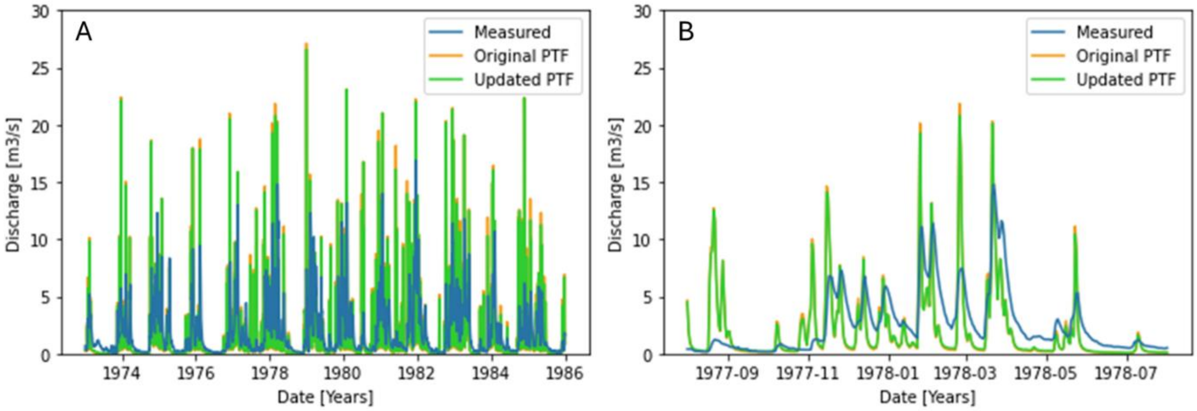


Figure 29: (A) Hydrograph of subbasin Blaise in the Seine basin using the updated PTF and (B) magnified section between 1977-08 and 1978-08

Table 9: np KGE statistics (np KGE, r, alpha, beta) and KsatHorFrac values of the optimised scenario, original PTF and updated PTF for the Blaise subbasin

	Np KGE	R	Alpha	Beta	KsatHorFrac
Optimised	0.82	0.87	0.94	0.89	10,000
Original PTF	0.69	0.81	0.81	0.85	539.22
Updated PTF	0.71	0.82	0.83	0.85	884.08

5. Discussion

The discussion consists of two main sections. First, the limitations of the data and research methods are stated in Section 5.1. Subsequently, results which require extra discussion are addressed in Section 5.2. Where possible, comparisons with existing literature is made.

5.1. Limitations

5.1.1. Data and models

SoilGrids v1.0

The SoilGrids variables played a significant role in this research as they were used to create the original and updated PTF. Additionally, they were used to determine the set of validation subbasins. The maps of the SoilGrids variables are predicted by combining soil samples and observations with multiple machine learning methods such as Random Forest, Gradient Boosting and Neural Networks, etc. (Hengl et al., 2017). Furthermore, the dataset only encompasses information about the first two meters of soil thickness at maximum. Even though the developers of the SoilGrids v1.0 database claim they reached the effective limit of software and remote sensing data, improvements in the dataset can still be made. The soil thickness limitation of two meters plays a large role in basins with deep groundwater dynamics such as the Seine. For several subbasins in the Seine the soil depth does not even reach two meters, such as in the discarded subbasin Blennes (Figure 30). The simulated discharge in this subbasin overestimated the observed discharge. A reason for this could be that in reality the water infiltrates beyond the limit of the soil thickness, while in the model, this water ends up in the river.

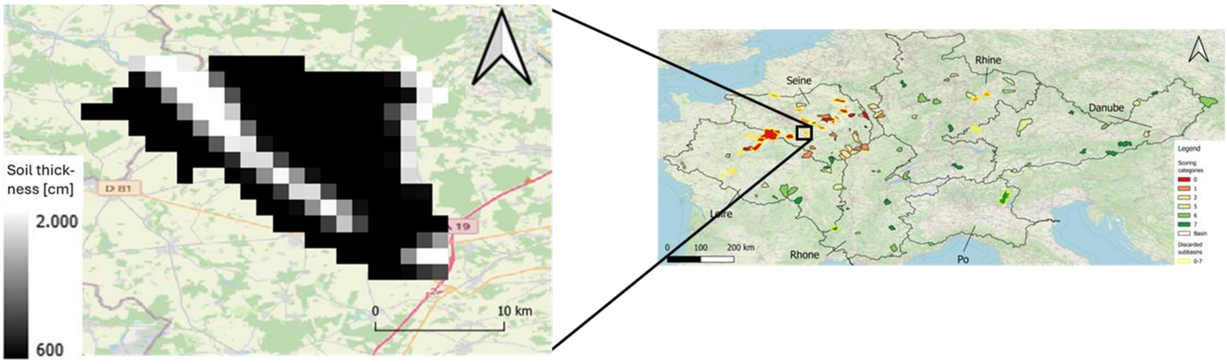


Figure 30: Soil thickness of the discarded subbasin Blennes in the Seine basin

Discharge data

Even though GRDC cooperates with the World Meteorological Organization (WMO) and claims the global discharge timeseries is quality controlled (GRDC, 2024), there can be uncertainties in the measurements. This uncertainty can affect the optimized KsatHorFrac value since the sensitivity analysis is related to the model performance of wflow_sbm in terms of $npKGE$. Hence, the hydrographs should manually be checked and anomalies such as showed in Figure 31 should be addressed accordingly. Here the optimised KsatHorFrac was 6,000 according to the sensitivity analysis. However, since the complete period was considered this value is not representative as it appears an intervention was applied around 1974 which reduced the measured discharge.

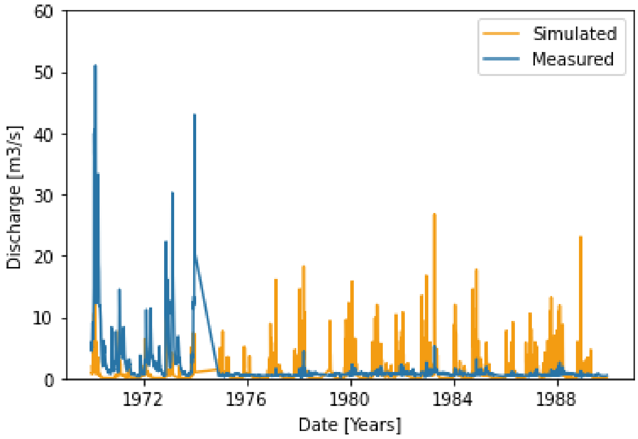


Figure 31: Measured and observed discharge in the discarded subbasin Le Moulin de L'Etang in the Seine basin with optimised KsatHorFrac = 6000

5.1.2. Methods

Basin selection

Early in the research it was unknown how many potential subbasins would be discarded after the selection. Since the selection was a time-consuming step in the research, it was decided to focus on the already available wflow_sbm models (Rhine, Po, Loire) and their neighbouring basins. The drawback of this decision is that the number of potential validation subbasins is largely reduced compared to an even larger study area. Especially the subbasins with a low score, when comparing the SoilGrids variables to the GB training subbasins. To overcome this limitation more basins could be included in the longlist. However, it is important that these basins have sufficient and reliable discharge and forcing data such that less subbasins are to be removed throughout the study. Another important factor which plays a role in extending the study area is the build-up of the soil. During the investigation of the dominant soil types per basin in Section 4.1.1 it became clear that the subbasins in the Seine basin have are different compared to the other subbasins, which translated in generally low scores in that basin. Subbasins with a low score were the bottleneck throughout the study. Hence, in order to mitigate this limitation it would be wise to select subbasins with similar dominant soil types as can be found in the Seine basin.

When determining the scoring categories of the subbasins, different ranges of the SoilGrids parameters were analysed. Eventually it was chosen to only consider the 90th percentile of the data. Because of this, there is a possibility that too much representative data is discarded, meaning that the subbasins with a low score (0-2) were sufficiently included into the RF algorithm. If this is the case, the differences between the original PTF and results from the sensitivity analysis in terms of both KsatHorFrac predictions and $npKGE$ are conservative for those scoring categories.

Model building

Al subbasins were handpicked from the six basins considered in the research, the river system in these basins are based on a river geometry dataset (Lin et al., 2020). During the building phase of the models it was discovered that not all subbasins could be build due to a limitation in coverage of that river geometry dataset, which only covers rivers which are wider than 30 meters. Which is a problem as this research only considers headwaters, with most often a low stream order and thus smaller rivers. Hence, not all discharge gauges are located in a river within the dataset. In order to bypass this problem, an alternative method was used to generate the river geometry. This method makes an estimate of the geometry based on a power law and the discharge.

A drawback of applying this method is that the river geometry is slightly different compared to the network from the rivers_lin2019 database. This change can clearly be seen in Figure 32 A and B. Here the blue rivers are from the rivers_lin2019 database and the orange rivers are created using the alternative method. Because of this, the geometry and location of the subbasins also changes. For most subbasins this change is minimal (Figure 32 A). However, some subbasins change as the GRDC gauges snap to an adjacent river. As a result, some subbasins did not correspond with the upstream area attributed to the GRDC gauges (Figure 32 B). As the river_lin2019 dataset is only limited to a number of subbasins, it would be preferred to create the Rhine, Seine, Loire, Rhone, Po and Danube basins also with the alternative method instead. Because of this, when generating the subbasins separately, they would be within the 15% margin when selecting them from the basins they are in.

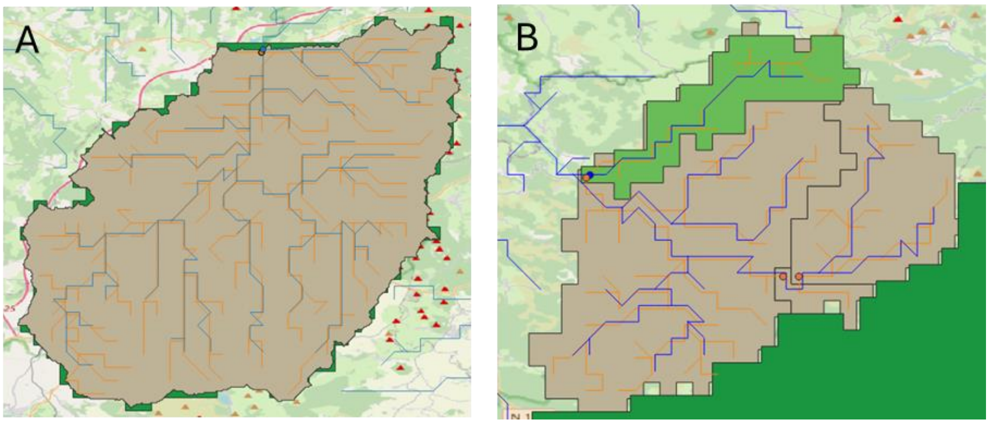


Figure 32: Example of (A) sufficiently overlapping models from Pontgibaud subbasins in Loire basin and (B) insufficiently overlapping model form subbasin Pont de la Borie in the Loire basin. Here green is the subbasin selected from the Seine basin (using river_lin2019) and grey is subbasin generated as separately (using alternative method)

Sensitivity analysis

In order to determine the optimized KsatHorFrac, the non-parametric Kling Gupta Efficiency (Pool et al., 2018) was chosen as performance indicator. It was found that the $npKGE$ was less sensitive to changes in KsatHorFrac compared to the KGE (H. V. Gupta et al., 2009). This means that the maximum difference between the lowest and highest values is smaller for $npKGE$ compared to KGE. There could be a possibility that the sensitivity analysis would result into a different optimized KsatHorFrac if the KGE was used, instead of the $npKGE$. The difference for the Souspierre subbasin, located in the Rhone basin can be seen in Figure 33. In the example below, the optimised KsatHorFrac reduces from 10,000 to 4,000 when using KGE. Since the predicted KsatHorFrac is generally underestimated compared to the optimised value for lower subbasins, using the KGE instead of the $npKGE$ could result in a smaller difference.

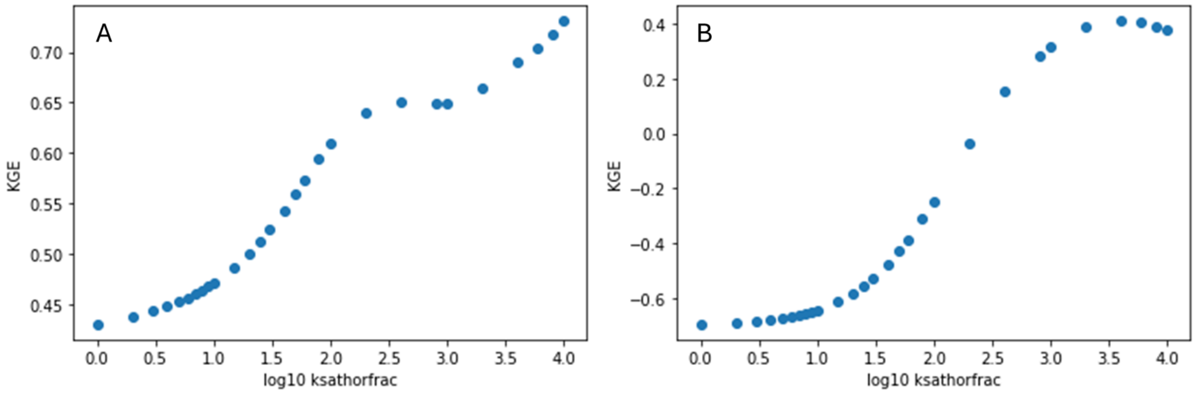


Figure 33: Optimisation curve when (A) np KGE and (B) KGE is the performance indicator for the Souspierre subbasin in the Rhone basin

Randomness of the Random Forest algorithm

Due to the random nature of the RF algorithm, the relative importance of the predictors changed after each run of the algorithm. Because of this, it was decided to run the algorithm ten times and take the average of the resulting KsatHorFrac maps. Figure 34 shows the spread of the relative importance for the predictors of the updated PTF after 100 runs (blue) together with the ten run average values which were used during the study (orange). From the figure it can be concluded that the spread of the predictors can be large, especially for the bulk density (BLDFIE). However, the median values of the boxplots correspond with the ten run average value for most of the predictors. Only fraction of silt (SLTPPT), elevation and slope are slightly overrepresented. Since no predictors show large differences between the ten run average and the mean value of the 100 runs, the choice to use a ten run average is substantiated.

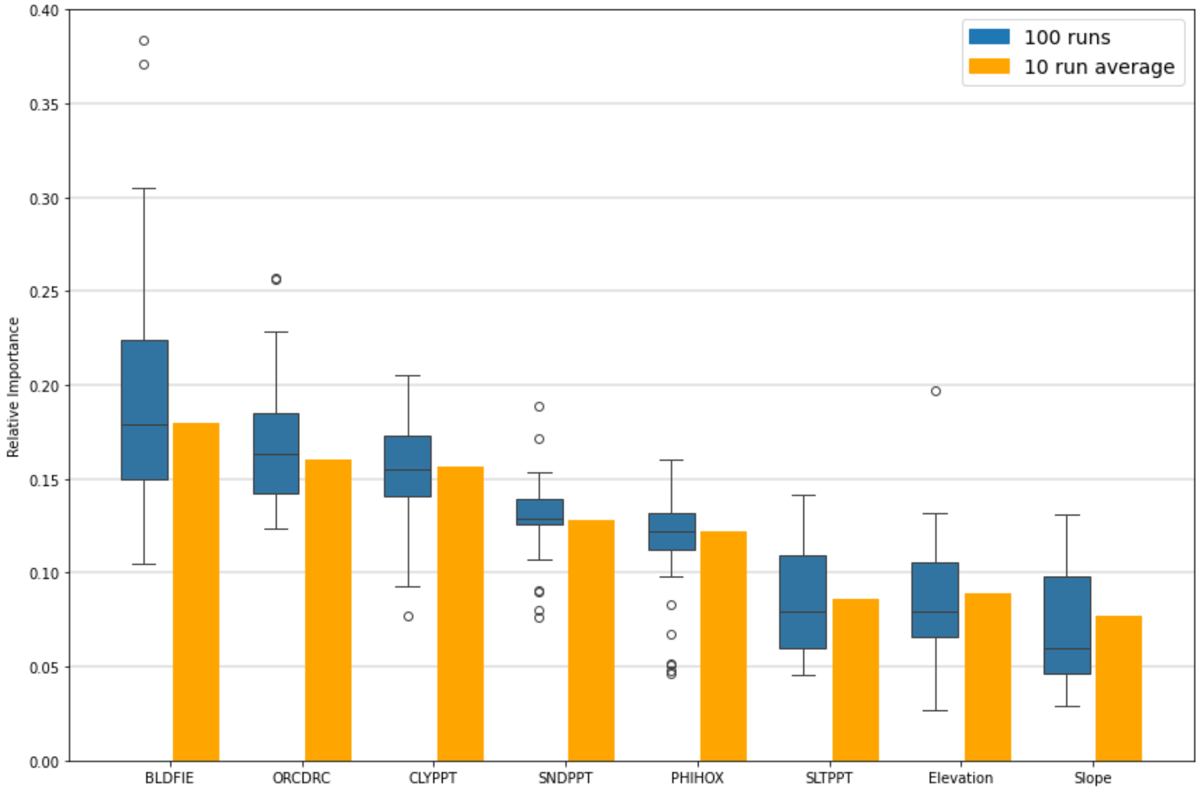


Figure 34: Spread of relative importance for the predictors of the updated PTF after conducting 100 runs (boxplot) and the 10 run average value (bar graph).

5.2. Interpretation of results

5.2.1. Removal of subbasins

In Research Question 2, 26 subbasins were removed from the validation set as a result of incorrect discharge simulations. The majority of these subbasins were removed because of a large overestimation of the observed discharge (Figure 35). Especially the scoring categories 0 and 1 were affected by this. The majority of these subbasins are located in the Seine basin (Figure 12). A possible explanation is that the Seine basin has a dominant interaction between the river and groundwater. Most of the basin lies within the Paris river basin, which is the largest ground water reservoir of Europe (Flipo et al., 2023). According to Flipo et al. (2020) more than 50% of the effective rainfall is infiltrated towards deeper groundwater. During the simulation in wflow_sbm no leakage towards deeper groundwater was considered. Hence, explaining the large overestimation. Additionally, the Seine basin is exposed to a large anthropogenic pressure on the groundwater as the area includes industries and agriculture. Flipo et al. (2020) also state that the functioning of the interaction between the surface and subsurface water is significantly affected by the groundwater withdrawal. Pryet et al. (2015) states that the groundwater supplies 82% of the main stream network. Additionally, the groundwater withdrawals reduce the groundwater to stream flow by 19% and even causes a stream to groundwater flow near pumping facilities. These processes are not simulated in wflow_sbm. This explains the relatively poor performance in the Seine basin independent from the KsatHorFrac values.

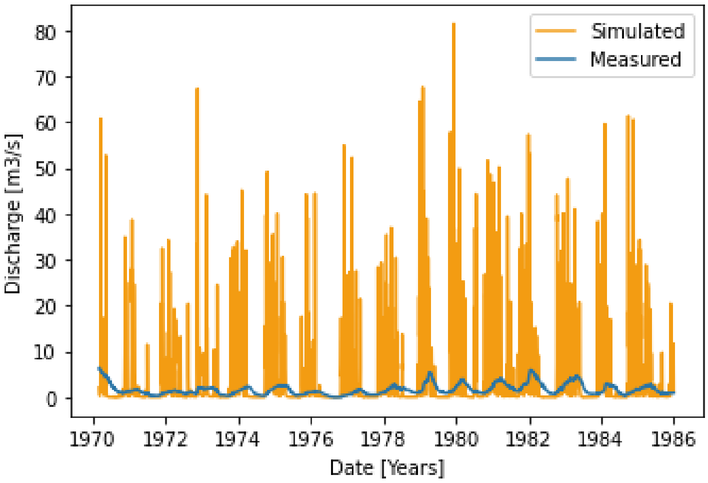


Figure 35: hydrograph of discarded subbasin Poilcourt Sydney in the Seine basin with a score of 0

5.2.2. Relation between KsatHorFrac and model performance

When looking at Figure 16, it can be said that overall, the $npKGE_{org}$ appears to be between the default and optimised np KGE, which suggests the algorithm seems to function as expected. In 26 subbasins, the $npKGE_{def}$ (uniformly = 100) outperforms $npKGE_{org}$. However, in 14 of these cases the difference is less than 0.01. Additionally, in 14 subbasins, the default KsatHorFrac is equal to the optimised KsatHorFrac resulting from the sensitivity analysis. In those subbasins the $npKGE_{def}$ outperforms that of the predicted KsatHorFrac, as the $npKGE_{opt}$ always outperforms that of the predicted KsatHorFrac.

Furthermore, it can be seen that in the subbasins where the optimised KsatHorFrac is much higher compared to the predicted KsatHorFrac (Figure 13 C), the model performance follows the same observation (Figure 16 F). However, this mainly occurs where the optimised KsatHorFrac value approaches the upper regime of the range (2,000-10,000). Figure 14 A shows that the predicted KsatHorFrac of the original PTF does not exceed 699 ($10^{2.84}$), This is in line with the research of Ali et al. (2023), where the predicted KsatHorFrac in the test subbasins is lower compared to the optimised KsatHorFrac and appears to be limited to 1000 (10^3). As shown in Figure 33, the optimised KsatHorFrac reduces when KGE is used instead of $npKGE$. Therefore, it is uncertain how the KGE resulting from the optimised and predicted KsatHorFrac will compare.

5.2.3. Relative importance of the predictors

As can be seen in Figure 23, the order of importance for the predictors has changed between the original and updated PTF. Percentage of clay and pH index got less important while organic carbon content and percentage of sand got more important. Both the topographic predictors: elevation and slope are amongst the least important predictors. Even though KsatHorFrac and Ksat are not the same, similar results have been found by Gupta et al. (2021) who studies the latter. The researchers found that over all soil predictors were more important compared to topographic predictors. However, Gupta et al. (2021) did not consider the exact same predictors. Additionally, they found that elevation contributed more compared to slope. Which has also been found in this research.

6. Conclusion and recommendations

In this study, the transferability of data-driven PTFs for KsatHorFrac were evaluated. The aim of the research consisted of two parts: First, the aim is to get more insight into the transferability of the original PTF of Ali et al. (2023), Subsequently, the transferability of an updated PTF which includes topographic predictors is evaluated. The transferability was evaluated by looking into KsatHorFrac prediction and resulting model performance of wflow_sbm in subbasin throughout central and Western Europa. In this chapter, the conclusions for each research question are stated. Subsequently, it is concluded how these questions relate to the knowledge gap and research aim.

6.1. Conclusion

6.1.1. Subbasin selection

'Which catchments should be considered for validating the original and updated PTF?'

The study area included the following basins in central and Western Europa: Seine, Loire, Rhone, Po, Rhine and upper section of the Danube. These basins were selected for the research since it was known a reliable and sufficient amount of discharge time series were available through GRDC (GRDC, 2024). After selecting only the headwaters in the basins, each subbasin received a score depending on how the SoilGrids variables compared with the GB training subbasins. After analysing the hydrographs in RQ2 and removing subbasins with unrealistic simulation results, the set of validation subbasins consisted of 76 subbasins divided into five scoring categories: 0-1, 2, 5, 6 and 7.

6.1.2. Evaluating transferability of the original PTF

'How does the KsatHorFrac prediction of the original PTF and resulting model performance of wflow_sbm change in comparison to the optimised scenario, when the original PTF is implemented for subbasins for which it was not developed?'

When evaluating the KsatHorFrac prediction, it was found that there was no statistically significant difference between the optimised and predicted KsatHorFrac for subbasins with a high score (5-7). However, there was a statistically significant difference for subbasins with a low score (0-2). This was as expected since subbasins with a high score share the same soil characteristic with the GB training subbasins, whereas subbasins with a low score do not. However, this observation could not be seen in the difference in model performance between the optimised simulation and when applying the original PTF. There, it was found that the $npKGE_{org}$ was significantly lower compared to the $npKGE_{opt}$, for all scoring categories.

6.1.3. Creating an updated PTF

'How does the KsatHorFrac prediction of the PTFs change in the GB training subbasins when topographic predictors are implemented in the RF training phase?'

The updated PTF showed a negligible improvement in terms of performance indicators RMSE and R^2 with respect to the original PTF. The first iteration (including elevation, slope and drainage density) and second iteration (including elevation and slope, but excluding drainage density and bedrock depth) also matched in terms of performance indicators. However, since two predictors (bedrock depth and drainage density) had a low contribution towards the prediction, it was decided to continue with the second iteration. Even though, the change in performance between the original and updated PTF was negligible, the overall increase of predicted KsatHorFrac was clearly visible in the GB training subbasins. On average, the updated PTF predicted 1.27 times higher KsatHorFrac values compared to the original PTF.

6.1.4. Comparing the original and updated PTF

'How does the KsatHorFrac prediction of the updated PTF and resulting performance of wflow_sbm change, with respect to the original PTF, when the updated PTF is implemented for subbasins for which it was not developed?'

The overall increase in predicted KsatHorFrac values in the GB training subbasins was translated to the validation subbasins. Significantly higher KsatHorFrac values were predicted in three out of the five (0-1, 2 and 5) scoring categories compared to the original PTF. As a result, no significant differences between the predicted and optimised KsatHorFrac values could be observed anymore. When comparing the $npKGE_{upd}$ and $npKGE_{org}$, it was found that the updated PTF significantly increased the model performance for subbasins with a score of 0-1. The change in $npKGE$ was not statistically significant in the remaining scoring categories. Overall, the model performance resulting from both the original and updated PTF were significantly lower compared to the optimised simulation. However, the KsatHorFrac prediction in the validation subbasins was better when comparing the updated PTF to the original PTF, suggesting that including topographic predictors in the RF algorithm improves the predictive capability.

6.2. Recommendations

6.2.1. Practical recommendations for hydrological modelling using data-driven PTFs for KsatHorFrac

Both results of RQ 2 and RQ 4 show that the $npKGE_{opt}$ outperforms the $npKGE_{org}$ and $npKGE_{upd}$ in almost every situation. Therefore it would be recommended to conduct a sensitivity analysis in order to find the optimised KsatHorFrac when time allows. This would be suitable for small subbasins which have short computational times. However, for large subbasins or on basin scale, it is not advised to conduct the sensitivity analysis to find the optimal KsatHorFrac value as this is too time consuming. Even though the default KsatHorFrac value of 100 resulted in similar model performance compared to the $npKGE$ resulting from both PTFs and $npKGE_{opt}$, it is advised not to conduct simulation with the default KsatHorFrac without orientation on beforehand.

When it is chosen to use a data-driven PTF instead of creating an optimised KsatHorFrac value it is recommended to use the updated PTF as presented in this study. The reason for this is that the updated PTF results in significantly a higher $npKGE$ compared to the original PTF for subbasins with a low score, without sacrificing model performance for subbasins with a high score. This makes the updated PTF more widely applicable.

6.2.2. Recommendations for future research

If a similar type of research were to be conducted, it would be advisable to evaluate both the $npKGE$ as well as the KGE. As highlighted in the discussion, the result of the sensitivity analysis is dependent on the type of performance indicator. In the discussion it could also be seen that the KGE is more sensitive to the KsatHorFrac. This suggests that the differences in KGE resulting from the predicted and optimized KsatHorFrac could different compared to the $npKGE$.

Even though this research provided relevant information about the transferability of RF based PTFs for the parameter KsatHorFrac, further improving the performance of the PTF will improve the model performance of wflow_sbm. First of all, the area in which the PTF is trained and tested could be extended to not only include GB. This would create a more diverse training set for the RF algorithm and hopefully create a wider range of KsatHorFrac predictions. This extension of the training set could for example include subbasins with different soil or topographic characteristics compared to GB, bypassing the limitation that PTFs are only applicable in a small representative area (Abdelbaki, 2021).

Finally, further investigation into the type and amount of predictors for the KsatHorFrac PTF could be conducted. The inclusion of topographic predictors was inspired by a PTF for Ksat developed by Gupta et al. (2021). Since the study has shown that including topographic predictors in the RF training phase has a positive effect on the prediction, experimentation with other predictors which are related to Ksat could be done. Both Gupta et al. (2021) and Ayele et al. (2020) included environmental and/or climate related predictors in the PTF training. Subsequently, different sets of the predictors can be tested. Leij et al. (2004) found that the PTF performance (RMSE) changes when different combinations of the predictors are made.

7. References

Abdelbaki, A. M. (2021). Selecting the most suitable pedotransfer functions for estimating saturated hydraulic conductivity according to the available soil inputs. *Ain Shams Engineering Journal*, *12*(3), 2603–2615. https://doi.org/10.1016/j.asej.2021.01.030

Águila, J. F., McDonnell, M. C., Flynn, R., Butler, A. P., Hamill, G. A., Etsias, G., Benner, E. M., & Donohue, S. (2023). Comparison of saturated hydraulic conductivity estimated by empirical, hydraulic and numerical modeling methods at different scales in a coastal sand aquifer in Northern Ireland. *Environmental Earth Sciences*, *82*(13). https://doi.org/10.1007/s12665-023-11019-6

Ali, A. M., Imhoff, R. O., & Weerts, A. H. (2023). *Machine learning for predicting spatially variable lateral hydraulic conductivity: a step towards efficient hydrological model calibration and global applicability*. http://www.hydro.eaufrance.fr/

Ameli, A. A., McDonnell, J. J., & Bishop, K. (2016). The exponential decline in saturated hydraulic conductivity with depth: a novel method for exploring its effect on water flow paths and transit time distribution. *Hydrological Processes*, *30*(14), 2438–2450. https://doi.org/10.1002/hyp.10777

Ayele, G. T., Demissie, S. S., Jemberrie, M. A., Jeong, J., & Hamilton, D. P. (2020). Terrain effects on the spatial variability of soil physical and chemical properties. *Soil Systems*, *4*(1), 1–21. https://doi.org/10.3390/soilsystems4010001

Bell, V. A., Kay, A. L., Jones, R. G., & Moore, R. J. (2007). Development of a high resolution grid-based river flow model for use with regional climate model output. In *Hydrol. Earth Syst. Sci* (Vol. 11, Issue 1). www.hydrol-earth-syst-sci.net/11/532/2007

Benning, R. G., Waterhuishouding, V., Kanaal, N., & Wageningen, P. A. (1995). *Towards a new lumped parameterization at catchment scale*.

Blakely, L., Reno, M. J., & Broderick, R. (2018). *Evaluation and Comparison of Machine Learning Techniques for Rapid QSTS Simulations*. https://www.researchgate.net/publication/326560291

Blyth, E. M., Arora, V. K., Clark, D. B., Dadson, S. J., De Kauwe, M. G., Lawrence, D. M., Melton, J. R., Pongratz, J., Turton, R. H., Yoshimura, K., & Yuan, H. (2021). ADVANCES AND FUTURE DIRECTIONS IN EARTH SYSTEM MODELLING (I SIMPSON, SECTION EDITOR) Advances in Land Surface Modelling. *Current Climate Change Reports*. https://doi.org/10.1007/s40641-021-00171-5/Published

Brakensiek, D. L., Rawls, W. J., & Stephenson, G. R. (1984). Modifying SCS hydrologic soil groups and curve numbers for rangeland soils. *American Society of Agricultural Engineers*.

Breiman, L. (2001). *Random Forests* (Vol. 45).

Chou, H. K., Ochoa-Tocachi, B. F., Moulds, S., & Buytaert, W. (2022). Parameterizing the JULES land surface model for different land covers in the tropical Andes. *Hydrological Sciences Journal*, *67*(10), 1516–1526. https://doi.org/10.1080/02626667.2022.2094709

Collins, D. B. G., & Bras, R. L. (2010). Climatic and ecological controls of equilibrium drainage density, relief, and channel concavity in dry lands. *Water Resources Research*, *46*(4). https://doi.org/10.1029/2009WR008615

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V., & Böhner, J. (2015). System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. *Geoscientific Model Development*, *8*(7), 1991–2007. https://doi.org/10.5194/gmd-8-1991-2015

De Bruin, H. A. R., Trigo, I. F., Bosveld, F. C., & Meirink, J. F. (2016). Thermodynamically based model for actual evapotranspiration of an extensive grass field close to FAO reference, suitable for remote sensing application. *Journal of Hydrometeorology*, *17*(5), 1373–1382. https://doi.org/10.1175/JHM-D-15-0006.1

Devia, G. K., Ganasri, B. P., & Dwarakish, G. S. (2015). A Review on Hydrological Models. *Aquatic Procedia*, *4*, 1001–1007. https://doi.org/10.1016/j.aqpro.2015.02.126

Eilander, D., Boisgontier, H., Bouaziz, L. J. E., Buitink, J., Couasnon, A., Dalmijn, B., Hegnauer, M., de Jong, T., Loos, S., Marth, I., & van Verseveld, W. (2023). HydroMT: Automated and reproducible model building and analysis. *Journal of Open Source Software*, *8*(83), 4897. https://doi.org/10.21105/joss.04897

European Environment Agency. (2011). *European water archive — European Environment Agency*. https://www.eea.europa.eu/data-and-maps/data/external/european-water-archive

Flipo, N., Gallois, N., Labarthe, B., Baratelli, F., Viennot, P., Schuite, J., Rivière, A., Bonnet, R., & Boé, J. (2020). Pluri-annual water budget on the Seine basin : past, current and future trends. *Handbook of Environmental Chemistry*, *90*, 59–89. https://doi.org/10.1007/698_2019_392

Flipo, N., Gallois, N., & Schuite, J. (2023). Regional coupled surface-subsurface hydrological model fitting based on a spatially distributed minimalist reduction of frequency domain discharge data. *Geoscientific Model Development*, *16*(1), 353–381. https://doi.org/10.5194/gmd-16-353-2023

GRDC. (2024). *GRDC Data Portal*. https://portal.grdc.bafg.de/applications/public.html?publicuser=PublicUser#dataDownload/Subregions

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*(4). https://doi.org/10.1029/2020MS002242

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Hateffard, F., Steinbuch, L., & Heuvelink, G. B. M. (2024). Evaluating the extrapolation potential of random forest digital soil mapping. *Geoderma*, *441*. https://doi.org/10.1016/j.geoderma.2023.116740

Hengl, T., De Jesus, J. M., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS ONE*, *12*(2). https://doi.org/10.1371/journal.pone.0169748

Horton, R. E. (1932). Drainage-basin characteristics. *Eos, Transactions American Geophysical Union*, *13*(1), 350–361. https://doi.org/10.1029/TR013I001P00350

Imhoff, R. O., van Verseveld, W. J., van Osnabrugge, B., & Weerts, A. H. (2020). Scaling Point-Scale (Pedo)transfer Functions to Seamless Large-Domain Parameter Estimates for High-Resolution Distributed Hydrologic Modeling: An Example for the Rhine River. *Water Resources Research*, *56*(4). https://doi.org/10.1029/2019WR026807

Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, *47*(260), 583–621. https://doi.org/10.1080/01621459.1952.10483441

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). https://doi.org/10.1029/2002WR001641

Lin, P., Pan, M., Allen, G. H., de Frasson, R. P., Zeng, Z., Yamazaki, D., & Wood, E. F. (2020). Global Estimates of Reach-Level Bankfull River Width Leveraging Big Data Geospatial Analysis. *Geophysical Research Letters*, *47*(7). https://doi.org/10.1029/2019GL086405

Pool, S., Vis, M., & Seibert, J. (2018). Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency. *Hydrological Sciences Journal*, *63*(13–14), 1941–1953. https://doi.org/10.1080/02626667.2018.1552002

Pryet, A., Labarthe, B., Saleh, F., Akopian, M., & Flipo, N. (2015). Reporting of Stream-Aquifer Flow Distribution at the Regional Scale with a Distributed Process-Based Model. *Water Resources Management*, *29*(1), 139–159. https://doi.org/10.1007/s11269-014-0832-7

Rogelis, M. C., Werner, M., Obregón, N., Wright, N., & Rogelis, M. C. (2016). *Hydrological model assessment for flood early warning in a tropical high mountain basin*. https://doi.org/10.5194/hess-2016-30

Samaniego, L., Kumar, R., & Attinger, S. (2010). Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resources Research*, *46*(5). https://doi.org/10.1029/2008WR007327

Scanlon, B. R., Zhang, Z., Save, H., Sun, A. Y., Schmied, H. M., Van Beek, L. P. H., Wiese, D. N., Wada, Y., Long, D., Reedy, R. C., Longuevergne, L., Döll, P., & Bierkens, M. F. P. (2018). Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(6), E1080–E1089. https://doi.org/10.1073/pnas.1704665115

Schellekens, J., Dutra, E., Martínez-De La Torre, A., Balsamo, G., Van Dijk, A., Sperna Weiland, F., Minvielle, M., Calvet, J. C., Decharme, B., Eisner, S., Fink, G., Flörke, M., Peßenteiner, S., Van Beek, R., Polcher, J., Beck, H., Orth, R., Calton, B., Burke, S., … Weedon, G. P. (2017). A global water resources ensemble of hydrological models: The eartH2Observe Tier-1 dataset. *Earth System Science Data*, *9*(2), 389–413. https://doi.org/10.5194/essd-9-389-2017

Singh, V. P. (2018). Hydrologic modeling: progress and future directions. In *Geoscience Letters* (Vol. 5, Issue 1). SpringerOpen. https://doi.org/10.1186/s40562-018-0113-z

Tanaka, T., & Tachikawa, Y. (2015). Test de l’applicabilité d’un modèle hydrologique distribué basé sur l’onde cinématique pour deux bassins climatologiquement contrastés. *Hydrological Sciences Journal*, *60*(7–8), 1361–1373. https://doi.org/10.1080/02626667.2014.967693

Van Looy, K., Bouma, J., Herbst, M., Koestel, J., Minasny, B., Mishra, U., Montzka, C., Nemes, A., Pachepsky, Y. A., Padarian, J., Schaap, M. G., Tóth, B., Verhoef, A., Vanderborght, J., van der Ploeg, M. J., Weihermüller, L., Zacharias, S., Zhang, Y., & Vereecken, H. (2017). Pedotransfer Functions in Earth System Science: Challenges and Perspectives. In *Reviews of Geophysics* (Vol. 55, Issue 4, pp. 1199–1256). Blackwell Publishing Ltd. https://doi.org/10.1002/2017RG000581

Van Verseveld, W. J., Weerts, A. H., Visser, M., Buitink, J., Imhoff, R. O., Boisgontier, H., Bouaziz, L., Eilander, D., Hegnauer, M., Ten Velden, C., & Russell, B. (2024). Wflow-sbm v0.7.3, a spatially distributed hydrological model: From global data to local applications. *Geoscientific Model Development*, *17*(8), 3199–3234. https://doi.org/10.5194/gmd-17-3199-2024

Vertessy, R. A., & Elsenbeer, H. (1999). Distributed modeling of storm flow generation in an Amazonian rain forest catchment: Effects of model parameterization. *Water Resources Research*, *35*(7), 2173–2187. https://doi.org/10.1029/1999WR900051

Wannasin, C., Brauer, C. C., Uijlenhoet, R., van Verseveld, W. J., & Weerts, A. H. (2021). Daily flow simulation in Thailand Part I: Testing a distributed hydrological model with seamless parameter maps based on global data. *Journal of Hydrology: Regional Studies*, *34*. https://doi.org/10.1016/j.ejrh.2021.100794

Weerts, A. (2024). *Dataset underlying the publication: Revealing spatial patterns of lateral hydraulic conductivity through sensitivity analysis*. Wageningen University and Research.

WMO. (2023). *State of Global Water Resources 2022*.

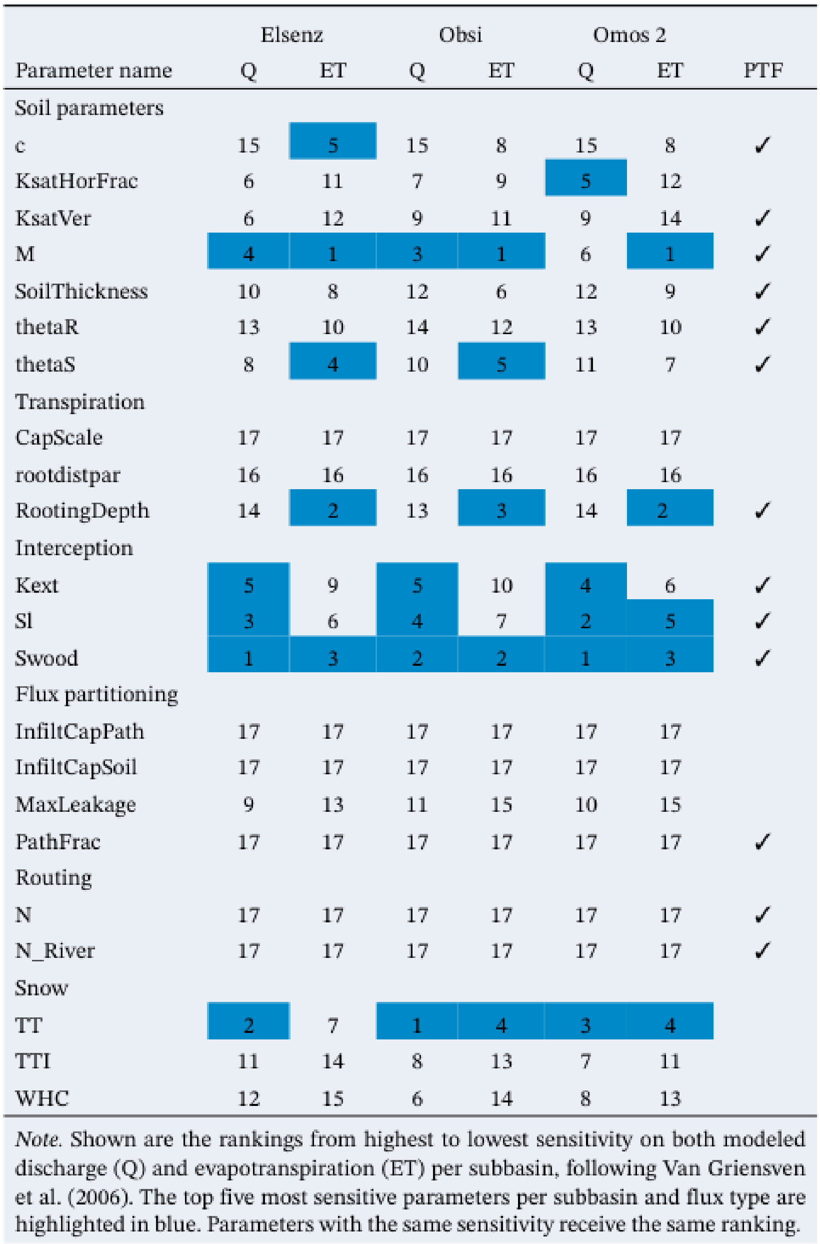
Zuo, Y., & He, K. (2021). Evaluation and development of pedo-transfer functions for predicting soil saturated hydraulic conductivity in the alpine frigid hilly region of qinghai province. *Agronomy*, *11*(8). https://doi.org/10.3390/agronomy11081581

Appendices

Appendix A: Sensitive parameters wflow_sbm

Table 10 shows the results for a sensitivity analysis of wflow_sbm for three sub-catchments in the Rhine basin (Imhoff et al., 2020). The sub-catchments were selected based on soil type, location and catchment size. It was made sure these three variables were different amongst the sub-catchments in order to mimic the variety in the Rhine basin.

Table 10: Results of the parameter sensitivity analysis of wflow_sbm for three sub-catchments the Rhine basin (Imhoff et al., 2020)



	Elsenz		Obsi		Omos 2		
Parameter name	Q	ET	Q	ET	Q	ET	PTF
Soil parameters							
c	15	5	15	8	15	8	✓
KsatHorFrac	6	11	7	9	5	12	
KsatVer	6	12	9	11	9	14	✓
M	4	1	3	1	6	1	✓
SoilThickness	10	8	12	6	12	9	✓
thetaR	13	10	14	12	13	10	✓
thetaS	8	4	10	5	11	7	✓
Transpiration							
CapScale	17	17	17	17	17	17	
rootdistpar	16	16	16	16	16	16	
RootingDepth	14	2	13	3	14	2	✓
Interception							
Kext	5	9	5	10	4	6	✓
Sl	3	6	4	7	2	5	✓
Swood	1	3	2	2	1	3	✓
Flux partitioning							
InfiltCapPath	17	17	17	17	17	17	
InfiltCapSoil	17	17	17	17	17	17	
MaxLeakage	9	13	11	15	10	15	
PathFrac	17	17	17	17	17	17	✓
Routing							
N	17	17	17	17	17	17	✓
N_River	17	17	17	17	17	17	✓
Snow							
TT	2	7	1	4	3	4	
TTI	11	14	8	13	7	11	
WHC	12	15	6	14	8	13	

Note. Shown are the rankings from highest to lowest sensitivity on both modeled discharge (Q) and evapotranspiration (ET) per subbasin, following Van Griensven et al. (2006). The top five most sensitive parameters per subbasin and flux type are highlighted in blue. Parameters with the same sensitivity receive the same ranking.

Appendix B: Basin selection

B.1. SoilGrids ranges

The ranges corresponding to the boxplots in Figure 10 are stated in Table 11. The upper and lower bounds per range are shown for each SoilGrids variable in the GB training subbasins.

Table 11: Four ranges of the seven SoilGrids variable of the GB training subbasins

Variable	BDTICM [cm]	BLDFIE [kg m^{-3}]	CLYPPT [kg kg^{-1}]	ORCDRC [g kg^{-1}]	PHIHOX [-]	SLTPPT [kg kg^{-1}]	SNDPPT [kg kg^{-1}]
Complete dataset							
Max	2226.65	1504.47	31.97	162.72	72.91	46.27	66.80
Min	550.77	1110.41	7.45	9.95	50.15	23.66	30.06
Box plot whiskers							
Upper	1680.29	-	-	123.26	69.66	-	-
Lower	770.79	-	-	-	-	-	-
Middle 95 %							
97.5th	1939.65	1486.08	29.56	118.34	70.08	42.25	65.00
2.5th	916.76	1159.54	10.09	12.12	51.84	24.86	32.55
Middle 90 %							
95th	1801.00	1476.77	28.31	100.49	68.39	41.32	63.38
5th	947.46	1191.42	10.77	12.64	52.25	26.21	33.07

Appendix C: Updated PTF

C.1. Randomness of the RF algorithm containing seven SoilGrids variables in addition to elevation, slope and drainage density

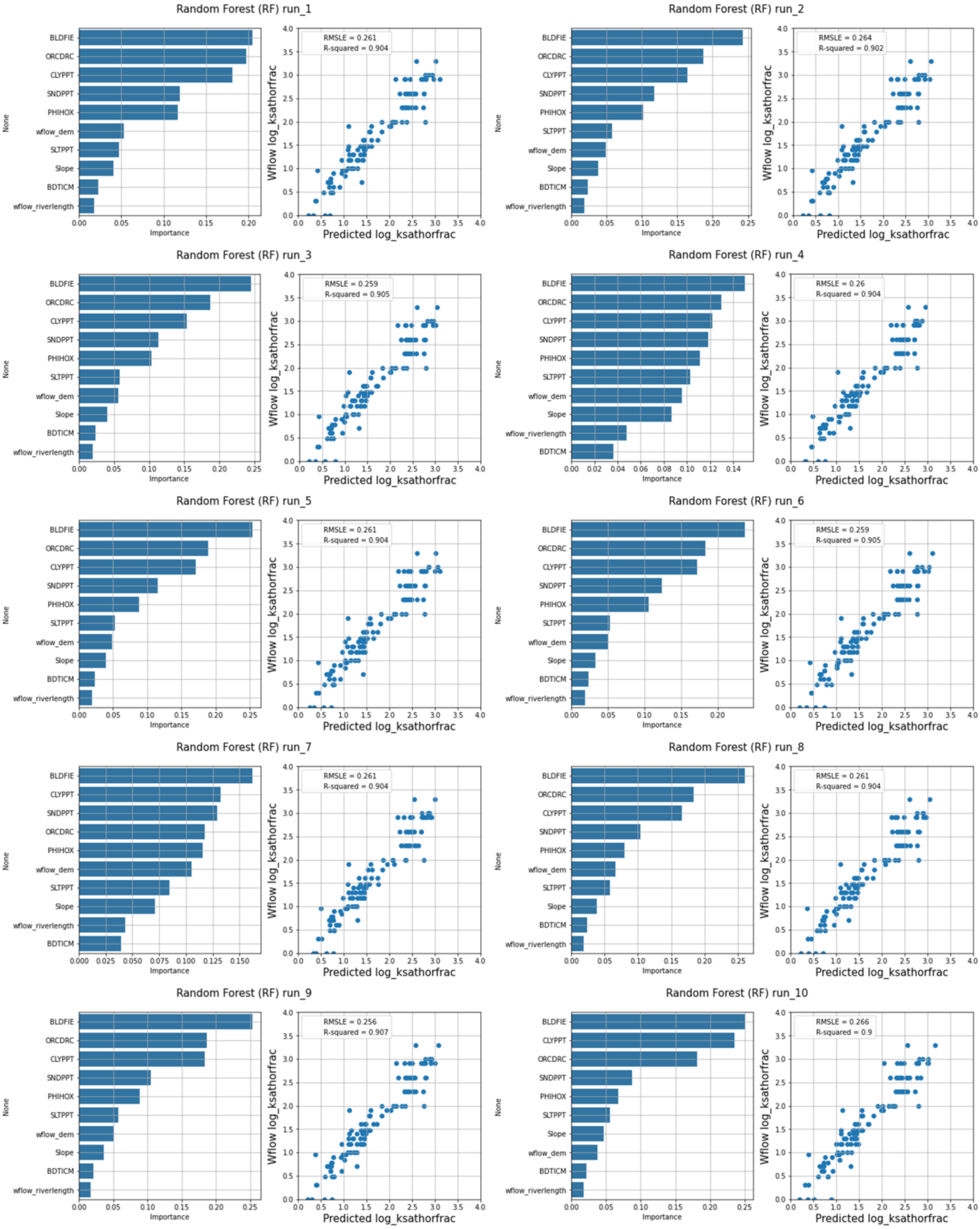


Figure 36: Distribution of relative importance for the predictors for ten runs of the first iteration of the updated PTF including the seven SoilGrids variables and elevation (wflow_dem), slope and drainage density (wflow_riverlength)

C.2. Randomness of the RF algorithm containing six SoilGrids variables in addition to elevation and slope

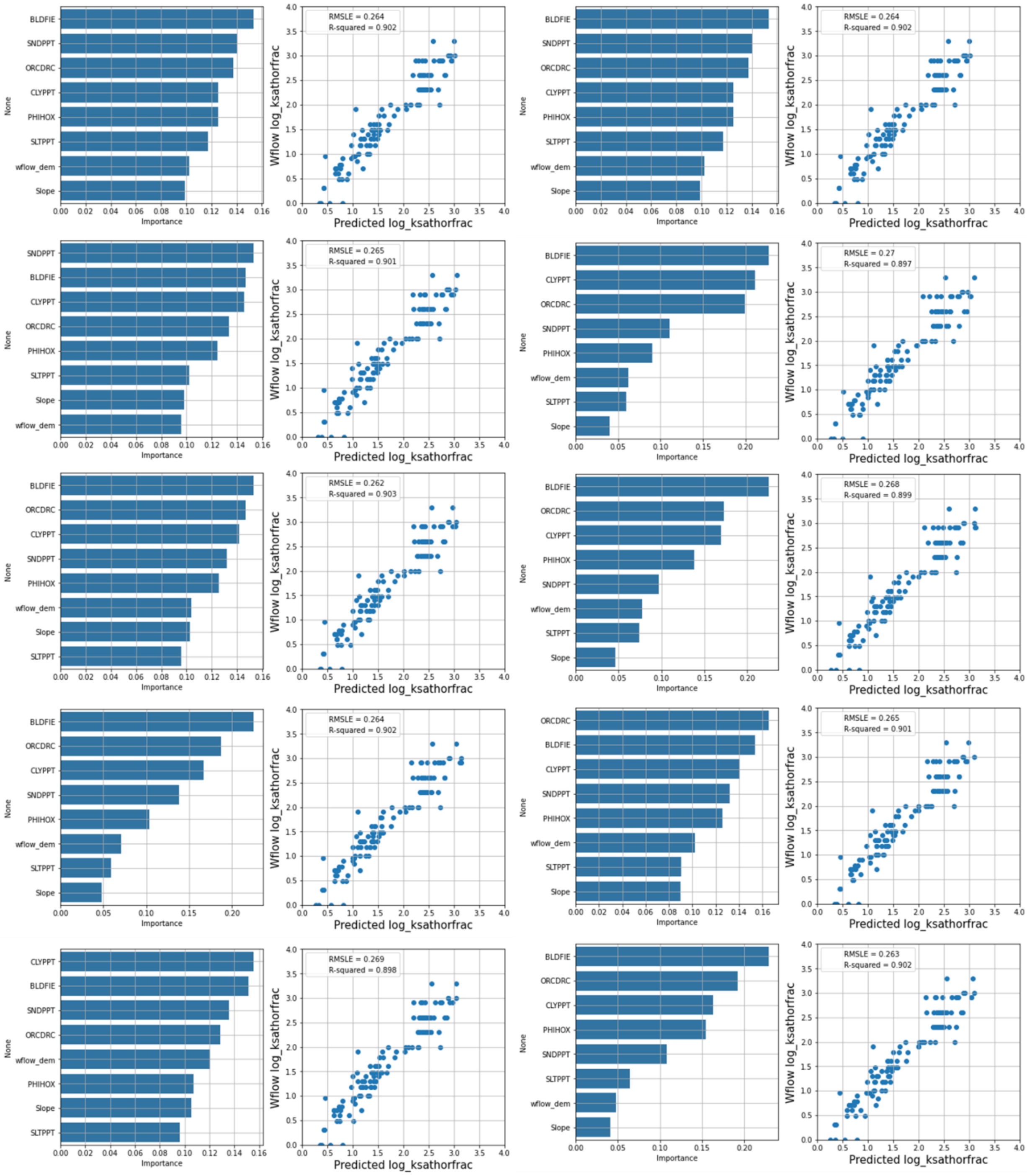


Figure 37: Distribution of relative importance for the predictors for ten runs the final selection of the updated PTF. Including the six SoilGrids variables (bedrock depth removed) and elevation (wflow_dem) and slope

C.3. spatial sensitivity of ten run KsatHorFrac map

Figure 38 shows the percentual difference between the lowest and highest value of KsatHorFrac for each grid cell across the ten distributed KsatHorFrac maps in the GB training subbasins. On average the difference between the extremes is 33% with 3% and 340% as minimum and maximum respectively.

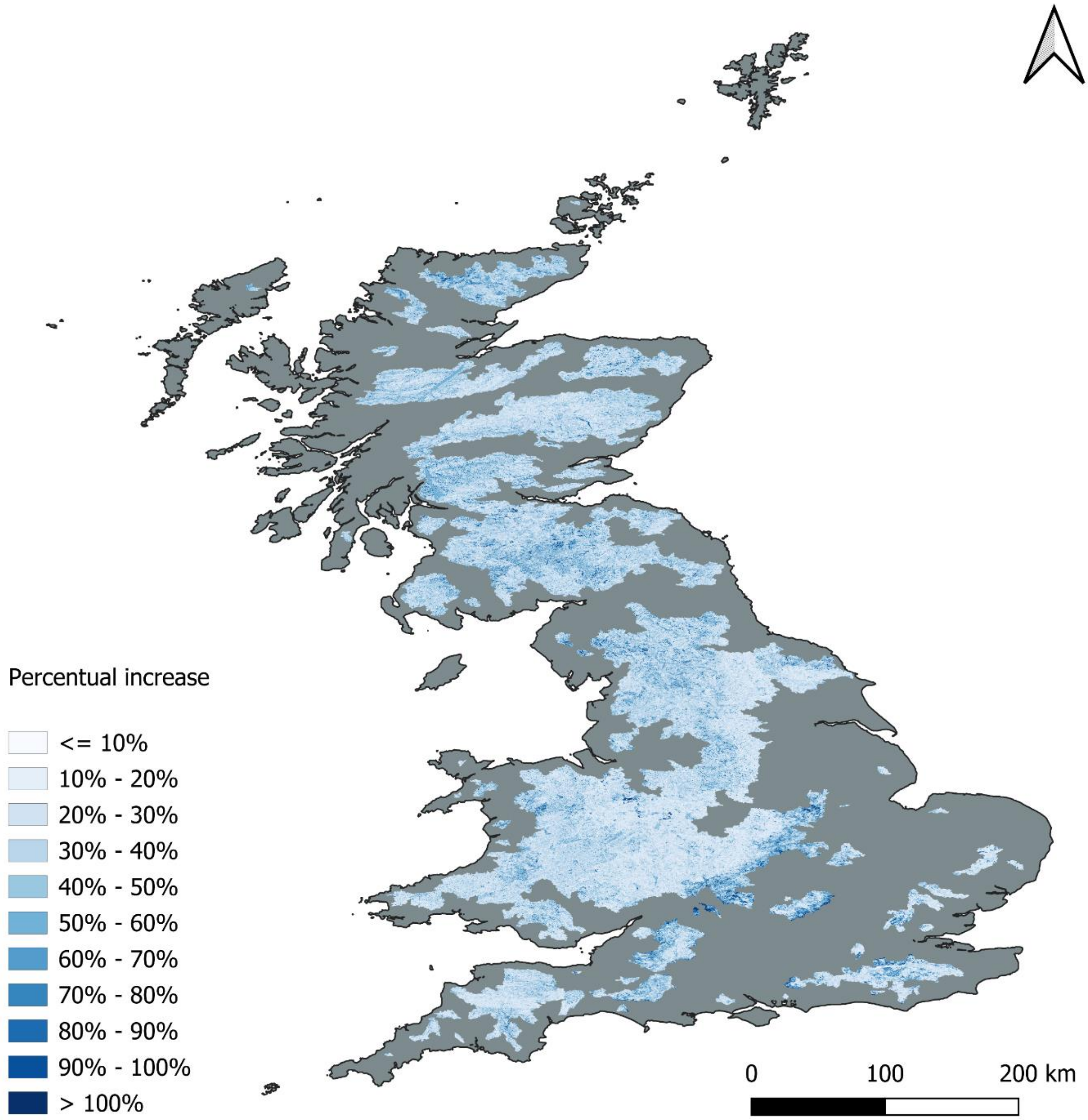


Figure 38: Percentual difference between the lowest and highest KsatHorFrac value per grid cell out of the ten runs for the GB training subbasins. Using the updated PTF

C.4. Ratio between updated and original PTF

The ratio of predicted KsatHorFrac between the updated and original PTF for the GB training subbasins can be seen in Figure 39. Here it can clearly be seen that the majority of the area under the histogram is above 1. Indicating that the updated PTF predicts higher KsatHorFrac values.

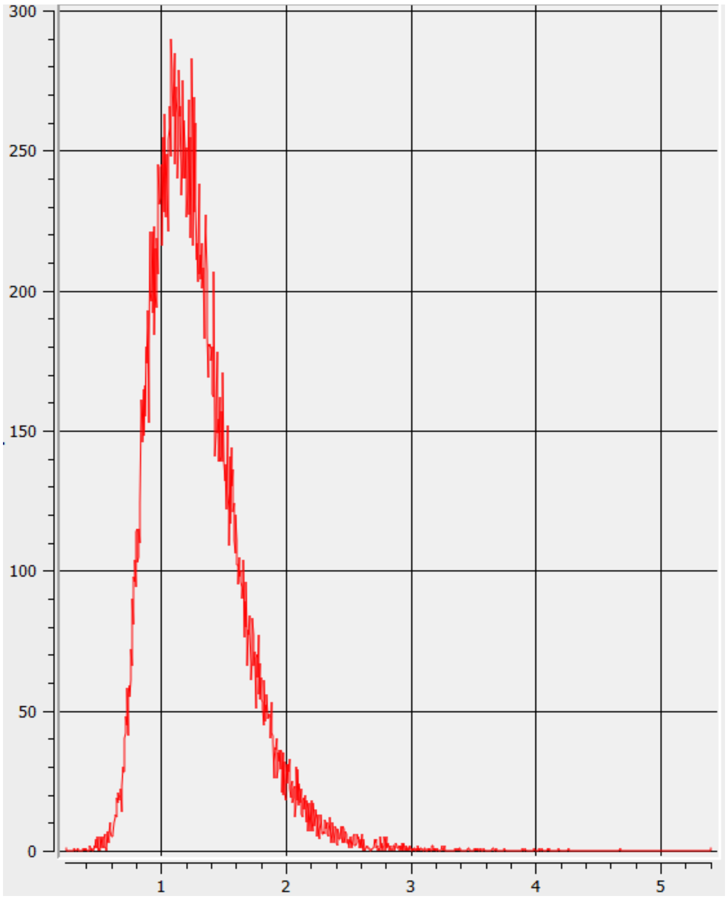


Figure 39: Histogram of the ratio of predicted KsatHorFrac between the updated and original PTF in the GB training subbasins