UTRECHT UNIVERSITY

FACULTY OF SCIENCE
Applied Data Science

Applied Data Science Master thesis

Hybrid streamflow modelling using machine learning and multi-model combination

First examiner:
Prof. dr. Derek Karssenberg (UU)

Candidate:
Hassan Ali (1635905)

Second examiners:
Dr. Edwin Sutanudjaja (UU)
Oriol Pomarol Moya, MSc (UU)
Michele Magni, MSc (UU)

30 June 2023

Abstract

Global hydrological models (GHMs) enable global estimation of freshwater availability, but their uncertainties and limitations hinder precise predictions. Multi-model combination (MMC) is a promising solution that combines the outputs of numerous hydrological models to create an ensembled output that surpasses the individual hydrological models. Moreover, the use of Machine Learning (ML) as a hybrid post-processing strategy is growing in popularity. However, there is a need to combine these two methods and investigate their performance in streamflow predictions. In this study, we demonstrate that using Random Forest (RF) as a non-linear MMC approach significantly enhances streamflow forecasts when multiple global hydrological models' outputs are combined. In streamflow forecasting, the RF-MMC method outperforms individual models and linear MMC approaches, demonstrating its potential. In addition, incorporating catchment attributes improved the generalizability of the RF-MMC method when tested on a river basin that was not in the training set. Significant potential exists for the application of RF-MMC to generate accurate streamflow forecasts, thereby providing valuable support for water resource management, flood mitigation, and decision-making processes. Future research can investigate additional machine learning algorithms and incorporate additional variables to improve the predictive ability and generalizability of MMC strategies in hydrological modelling.

Contents

1 Introduction

Water plays a vital role in numerous aspects of human well-being, ecological systems, economic operations, and geographical phenomena (Milly et al., 2005). The study pointed out that, some locations in the world could potentially see a significant reduction of up to 30% in runoff by 2050, which could have enormous regional and global implications. In their research, Zhang et al. (2023), highlighted the potential for water scarcity in different parts of the world to be exacerbated by climate change and increased social economic development. According to the same study, streamflow is critical in delivering key freshwater supplies for both humans and ecosystems. Reliable and accurate long-term streamflow prediction is essential to make sufficient and informed decisions in water resource management and flood mitigation (Shen et al., 2022).

There are various parameters that influence streamflow such as precipitation, evapotranspiration, air temperature and land use (Liu et al., 2016; Yaseen et al., 2017). The random variables involved in streamflow are complex and non-linear, posing challenges to a complete understanding of the hydrological cycle (Sharma & Machiwal, 2021). To simulate streamflow, hydrological models have been widely employed (Y. Chen et al., 2011; Jia et al., 2001; Yang et al., 2020). Global estimation of freshwater availability is made possible by global hydrological models (GHMs) (Eisner, 2016) and are characterized by a simplification of hydrological phenomena to be able to support multi-decadal hydrological simulations at a global scale (Zaherpour et al., 2019). However, the model outputs may contain significant errors despite the use of sophisticated calibration techniques (Shen et al., 2022). This is because hydrological models are imperfect representations of reality and they produce uncertain estimations even with access to highly accurate meteorological data (Beck et al., 2017). These uncertainties arise from the simplified structure of the models, inaccuracies in input data, and a lack of correct understanding of hydrological processes (Mohammadi et al., 2021).

A classic strategy for dealing with the uncertainties of conventional hydrological models is to combine the results of different models to produce a combined output that exceeds the individual hydrological models. This is known as Multi-Model Combination (MMC), and it involves combining the outputs of multiple models into a single variable using techniques such as mean or median. According to Shamseldin et al. (1997) and Xiong et al. (2001), no single model is universally optimal for all seasons and catchments, as they can vary significantly. As a result, a thorough combination of numerous estimates generated from different models can be expected to produce a more comprehensive and accurate depiction of catchment response than each individual model.

Shen et al. (2022) developed an error-updating procedure to correct streamflow predictions from PCR-GLOBWB using Random Forest (RF), an ensemble tree-based algorithm. The authors used both meteorological input and simulated hydrological state variables and runoff as predictors of observed streamflow for three catchments in the Rhine basin. The authors showed that RF can improve prediction errors for both calibrated and uncalibrated simulations. Magni et al. (2023) successfully extended such a hybrid framework to the global scale. The authors incorporated catchment attributes such as topography and river channel characteristics as additional predictors for observed discharge, along with state- and meteorological variables of PCR-GLOBWB. According to the findings, using static catchment features as additional predictors helped reduce the predictive error of the RF-based post-processing strategy.
Combining these two approaches, Zaherpour et al. (2019) proposed integrating machine learning (ML) and MMC in GHMs. The authors presented a method to blend simulated runoff

from five different GHMs in 40 large catchments across the globe, covering a total area of 100.000 km2. The authors defined two MMC methodologies: one using Gene Expression Programming (GEP) and one with a less sophisticated Ensemble Mean (EM). When compared to the top performing GHM, the GEP's median performance improvement exceeded 45%, reaching more than 100% when compared to the EM.

Given the uncertainties associated with traditional hydrological models and the potential benefits offered ML and MMC approaches, the aim of this study is to evaluate the use of a non-linear ML model such as RF as a MMC in improving streamflow prediction and model reliability. The decision to use RF as an MMC is predominantly influenced by Magni et al.'s (2023) demonstration of RF's performance in streamflow predictions by incorporating meteorological inputs, hydrological state variables, and catchment attributes. In comparison, we assess the performance of individual GHMs and a simple linear combination of their outputs using Multiple Linear Regression (MLR). Zaherpour et al. (2019) also employed MLR and compared it with a non-linear MMC approach using Gene Expression Programming (GEP), providing a relevant benchmark for our study.

This study thus aims at answering the following research question:

How can a machine-learning-powered multi-model combination approach improve streamflow predictions by combining outputs from different global hydrological models?

The following sub-questions are followed as a guide for the research process:

1. *Can a non-linear combination of GHMs simulations using RF achieve better performance for streamflow prediction compared to individual hydrological models?*
2. *How well does a RF-MMC generalize when trained and tested on different river basins?*
3. *How does combining streamflow simulations from different GHMs using MLR as a non-linear model improve streamflow prediction accuracy when compared to non-linear RF-MMC approach ?*
4. *How does the proposed RF-MMC strategy compare to the hybrid post-processing technique developed by* Magni et al. (2023).
5. *Does the inclusion of static PCR-GLOBWB attributes improve the generalization ability of the RF-MMC approach for predicting streamflow?*

To answer the first sub-question, we will combine the outputs from three GHM and use RF to predict streamflow. The predicted discharge will be evaluated using Kling-Gupta Efficiency (KGE) (Gupta et al., 2009) and compared to the KGE of the individual GHMs simulations. The RF model will be trained on stations in Rhine basin and subsequently tested on stations in Elbe and Maas basin to test its generalization ability. Additionally, MLR will be applied to forecast streamflow and examine whether a simple linear combination of GHM outputs can perform similar to the RF-MMC method. In addition, the hybrid post-processing method developed by Magni et al. (2023) will be compared to the proposed multi-model combination strategy. Finally, we examine whether catchment attributes enhance the generalizability of RF-MMC.

The report is organized as follows: we begin with a description of the study area and description of the dataset in section 2. We then explain the methodologies employed, including the GHMs, ML algorithms, and multi-model combination strategy in section 3. In section 4, the results of research are then presented. Section 5 concludes the research by discussing the research findings, limitations, and prospective areas for further research.

2 Study area and Data

2.1 Study area

In this study, we look at 50 European catchments with a variety of hydrological features. These catchments are spread throughout three major river basins: the Rhine, the Elbe, and the Maas. Thirty of the total fifty catchments are in the Rhine basin, thirteen in the Elbe basin, and the other seven in the Maas basin (Figure 1). These stations' catchment sizes range from 10,000 to 160,000 square kilometres, with stations in the Maas basin typically covering smaller catchment areas.

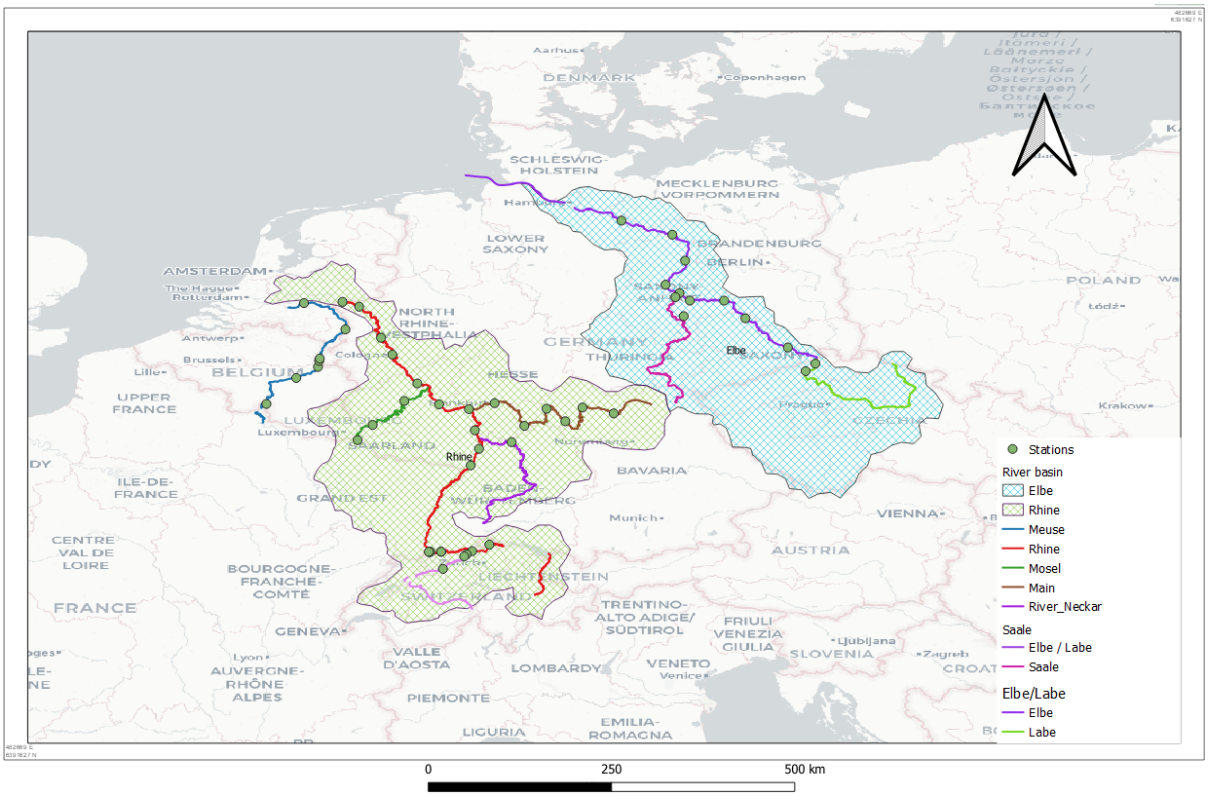

Figure 1: Study area, Rhine, Elbe and Meuse Basin. The green dots show the location of the gauging stations, and the lines indicate the rivers.

The Rhine basin is characterized by a variable flow regime that is governed by factors such as precipitation patterns, snowmelt, and tributary contributions. The average annual precipitation ranges from 500 mm in the low-lying areas to 2000 mm in the Alpine regions (Khanal et al., 2019). This river basin has a varied land use, such as agricultural areas, urbanized areas, and forested zones, with the area surrounding Basel (Switzerland) having more grassland, surface water and glaciers. The upper part of the Rhine basin is in the high altitudes of the Alpine region. As a result of this, the streamflow is mainly influenced by snow and glacier melt in spring and summer (Shen et al., 2022). The streamflow regime of the lower parts of the Rhine basin is predominantly influenced by heavy rainfall and reaches its peak in late winter and is lowest in the summer.

The Elbe River basin flows across Germany from the Czech Republic to the North Sea. The Elbe River, like the Rhine, has a discharge regime that is impacted both by rainfall and snow. Water levels are normally higher in the winter and spring, and lower in the late summer and autumn. Precipitation in the Elbe River basin varies with elevation. Higher altitude place getting more rain compared to lower altitude. This fluctuation adds to the Elbe River's distinct

discharge characteristics, influencing its flow regime and water levels throughout the year (Hesse, 2018). As mentioned by the same author, the yearly rainfall in the Elbe River varies across different altitudes, ranging from 1700 mm in higher altitude places to as low as 450 mm in lower altitude areas. It is worth noting that the lower altitude zone, which accounts for around one-third of the Elbe River basin and is in both Germany and the Czech Republic.

The Maas River begins in the French Alps and runs through Luxembourg, Belgium, and the Netherlands before reaching the North Sea. However, for computational considerations, this study concentrates primarily on the Maas basin's Dutch and Flemish regions. The average annual precipitation in this part of the river basin is 700-800 mm (de Wit et al., 2007). The discharge regime of the Maas River basin is primarily influenced by rainfall and is highest in the winter and at the beginning of spring. According to de Wit et al. (2007), a decrease in the average discharge occurs naturally in the autumn months. Unlike the Elbe and Rhine rivers, the Meuse River's streamflow is consistent throughout the year and does not have high peaks caused by snowmelt or glacier runoff.

2.2 Data

As mentioned in the introduction, in this study we will combine the outputs of three GHMs to improve streamflow predictions. The outputs of the GHMs have been made available on the eartH2Observer Water Cycle Integrator (WCI) portal (Schellekens et al., 2017). The models in the eartH2Observer project, including the ones used in this study, are fundamentally different from each other, but they are all driven by the same daily 0.5° WATCH Forcing Data ERA-Interim (WFDEI) meteorological dataset (Weedon et al., 2014). This allows for the comparison and study of the outputs of the various models. The retrieved dataset has a monthly temporal resolution and covers 1979 to 2012.

A preliminary investigation was conducted to examine the data resources accessible on the WCI data portal that were relevant to the study. This review provided insight into the data availability, quality, and overall fit of the data for the research aims. Based on the preliminary research, the following three GHMs have been selected: WaterGAP3, PCR-GLOBWB and LISFLOOD. Before the final run of the GHMs, Schellekens et al. (2017) used specific initialization processes for each model to account for the distinct characteristics of each model and to ensure that they accurately depicted the climatic conditions. Beck et al. (2017) built upon the work of Schellekens et al. (2017) and evaluated a set of model outputs described in the same study with discharge from 966 medium-sized watersheds and found that the calibrated models performed best. In their effort, the scientists calibrated several GHMs, including WaterGAP3 and LISFLOOD (excluding PCR-GLOBWB) and made updates to the outputs of these GHMs on the eartH2Observe Water Cycle Integrator (WCI) data portal.

The variable importance found by Shen et al. (2022) is considered to select two to three state-variables for each of the GHMs. The authors discovered that the influence of state variables varied depending on the characteristics of the catchment. Snow water equivalent and surface water storage variables were shown to be critical in minimizing prediction errors in catchments where glacier melt contributed considerably to the discharge regime. In contrast, the authors stressed the relevance of groundwater components as important contributors in places where rainfall plays a dominant role in the discharge regime. The number of variables investigated in this study is limited due to computing restrictions and the short period of the thesis. Besides the discharge and meteorological variables, the following three state variables are considered for each GHM: snow water equivalent, soil moisture, and surface water storage. Table 1 gives

a complete description of these variables. Furthermore, the GHMs considered in this study are described in paragraphs 2.2.1 to 2.2.3.

2.2.1 WaterGAP3

Water – Global Assessment and Prognosis-3 (WaterGAP3) hydrological model is part of the WaterGAP suite of models used to assess and predict global water resources and their use. The first version was initially proposed in 1996 and is described in Alcamo et al. (1997). The model included a simpler representation of water use and availability (ALCAMO et al., 2003) and has since been further developed and improved (Müller Schmied et al., 2021). Additionally, WaterGAP consists of five sectoral water use models (irrigation livestock, household, manufacturing, and thermal power plant cooling) (Döll & Siebert, 2002; Flörke et al., 2013) and a large-scale water quality model (Eisner, 2016; Schellekens et al., 2017). WaterGAP3 is a grid-based, integrative assessment tool that is used to investigate the current state of global freshwater resources. The model is also used to examine the potential impacts of global changes in the water sector, particularly in the context of climate change and human interventions (Döll et al., 2003; Eisner, 2016; Müller Schmied et al., 2021b). The WaterGAP3 hydrological model is simulated at daily timestep; however, monthly data aggregation is also available. The model is used to simulate both the global and local water cycles. This is a water balance model that incorporates interception, soil water, snow, groundwater, and surface water as the key components of water storage. As elaborated by Schellekens et al. (2017), the WaterGAP3 model made available on the eartH2Observe data portal was first re-run ten times using the first year of accessible meteorological forcing to initialize the storage component before the final run was made. Furthermore, the model was calibrated before being regionalized to ungauged catchments by multiple linear regression (Beck et al., 2017).

2.2.2 PCR-GLOBWB

PCR-GLOBWB (Sutanudjaja et al., 2018) is a global hydrological model that uses a grid-based technique to represent the Earth's water cycle dynamics. It simulates streamflow and a wide range of state variables, allowing for a complete analysis of global hydrological dynamics. The framework structure in interconnected modules, namely: an irrigation and water consumption model, a meteorological model, a land surface model, a groundwater model, a surface water routing component and irrigation and water use model. However, it should be noted that the data provided on the eartH2Observe WCI data portal (tier 1) does not encompass this particular water use component from the standard PCR-GLOBWB model. Before the final run of PCR-GLOBWB, Schellekens et al. (2017) carried out a 68-year initialization period by conducting two consecutive preliminary runs from 1979 to 2012. The parameters rely solely on previous variable estimation and were not calibrated.

2.2.3 LISFLOOD

LISFLOOD (Van Der Knijff et al., 2010) is a grid-based hydrological model that simulates hydrological cycle. The model can replicate the full water cycle, from rainfall to water in rivers, lakes, and groundwater. The model has a wide range of applications, including water and climate studies as well as forecasting floods and droughts. LISFLOOD consist of numerous modules that simulate different hydrological processes. This platform offers complete hydrological dynamics study and simulation. These models contain potential and actual evapotranspiration, snow cover and melt assessment, and soil water balance analysis across three layers. Prior to starting the main simulation, a complete run from 1979 to 2012 was performed to create the model (Schellekens et al., 2017). The model was then calibrated using

WFDEI forcing, and its performance was thoroughly evaluated using daily streamflow data (Beck et al., 2017).

Table 1: Hydrological variables from GHMs and Meteorological Forcing variables used for the study.

Variable	Source	Unit	Explanation
lis_dis	LISFLOOD	m/day	Simulated river discharge.
lis_SWE	LISFLOOD	kg/m^2	Snow water Equivalent
lis_SurfMoist	LISFLOOD	kg/m^2	Surface soil moisture
pcr_dis	PCRG-LOBWB	m/day	Simulated river discharge.
pcr_SWE	PCR-GLOBWB	kg/m2	Snow water Equivalent
pcr_SurfMoist	PCR-GLOBWB	kg/m2	Surface soil moisture
pcr_SurfStor	PCR-GLOBWB	kg/m2	Surface water storage (lakes, reservoirs, rivers, and inundated water)
wg3_dis	WaterGAP3	m/day	Simulated river discharge.
wg3_SurfStor	WaterGAP3	kg/m2	Surface water storage (lakes, reservoirs, rivers, and inundated water)
wg3_SWE	WaterGAP3	kg/m2	Snow water Equivalent
wg3_RootMoist	WaterGAP3	kg/m2	Root zone soil Moisture
meteo_rain	Meteorology	kg m-2 s-1	Rainfall rate.
meteo_tair	Meteorology	K	Temperature measured in Kelvin degrees.

3 Methods

In this research, two ML models are applied: Random Forest and Multiple Linear Regression. In the following sections, we will go over the data pre-processing approach, the ML models used, and the setup and evaluation of the models.

3.1 Pre-processing

3.1.1 Observed discharge

Observed discharge data was obtained from the dataset made public by Magni et al. (2023). This is the target variable in the ML modelling. The authors downloaded the data from the Global Runoff Data Center (GRDC) and chose stations based on two constrains: stations with at least a 10,000 km2 upstream area, and at least one year of data between 1979 and 2019. They also converted the discharge (m3/s) into flow depth (m/day). This transformation is given as:

$$f_d = q\frac{86400}{A_{\text{ups}}}$$

Whereby "f_d" denotes the flow depth in meters per day (m/d), "q"defines the discharge in cubic meters per second (m3/s). The numerator 86400 denotes the number of seconds in a day (s/d-1) and "A_{ups}" expresses the upstream area in square meters (m2).

3.1.2 Meteorological and GHM variables

For each variable from Table 1, a separate NetCDF file was retrieved from the eartH2Observe WCI data portal. The files were then divided into single monthly timesteps to minimize the computational time for subsequent steps. Each variable was then resampled to 0.5-degree spatial resolution and uniform spatial extent, allowing for accurate and comparative analysis on a unified scale. Upstream normalization was performed on all the variables (except discharge variables) using PCRaster Python framework (Karssenberg et al., 2010).

For the extraction of station specific values, the GRDC stations within the study region that satisfied the constraints defined above were used. The extraction of station values was done by locating the nearest pixel in the dataset that matched the coordinates of each GRDC station. This is repeated for each timestep and variable. The extraction procedure concludes by saving the extracted values along with the corresponding date of each station into a Comma-separated values file (csv). Subsequently one predictor table was created for each station which contains all the variables by combining the extracted csv files. The discharge variables form the GHMs were also converted into flow depth. Furthermore, the extracted GRDC discharge is added into the predictor table, creating a complete dataset for training and validation. The state variables and meteorological input variables were standardized using z-score normalization. This ensures a zero mean and a standard deviation of one allowing for a clearer input to the machine learning models. During this stage, we also cleaned the dataset by removing observations with no data. Finally, the dataset was divided into separate subsamples based on GRDC stations for cross-validation purpose in order to overcome overfitting. This is required to evaluate the model's performance on unobserved data and to assure its generalizability. Figure 2 shows the entire workflow of the data-processing steps that were carried out.

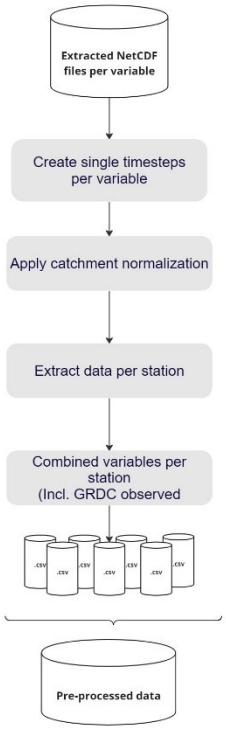

Figure 2: Data pre-processing workflow for creating predictors table. This was performed using Python.

3.2 Machine Learning models

3.2.1 Random Forest

Random Forest (RF) (Breiman, 2001) is a ML algorithm that combines bagging and feature sampling strategies to address the problem of overfitting in ML models. Bagged trees are formed by taking a random sample from the dataset with replacement. These trees are frequently high in variance, and reduced bias. However, this complexity can result in overfitting, which can be prevented by taking the average of all projections. Furthermore, with RF, features are sampled randomly at each tree node. Because an independent sample of the attributes is chosen at each node, the tree will overcome overfitting on features that are corelated. In this study, RF was implemented using ranger package from R (Wright & Ziegler, 2017).

To enhance the performance of the RF model, hyperparameter tunning was performed on two important parameters: the number of trees (ntrees) and the number of features (mtry). The tuning procedure was divided into two sections. The initial focus was on fine-tuning the mtry parameter while maintaining the number of trees fixed at 200. This method enables us to determine the best number of attributes to consider at each node. The same method was then used to tune the number of trees while using the mtry value obtained in the previous phase. This two-step tuning process was adopted from the work of Magni et al. (2023).

3.2.2 Multiple linear regression

Multiple linear regression is executed by simply taking only linear dependencies between the response variable and each of the independent variables. The MLR model was implemented using the *lm* function from the base package in R. The general equation of the MLR models is given below.

$$y = \beta_0 + \beta_1 x_1 \cdot \beta_2 x_2 \cdots\cdots \beta_n x_n + \epsilon$$

The term y is the dependent variable, the letter β_0 represents the constant term, β_1.... β_n are the regression coefficients, and ϵ is the error term.

MLR may be limited by multicollinearity. When multicollinearity exists, x_1 may be linearly dependent on another explanatory variable, such as x_n, resulting in inaccurate coefficient estimates and an unreliable model. Lafi and Kanene (1992) identified four main symptoms of multicollinearity, including: (1) large standard errors, (2) unexpected coefficient signs due to high correlation with other variables, (3) high correlations between predictor and response variables without statistical significance, and (4) correlation coefficients among explanatory/predictor variables are large but the overall ability of the model's ability to explain the variation in the response variable is considerably low.

In order to test for multicollinearity, the Variation Inflation Factor (VIF) (Marriott et al., 1985) will be used. The VIF is defined as:

$$VIF_j = \frac{1}{1 - {R_j}^2}$$

where ${R_j}^2$ is the coefficient of determination for the regression of xj on the remaining variables. There is a multicollinearity if the VIF is larger than 10. The VIF factor is calculated using the VIF function in the car package in R.

3.3 Model setup and evaluation

3.3.1 Experiment setup

To investigate the best combination of variables, we will explore three distinct variable combinations, each of which will provide useful insights into the predictive capabilities. We will then run for each variable setup a separate model and investigate how well the setup can predict observed streamflow. Besides the predicted streamflow using the ML models, we will also compute the arithmetic mean of the GHMs. The arithmetic mean of the discharge of GHMs will be given as average in the results section. A visual representation of the modelling strategy with different variable setups is shown in Figure 3.

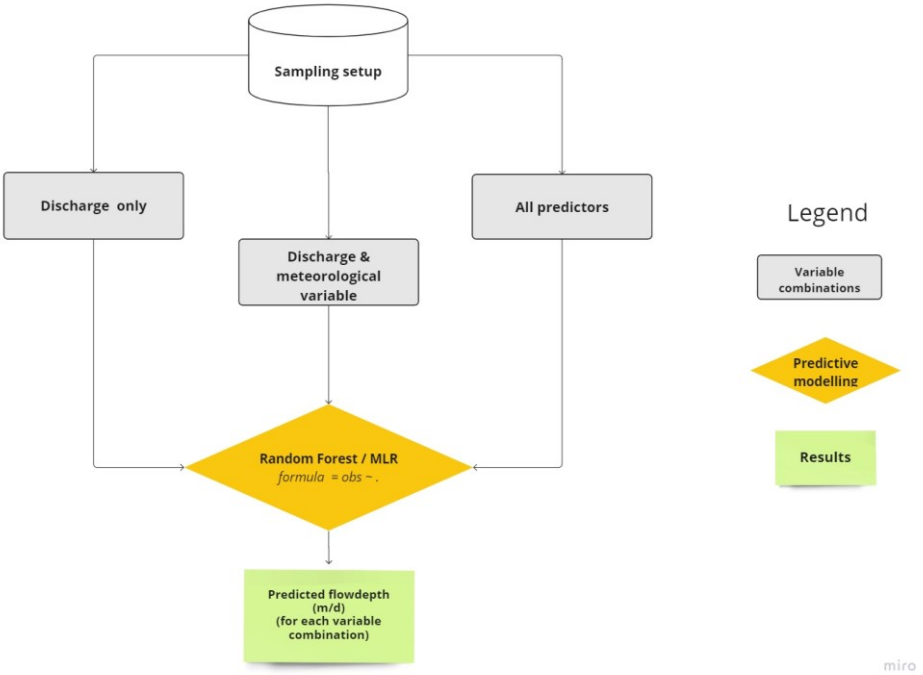

Figure 3: Experiment setup and various variable configurations

3.3.2 Evaluation metric

The performance of the ML in predicting streamflow is evaluated using Kling-Gupta Efficiency (Gupta et al., 2009). This metric quantifies the similarity between measured and predicted discharge with a single number. The equation of KGE and its components is given below.

$$KGE = \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}$$

$$r = Cov\left(\frac{q_p,\ q_o}{\sigma_p \cdot \sigma_o}\right)$$

$$a = \frac{\sigma_p}{\sigma_o}$$

$$\beta = \frac{\mu_p}{\mu_o}$$

Where "r" denotes the correlation coefficient between predicted and observed discharge, "a" represents bias ratio based on standard deviation and "β" represent variability ratio based on mean. The resulting KGE value from the above equation ranges from -∞ to 1, with a perfect performance having a value of 1. KGE values that are less than -0.41 indicate that a model fails to outperform the mean flow benchmark and is therefore viewed as bad model performance (Knoben et al., 2019).

3.3.3 Validation setup

To validate the ML- MMC, four distinct validation setups were used to evaluate the model's performance and generalization capabilities. For the first two setups a 5-fold cross-validation setup was applied while the last two setups were designed to test the model's ability to generalize to new river basins. Figure 4 shows these validation setups visually. For each setup, the validation is carried out by applying the trained ML model to its matching test dataset on station-by-station basis.

The first setup, *all_stations*, utilizes a 5-fold cross-validation in which 35 training set stations were randomly picked from all available stations for each fold. The remaining stations were employed to test the trained model. This method allowed the model's efficacy to be assessed across a diverse set of stations with various features. The second setup is called *rhine_only* setup and uses 5-fold cross-validation as well. Using this set up, we only sample from stations in the Rhine basin. This will enable us to investigate the model's performance on a specific region of interest.

To assess model generalization capabilities, setups C and D were created. The model was first trained on the catchments in the Rhine watershed and then tested on the catchments in Elbe and Maas for setup *rhine_elbe* and *rhine_maas* respectively. These two setups were evaluated only once (without cross-validation) for both configurations to quickly examine how well the model performed when applied to entirely different river basins.

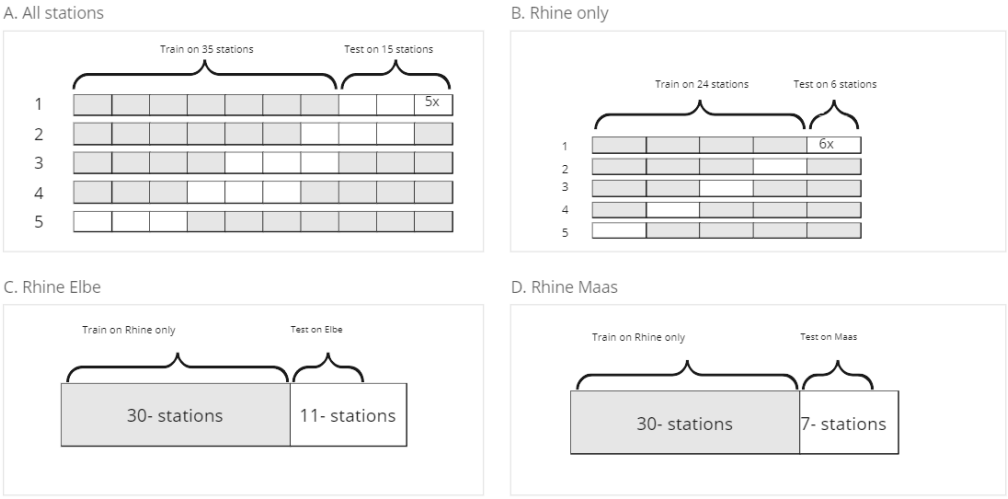

Figure 4: Four validation setups. (A) applies 5-fold cross-validation on all stations, sampling 35 stations for training. (B) focuses on Rhine stations using 5-fold cross-validation. C and D test model generalization by training on Rhine and testing on Elbe

3.3.4 Benchmark setup

The performance of the RF-MMC (*allpredictors*) is compared with the alternative hybrid post-processing strategies proposed by Magni et al. (2023) (*PCR-allpredictors*). The purpose of this comparison is to see how well the RF-MMC predicts streamflow compared to the method developed by Magni et al. (2023). The benchmarking dataset contains a total of 53 attributes, 27 of which are specific to catchment characteristics. The remaining variables include the full set of PCR-GLOBWB time-dependent state variables as well as meteorological variables. We can limit the possible influence of catchment features on benchmarking results by concentrating on only one river basin, allowing for a more precise and specific evaluation of the MMC technique. For that reason, the *rhine_only* setup will be used as a cross-validation setup.

Additionally, we investigate whether additional catchment attributes could improve the generalization ability of the RF-MMC approach. To achieve this, we added the static attributes from PCR-GLOBWB to the *allpredictors* setup, creating a new dataset with an additional 27 features (*allpredictors_catch*). The validation configurations *rhine_elbe* and *rhine_maas* will be utilized to test the performance of the enhanced data set. All the different variable setups considered in this study including the two benchmarking setups are given in Table 2.

Both setups described above were run with mtry of 25 and ntrees of 300, based on the findings of Magni et al. (2023).

Table 2: All the variable setups considered in the study, including the benchmarking dataset retrieved from Magni et al. (2023) study.

Setup	**GHM-Discharge**	**GHM-Meteo**	**GHM-State variables**	**PCR-GLOBWB Variables[1]**	**PCR-GLOBWB Catchment attributes[2]**
dis_only	*X*				
dis_meteo	*X*	*X*			
allpredictors	*X*	*X*	*X*		
allpredictors_cacth	*X*	*X*	*X*		*X*
PCR_allpredictors				*X*	*X*

[1] Variables from Magni et al. (2023)
[2] Variables from Magni et al. (2023)

4 Results

4.1 Random Forest

4.1.1 Tunning and training

The optimum hyperparameters applied for each variable setup are given in Table 3. Due to the limited number of predictors, hyperparameter optimization, as described in section 3.2.1, was not conducted for *all_dis* setup. Therefore, for this configuration, the mtry parameter was set to 1 and the ntrees value was set to 300. Please see Appendix A for a plot demonstrating the RF hyperparameter tuning outcomes for *all_dis_meteo* and *allpredictors* setup.

Table 3: Optimal RF hyperparameters for the different variable setups

Variable setup	ntrees	mtry
all_dis	300	1
all_dis_meteo	300	3
allpredictors	300	4

During training, each variable setup was run separately for the different cross-validation setups. After training the RF on the dataset, the variable importance was extracted from the model and saved for further investigation. Figure 5 shows the variable importance plot of the different variable setups and when *all_stations* cross-validation is used. There was no difference found in the variable importance between the different cross-validation setups. From this figure, we can observe that three out of the four WaterGAP3 variables to be positioned in the top half of the most informative variables, and three out of four PCR-GLOBWB variables were in the bottom half. Following that, the KGE plots of the discharge variables provided further proof that WaterGAP3 outperformed PCR-GLOBWB in terms of simulating observed discharge. This distinction stems from the fact that the parameters in WaterGAP3 and LISFLOOD were calibrated, whereas the parameters in PCR-GLOBWB were not.

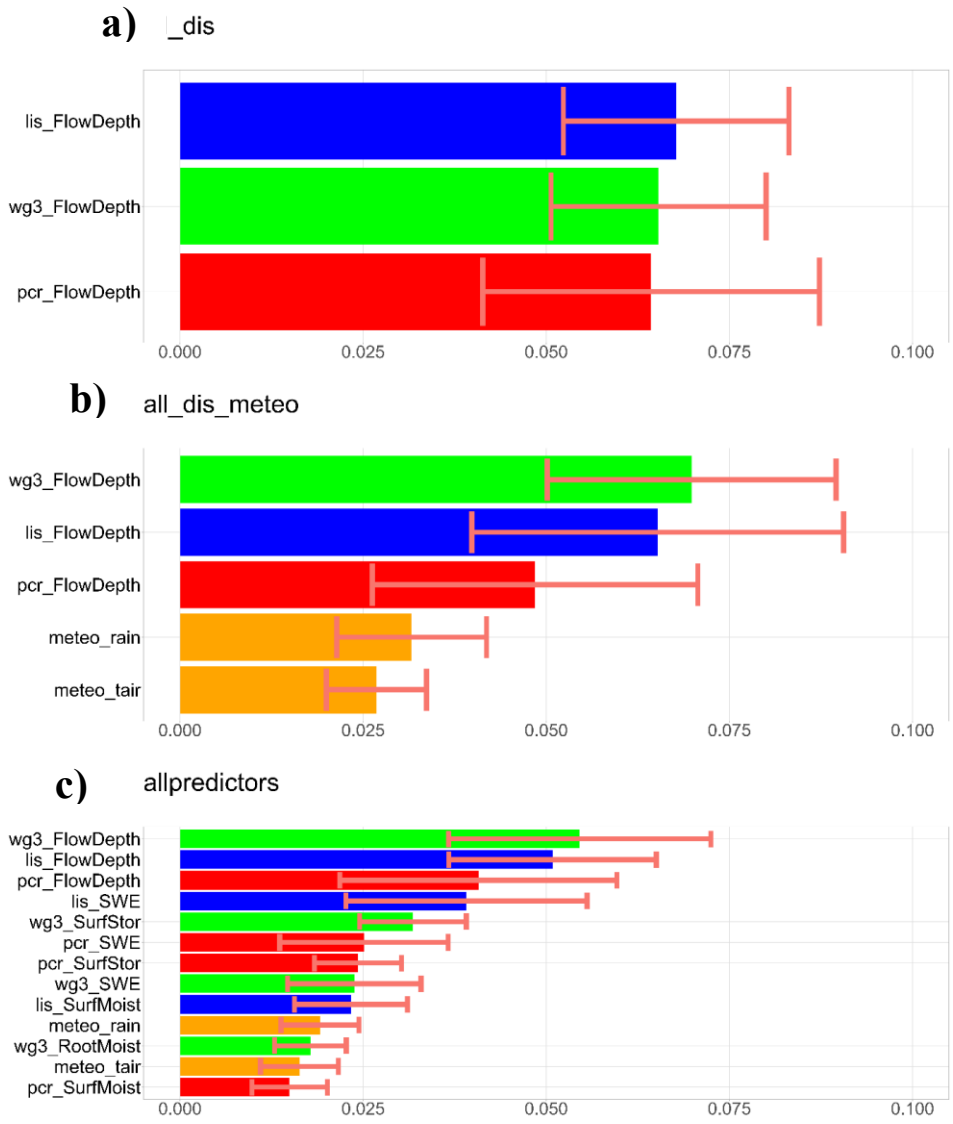

Figure 5: Square rooted mean decrease in impurity values for the three variable setups. Green colour indicate WaterGAP3 variable, blue colour represent LISFLOOD variables, red colour show the PCR-GLOBWB variables and orange colour indicate the meteorological input. The orange bar indicates the standard deviation setup. .

4.1.2 Validation setup: all_stations and rhine_only

The results for the RF model for *all_stations* and *rhine_only* setups are shown in Figure 6 and Figure 7 respectively. The figure shows the cumulative distribution (CDF) of KGE values for each of the RF models (solid lines) in combination with the individual GHMs discharges and their mean (dashed lines). For all three variable configurations *(all_dis, all_dis_meteo and allpredictors*), the RF predictions showed a clear and considerable improvement over the individual process-based GHMs. WaterGAP3 and the average discharge performed better than the LISFLOOD and PCR-GLOBWB discharge.

Table 4 shows the percentage gains in KGE values for each setup compared to the average discharge of the GHMs. The median KGE values in this table and subsequent tables are calculated in two steps. First, sub-sample medians are calculated to mitigate the effect of

outliers on the averaging procedure. The average of these median values is then presented for analysis. Under the setup *all_stations* sampling strategy, the *all_dis* variable setup improved by 18% while *all_dis_meteo* improved by 25%. For the cross-validation rhine_only, *all_dis* setup improves the KGE by 8% while *all_dis_meteo* had an improvement of 19% over the average discharge. The *allpredictors* configuration shows the greatest improvement of 39% and 34% for cross-validation setup *all_stations* and *rhine_only* respectively.

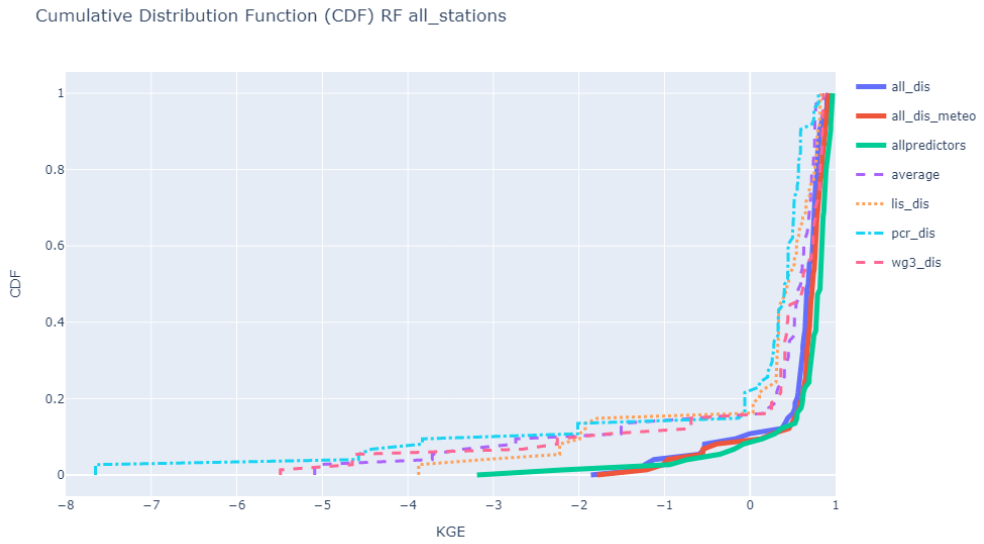

Figure 6: Cumulative distribution Function (CDF) of KGE values for all_stations setup. The bold lines show the predicted discharge of the different variable setups, and the dashed lines indicate the discharge of the individual GHMs and their mean.

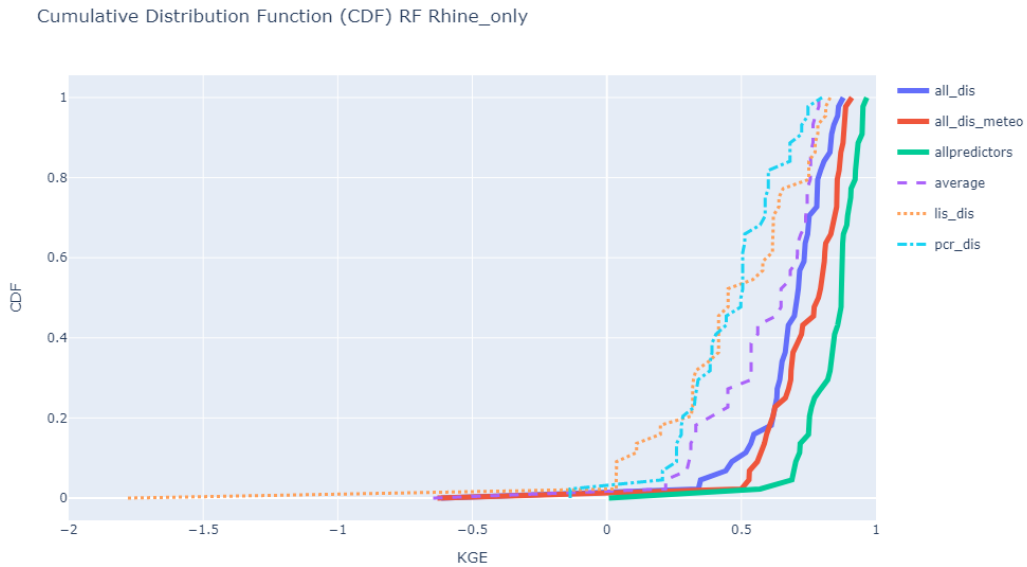

Figure 7: Cumulative distribution Function (CDF) of KGE values for rhine_only setup. The bold lines show the predicted discharge of the different variable setups, and the dashed lines indicate the discharge of the individual GHMs and their mean.

Table 4: RF results: Percentage KGE improvement over average discharge, setup All stations and Rhine only.

Setup	Median KGE All stations	KGE performance All stations	Median KGE rhine only	KGE performance rhine only
average	0.58	-	0.64	-
all_dis	0.69	+18%	0.71	+8%
all_dis_meteo	0.73	+25%	0.79	+19%
allpredictors	0.81	+39%	0.87	+34%

4.1.3 Validation setup: rhine_elbe and rhine_maas

This validation setup is designed to examine the generalization ability of the RF-MMC for streamflow prediction. This is done by training the RF-MMC with catchments in Rhine basin and was validated on catchments in Elbe and Maas basins. The results of these two setups are given in Figure 8 and Figure 9.

For the additional setups, the RF-MMC performance decreased significantly for all three variable configurations. The predicted discharge with *all_dis* showed a decrease of 10% and 27% for *rhine_elbe* and *rhine _maas* respectively. Model setups where meteorological variables and discharge variables were combined showed a decrease of performance of 26% and 44% compared to the average discharge. Surprisingly, *allpredictors* had the largest decrease in performance with a 144% decrease in performance compared to the average in *rhine_elbe* setup. This is because the model captures the specific patterns and noise that are present in the Rhine catchments which are not present in the Elbe or Maas. Based on these results, the RF fails to generalize to new river basin that was not included in the training sets, resulting in poor performance. It is worth mentioning that some transfer of knowledge can be observed in the *rhine_maas* setup (see Figure 9) especially in stations where the individual hydrological models are underperforming. Additionally, some GHMs have KGE values of less than -5, whereas all forecasts have KGE values greater than -1.

Table 5: RF results: Percentage KGE improvement over average discharge, setup Rhine Elbe and Rhine Maas.

Setup	Median KGE Elbe	KGE performance Elbe	Median KGE Maas	KGE performance Maas
average	0.59	-	0.52	-
all_dis	0.53	-10%	0.38	-27%
all_dis_meteo	0.44	-26%	0.29	-44%
allpredictors	-0.26	-144%	0.52	0

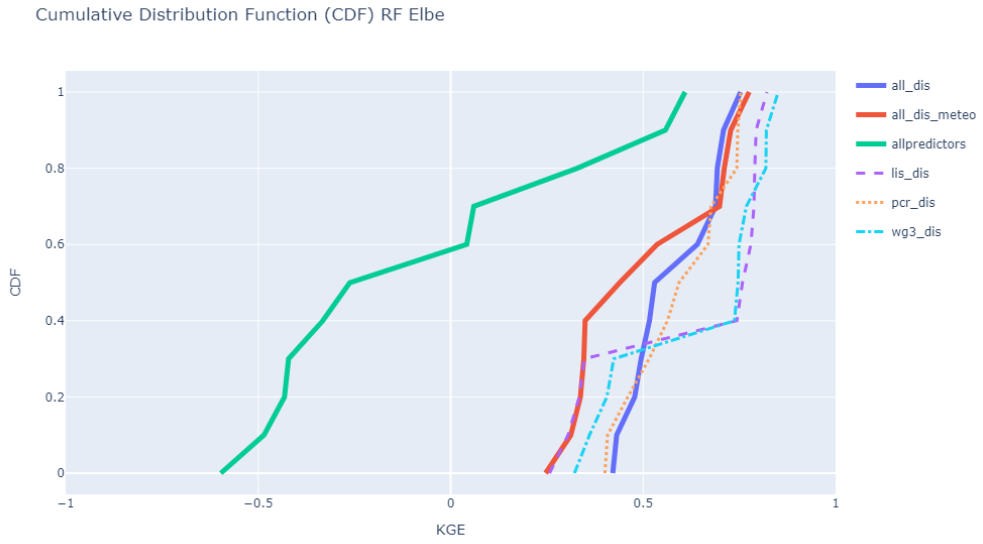

Figure 8: Cumulative distribution Function (CDF) of KGE values for rhine_elbe setup. The bold lines show the predicted discharge of the different variable setups, and the dashed lines indicate the discharge of the individual GHMs and their mean.

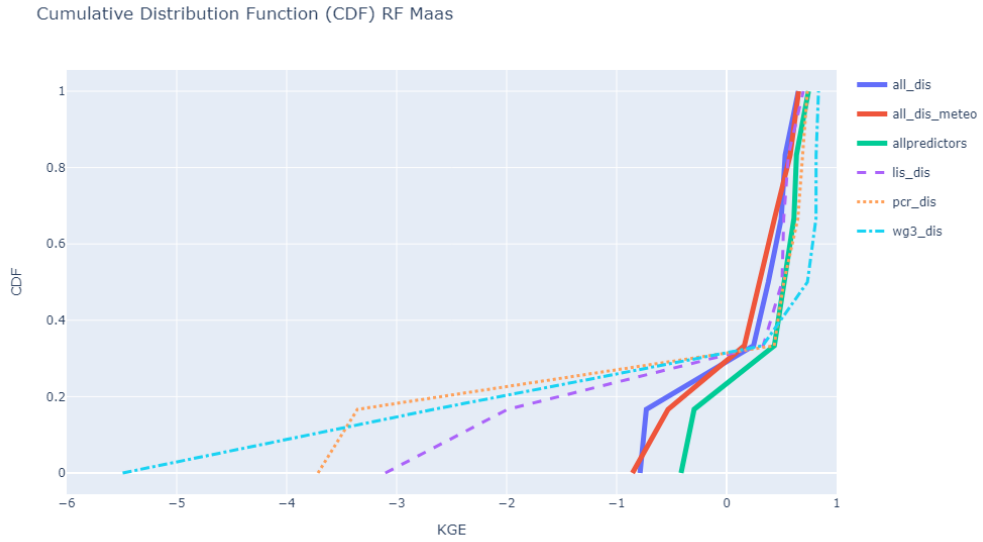

Figure 9: Cumulative distribution Function (CDF) of KGE values for rhine_maas setup. The bold lines show the predicted discharge of the different variable setups, and the dashed lines indicate the discharge of the individual GHMs and their mean.

4.2 Multiple linear regression

The results of the MLR are presented in Table 6 and Table 7. The MLR streamflow predictions failed to show any improvements over the individual GHMs. For variable setup *all_dis*, the MLR performed worse than all the individual GHMs. Appendix D shows the plots of the cumulative distribution (CDF) of KGE values for each of variable setup.

The MLR approach underperformed when compared the RF-MMC approach in all setups. Additionally, the model performed worse than the average discharge in all three configurations. The *All_dis* variable setup resulted in a 63% loss in performance for *all_stations* setup and a 71% decrease for *rhine_only* setup. The setup *all_dis_meteo* reduced performance by 30% for

all_stations and 40% for *rhine_only* setup. Moreover, using *allpredictors* setting resulted in a 17% loss in performance in *all_stations* and a 21% decrease in *rhine_only* performance. The performance of the MLR when tested on a new river basin were significantly worse than all other models, with *rhine_elbe* setup having a median KGE of -0.62 which is more than 200% decrease from the average discharge.

Table 6: RF results: Percentage KGE improvement over average discharge, setup All stations and Rhine only

Setup	Median KGE All stations	KGE performance All stations	Median KGE rhine only	KGE performance rhine only
average	0.58	-	0.64	-
all_dis	0.22	-63%	0.19	-71%
all_dis_meteo	0.42	-30%	0.39	-40%
allpredictors	0.49	-17%	0.51	-21%

Table 7: RF results: Percentage KGE improvement over average discharge, MLR, setup Rhine Elbe and Rhine Maas

Setup	Median KGE Elbe	KGE performance All stations	Median KGE rhine only	KGE performance rhine only
average	0.59	-	0.52	-
all_dis	-0.46	178%	0.15	-72%
all_dis_meteo	-0.59	199%	0.28	-47%
allpredictors	-0.62	205%	0.38	-27%

Summary of the coefficients generated from the MLR model for *all_stations* setup is presented in Appendix C. The results of the coefficients demonstrates that the three discharge variables, namely *wg3_dis, lis_dis, and pcr_dis*, have the most significant impact among the variables investigated. Surprisingly, the coefficient associated with *wg3_dis* and *pcr_dis* were found to be negative, contradicting the observed positive association between simulated discharge and observed discharge.

Based on the large difference in KGE values between the MLR and the standard GHMs, as well as the unexpected negative coefficient sign of discharge variables and the large correlation between the variables, a VIF test was performed to explore whether there was a violation of the multicollinearity assumption. Appendix B shows the results of VIF along with the correlation analysis. The results show that the collinearity assumption is violated by the MLR model. The VIF scores reveals the presence of collinearity issues among the model's predictor variables.

4.3 Benchmarking RF results

In this section we present the results of the comparative analysis between the RF-MMC (*allpredictors*) and the alternative hybrid post-processing strategies proposed by Magni et al. (2023) (*PCR-allpredictors*). Figure 10 depicts the outcomes of these two models. The figure shows the Cumulative Distribution Function (CDF) of KGE values in the two RF predictions. The results show that the MMC setup, *allpredictors,* outperforms the hybrid post-processing strategy across most of the predictions. It is worth noting that the *PCR_allpredictors* model catches up to the MMC-based models at around KGE value of 0.87 and subsequently outperforms the MMC-based model in terms of performance. The hydrograph plot depicted in Appendix E demonstrates that the MMC method outperforms the PCR-allpredictors dataset, especially at low discharge volumes.

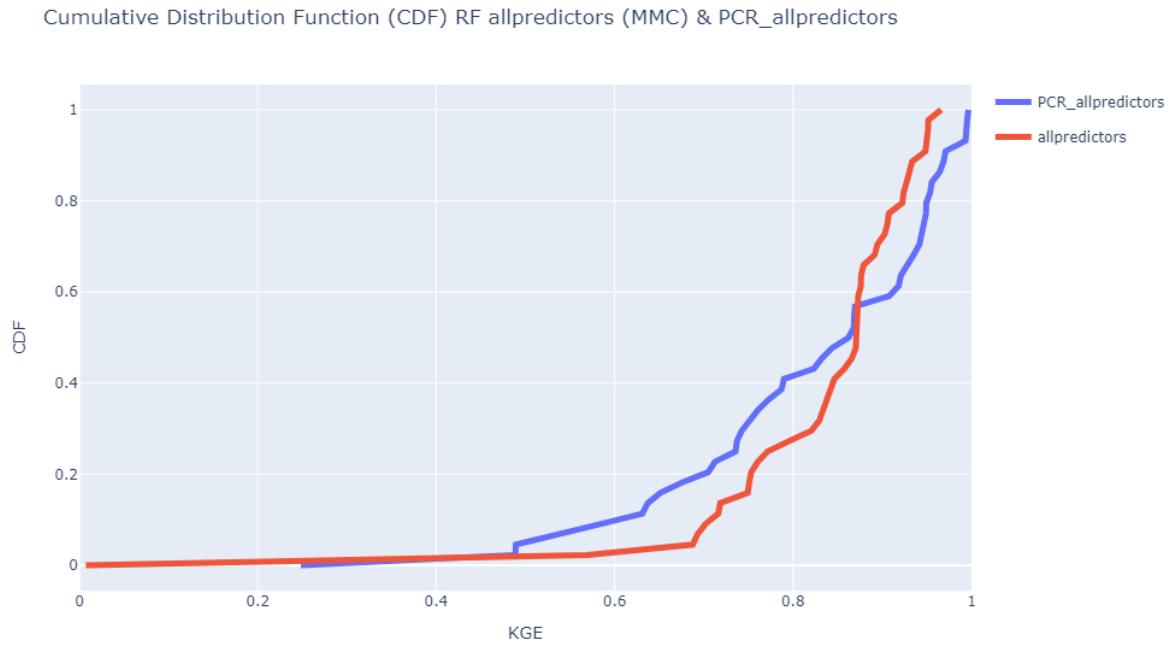

Figure 10Cumulative Distribution Function (CDF) of KGE values of RF-MMC (allpredictors) and the alternative hybrid post-processing strategies (PCR_allpredictors). The blue line depicts the PCR_allpredictors setup and the red line shows the allpredictors setup.

Besides the above comparison, we also investigate weather catchment attributes could enhance the generalization ability of the RF-MMC approach. The results of this analysis can be found in Appendix F For *rhine_elbe* setup, *allpredictors_catch* improved 100% of all bad performing stations under the MMC *allpredictors* setup (KGE < -0.41). However, *allpredictors_catch* did not have the same success with the *rhine_maas* setup. This is mainly because Elbe basin is more similar the Rhine basin than the Maas basin in terms of catchment characteristics. Due to this, the model was able to transfer the patterns learned from the Rhine basin to the Elbe, but not the Maas basin.

5 Discussion and conclusion

This study aimed to evaluate the performance of using a non-linear machine learning model, specifically Random Forest (RF), as a Multi-Model Combination (MMC). The outputs of three Global Hydrological Models (GHM) have been combined, namely: PCR-GLOBWB, LISFLOOD and WaterGAP3 and RF was applied as a MMC solution. The main research question was: How can a machine-learning-powered multi-model combination approach improve streamflow predictions by combining outputs from different global hydrological models?

We evaluated the efficacy of the RF-MMC method in improving streamflow forecasts relative to the discharge of individual global hydrological models. Using MLR, the performance of the nonlinear RF-MMC approach was compared to that of a linear MMC. In addition, the generalizability of the RF-MMC method when tested on a river basin which was not included in the training set. Moreover, we compared the performance of the RF-MMC method with that of the hybrid post-processing method devised by Magni et al. (2023). Finally, the effect of incorporating static PCR-GLOBWB attributes on the generalization ability of the RF-MMC approach for streamflow prediction was investigated.

The results of RF-MMC demonstrated a clear improvement over the individual GHMs in the setup where all available stations were used in the cross-validation setting. Similarly, when trained and test on a single river basin, the RF-MMC worked remarkably better than the individual GHMs. The best performing variable setup was the *allpredictors*, which achieved a median KGE of 0.81 and 0.87 for *all_stations* and *rhine_only* setups, corresponding to a 39% and 34% improvement over the average GHMs discharge, respectively. These findings confirm the benefits of applying RF as a non-linear multi-model combination of the outputs of global hydrological models for streamflow prediction. This finding validates the premise of Ajami et al. (2006) who argued that, combining uncalibrated multiple simulations predictions is more effective than relying on the best-calibrated individual model. The authors compared several MMC approaches that are used for streamflow forecasting. Similar results were also found by Zaherpour et al. (2019) who investigated the application of Gene Expression Programming (GEP) as a non-linear MMC solution and compared it with the individual GHMs and their mean. The authors found that the non-linear multi-model combination has considerable performance gain when compared to the individual GHMs and their ensembled mean, further supporting the findings of this study.

Despite the promising performance of RF-MMC approach, it underperforms when its generalization potential is tested on a different river basin. This method showed limitations when extrapolating to stations whose catchments were not in the training set, resulting in poor performance. However, because the RF-MMC model in this validation setup was trained only on the Rhine basin, the dataset may be weak in diversity, resulting in a limited representation of different catchment characteristics. This shortcoming can be addressed by increasing the dataset to encompass a diverse variety of catchments. As a result, the model may be able to capture overall discharge patterns across multiple catchments rather than focusing primarily on unique noise found in a single river basin.

According to the findings, the MLR-MMC approach underperformed across all the validation setups when compared to the RF-MMC approach. This is likely due the violation of the collinearity assumption in the model. Multicollinearity in regression analysis decreases

prediction stability and raises coefficient standard errors, resulting in less accurate and unreliable forecasts (Chan et al., 2022). When *all_stations* and *allpredictors* setup are evaluated, the VIF results of the MLR model are alarming. The discharge factors have a VIF greater than 5, which is cause for worry according to Menard (2002). Furthermore, the VIF of the snow water equivalent variables from PCR-GLOBWB and WaterGAP3 are both greater than 70, indicating a significant collinearity issue. It is notable that collinearity evaluation was conducted after MLR model application, which is a limitation of the study because it should ideally precede model application. To adequately address the question of whether GHM outputs can be extrapolated using MLR, the issue of collinearity must be addressed. In the future, efforts could be made to investigate approaches such as ridge regression (Hoerl, 1962), to increase the stability and dependability of MLR predictions. Similarly, another possible area for investigation is to transform the features using Principal Component Analysis (PCA) as proposed by Lafi & Kaneene (1992) to mitigate the collinearity effect prior to testing MLR.

Furthermore, when the RF-MMC was benchmarked against the hybrid method, *PCR_allpredictors*, developed by Magni et al. (2023), the RF-MMC had better KGE performance for predictions up to KGE of 0.87. This further demonstrates the potential of hybrid streamflow modelling using machine learning and multi-model combination approach. As an additional benchmark, we investigated whether incorporating catchment attributes from PCR-GLOBWB could enhance the generalization ability of the RF-MMC approach. We found that combining the outputs of global hydrological models with catchment attributes enhanced the ability of the model to generalize to a river basin that was not included in the training set. However, this is conditional on the inclusion of training datasets that have catchment characteristics similar to those of the validation dataset. Based on these findings, future research could consider incorporating HydroBASINS (Lehner & Grill, 2013), a global high-resolution data set that offers detailed data about river basins and their characteristics, as a dataset to improve the framework with additional catchment attributes.

The multiple-model combination strategy utilized in this study has some limitations. First, it only includes a subset of outputs from three GHMs, which does not represent the complete range of GHMs and their outputs. To enhance the approach, it is necessary to investigate the inclusion of additional variables such as groundwater storage and evapotranspiration. Expanding the scope of process-based GHMs and integrating a broader range of variables would improve the approach's predictive ability and generalizability. Secondly, the varying calibration status of the models had a direct effect on the predictive performance of each model. To accurately assess the variable importance and predictive ability induvial models, it is recommended that all models included in the analysis be calibrated or uncalibrated. Lastly, only one non-linear MMC approach (i.e., Random Forest (RF)) has been applied. It is important to investigate the effectiveness of other ML models such as XGBoost (T. Chen & Guestrin, 2016) as MMC approach.

In conclusion, this research showed the potential of generating accurate streamflow predictions when machine learning techniques, specifically RF is employed to combine outputs from different global hydrological models. We observed that the RF-MMC outperformed individual GHM for streamflow predictions. The strategy proposed was also compared to the established hybrid approach developed by (Magni et al., 2023) and showed better KGE performance for most of the predictions. Additionally, including catchment characteristics led to a better generalization of the RF-MMC approach. However, more research is needed to further improve the non-linear MMC approach for more reliable streamflow predictions.

Code availability
The source codes used in this study can be found at this repository on GitHub.

Appendix

Appendix A: RF hyperparameter tuning

Figure A 1 shows the results of the RF hyperparameter tunning for *all_stations* setup

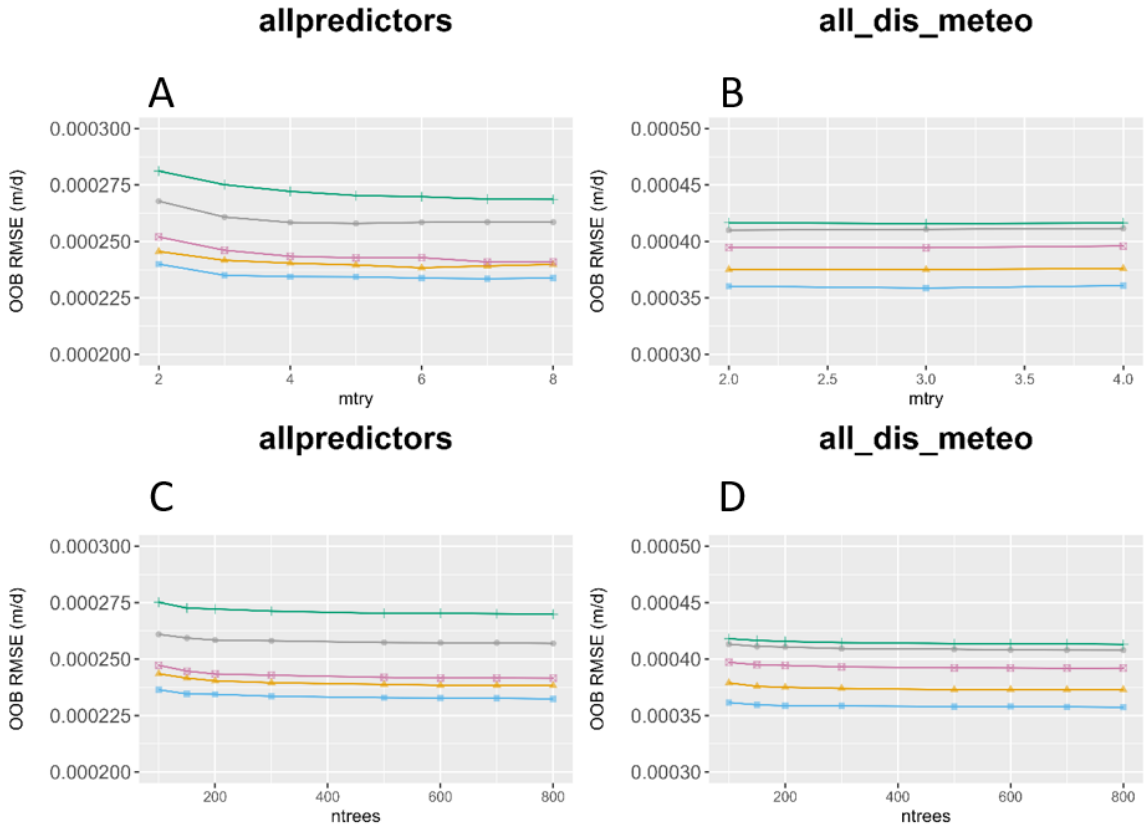

Figure A 1: RF hyperparameter tuning. (a-c) Tuning of mtry, using a fixed ntree of 200. (d-f) Tuning of ntrees, using the optimal mtry calculated in the previous step.

Appendix B: Analysis of Correlation and Variable Inflation Factor (VIF)

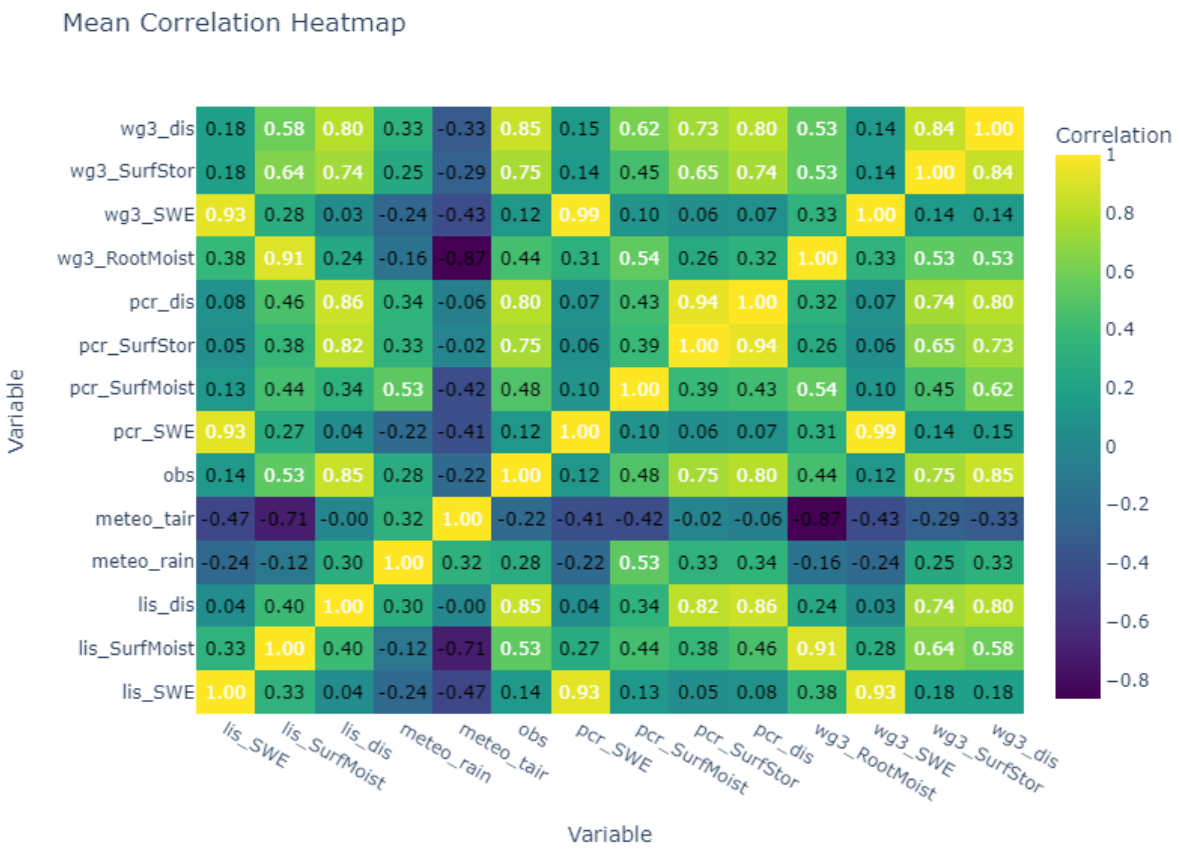

Figure B 1 mean correlation heatmap, Allpredictors

Figure B 2 shows the VIF test results. The red bars reflect VIF values greater than 10, indicating strong multicollinearity, whereas the dotted line represents a VIF value greater than 5.

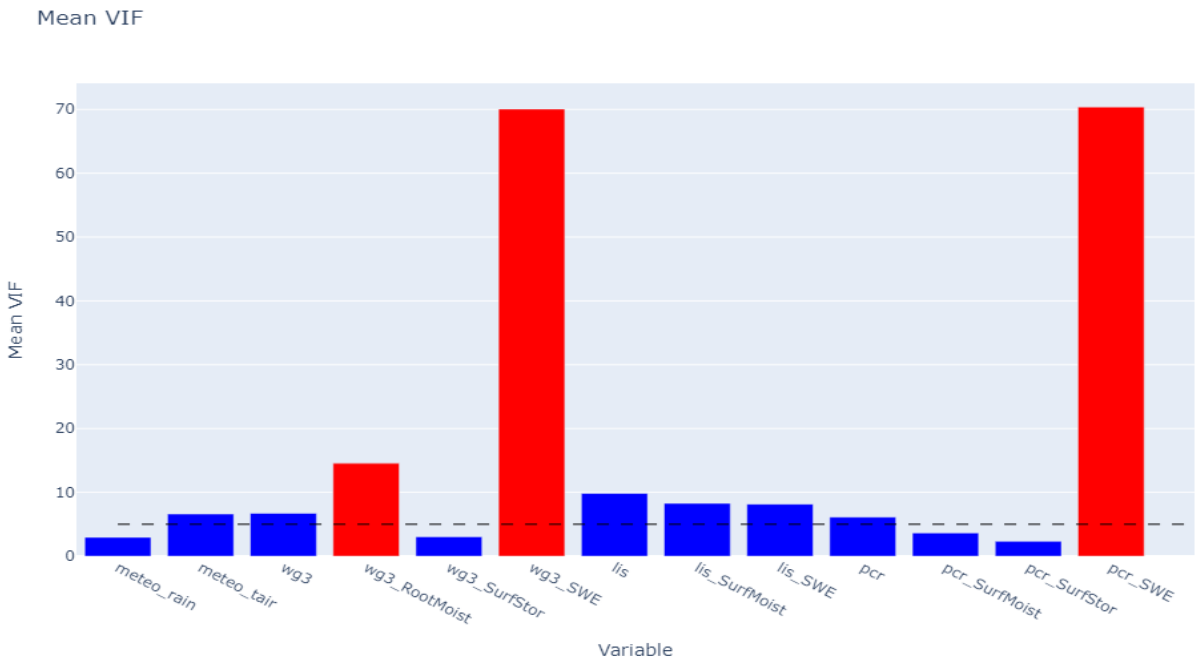

Figure B 2 Mean VIF Allpredictors

Appendix C: MLR Coefficients

Table 1 Coefficients MLR all_dis setup.

Variables	Beta	Lower CI	Upper CI	P-value
(Intercept)	6.73E-04	6.56E-04	6.91E-04	< 0.005
lis	0.42	4.06E-01	4.41E-01	< 0.005
pcr	-0.03	-4.11E-02	-1.58E-02	< 0.005
wg3	-0.02	-3.69E-02	-2.49E-03	0.03

Table 2 Coefficients MLR all_dis_meteo setup.

Variables	Beta	Lower CI	Upper CI	P-value
(Intercept)	7.25E-04	7.09E-04	7.41E-04	< 0.005
lis	0.50	4.84E-01	5.19E-01	< 0.005
pcr	-0.01	-2.35E-02	5.40E-04	0.20
wg3	-0.16	-1.80E-01	-1.44E-01	< 0.005
meteo_rain	1.77E-04	1.64E-04	1.89E-04	< 0.005
meteo_tair	-2.52E-04	-2.65E-04	-2.38E-04	< 0.005

Table 3 Coefficients MLR all_dis setup.

Variables	Beta	Lower CI	Upper CI	P-value
(Intercept)	7.75E-04	7.59E-04	7.91E-04	< 0.005
lis	0.42	4.01E-01	4.36E-01	< 0.005
pcr	0.03	1.62E-02	4.00E-02	< 0.005
wg3	-0.16	-1.73E-01	-1.38E-01	< 0.005
wg3_SurfStor	1.5E-04	1.36E-04	1.74E-04	< 0.005
pcr_SWE	1.0E-04	1.13E-05	1.93E-04	0.05
meteo_rain	9.1E-05	7.25E-05	1.10E-04	< 0.005
pcr_SurfStor	8.0E-05	6.36E-05	9.67E-05	< 0.005
lis_SWE	4.1E-05	1.05E-05	7.25E-05	0.012
wg3_RootMoist	2.5E-05	-1.65E-05	6.64E-05	0.27
lis_SurfMoist	8.9E-06	-2.24E-05	4.02E-05	0.35
pcr_SurfMoist	6.8E-06	-1.40E-05	2.75E-05	0.38
wg3_SWE	-1.4E-04	-2.27E-04	-4.48E-05	0.009
meteo_tair	-1.4E-04	-1.64E-04	-1.08E-04	< 0.005

Appendix D: Results Multiple linear regression

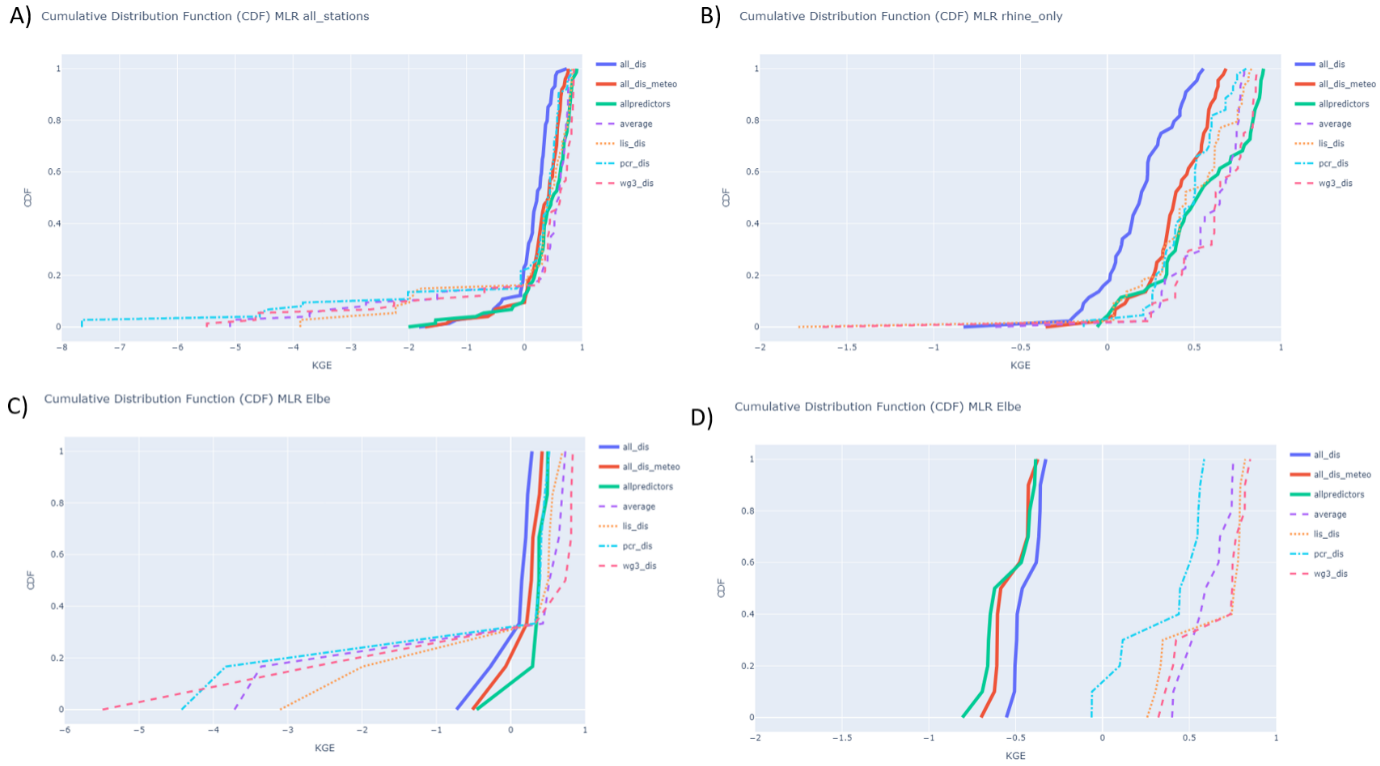

Figure D 1: Results MLR, Cumulative Distribution Function (CDF) of the KGE values. Panels A and B display the results of the All-stations and Rhine_only setup, while Panels C and D show the results of Rhine-Maas and Rhine Elbe respectively. Each Figure has its own margin that corresponds to its minimum KGE value.

Appendix E: Comparison of RF-MMC approach with PCR-Allpredictors

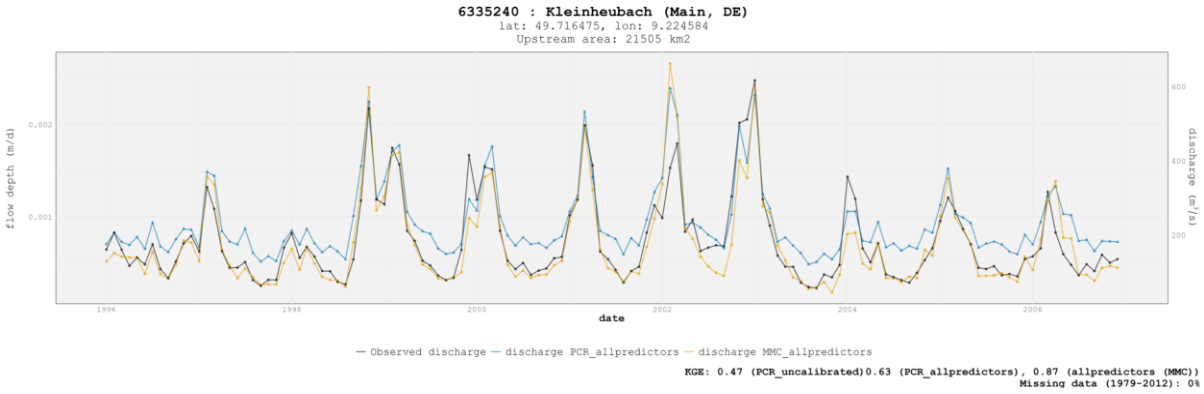

Figure E 1: Flow depth and discharge for the Kleinheubach station in Germany for MMC predictions (MMC_allpredictors) and benchmark model (PCR_allpredictors)

Appendix F: Generalization ability of RF-MMC approach with catchment attributes

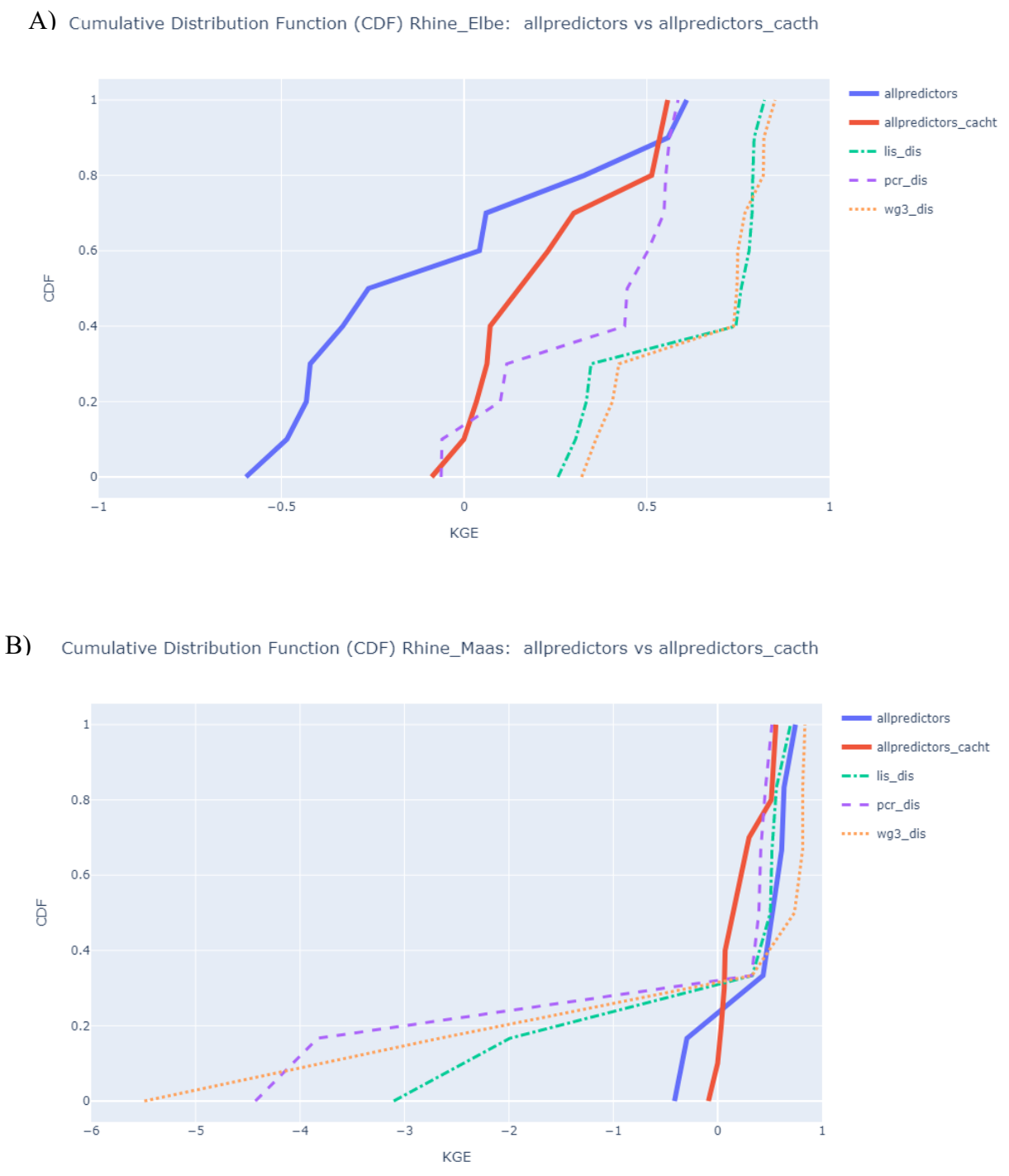

Figure F 1: Cumulative Distribution Function (CDF) of KGE values of RF-MMC (allpredictors) and the RF-MMC with catchment attributes from PCR-GLOBWB (allpredictors_catch). The bold lines show the predicted discharge of these two setups, and the dashed lines indicate the discharge of the individual GHMs. Pane A, B show the results of rhine_elbe and rhine_maas setup respectively. Each Figure has its own margin based on its minimum KGE value.

Bibliography

Ajami, N. K., Duan, Q., Gao, X., & Sorooshian, S. (2006). Multimodel Combination Techniques for Analysis of Hydrological Simulations: Application to Distributed Model Intercomparison Project Results. *Journal of Hydrometeorology*, *7*(4), 755–768. https://doi.org/https://doi.org/10.1175/JHM519.1

Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Dutra, E., Fink, G., Orth, R., & Schellekens, J. (2017). Global evaluation of runoff from 10 state-of-the-art hydrological models. *Hydrology and Earth System Sciences*, *21*(6), 2881–2903. https://doi.org/10.5194/hess-21-2881-2017

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chan, J. Y.-L., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z.-W., & Chen, Y.-L. (2022). Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. *Mathematics*, *10*(8). https://doi.org/10.3390/math10081283

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. 785–794. https://doi.org/10.1145/2939672.2939785

Chen, Y., Ren, Q., Huang, F., Xu, H., & Cluckie, I. (2011). Liuxihe Model and Its Modeling to River Basin Flood. *Journal of Hydrologic Engineering*, *16*(1), 33–50. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000286

de Wit, M. J. M., van den Hurk, B., Warmerdam, P. M. M., Torfs, P. J. J. F., Roulin, E., & van Deursen, W. P. A. (2007). Impact of climate change on low-flows in the river Meuse. *Climatic Change*, *82*(3–4), 351–372. https://doi.org/10.1007/s10584-006-9195-2

Eisner, S. (2016). *Comprehensive evaluation of the WaterGAP3 model across climatic, physiographic, and anthropogenic gradients*.

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

Hesse, C. (2018). *Integrated water quality modelling in meso- to large-scale catchments of the Elbe river basin under climate and land use change* [Universität Potsdam]. https://doi.org/10.25932/publishup-42295

Hoerl, A. E. (1962). Application of ridge analysis to regression problems. *Chemical Engineering Progress*, *58*, 54–59. https://cir.nii.ac.jp/crid/1571698600104425088

Jia, Y., Ni, G., Kawahara, Y., & Suetsugi, T. (2001). Development of WEP model and its application to an urban watershed. *Hydrological Processes*, *15*(11), 2175–2194. https://doi.org/10.1002/hyp.275

Karssenberg, D., Schmitz, O., Salamon, P., de Jong, K., & Bierkens, M. F. P. (2010). A software framework for construction of process-based stochastic spatio-temporal models and data assimilation. *Environmental Modelling & Software*, *25*(4), 489–502. https://doi.org/https://doi.org/10.1016/j.envsoft.2009.10.004

Khanal, S., Lutz, A. F., Immerzeel, W. W., Vries, H. de, Wanders, N., & Hurk, B. van den. (2019). The Impact of Meteorological and Hydrological Memory on Compound Peak Flows in the Rhine River Basin. *Atmosphere*, *10*(4). https://doi.org/10.3390/atmos10040171

Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, *23*(10), 4323–4331. https://doi.org/10.5194/hess-23-4323-2019

Lafi, S. Q., & Kaneene, J. B. (1992). An explanation of the use of principal-components analysis to detect and correct for multicollinearity. *Preventive Veterinary Medicine*, *13*(4), 261–275. https://doi.org/https://doi.org/10.1016/0167-5877(92)90041-D

Lehner, B., & Grill, G. (2013). Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. *Hydrological Processes*, *27*(15), 2171–2186. www.hydrosheds.org

Liu, Y., Sang, Y.-F., Li, X., Hu, J., & Liang, K. (2016). Long-Term Streamflow Forecasting Based on Relevance Vector Machine Model. *Water*, *9*(1), 9. https://doi.org/10.3390/w9010009

Magni, M., Sutanudjaja, E., Shen, Y., & Karssenberg, D. (2023). Global streamflow modelling using process-informed machine learning. *EGU General Assembly 2023*. https://doi.org/10.5194/egusphere-egu23-1160

Marriott, F. H. C., Neter, J., Wasserman, W., & Kutner, M. H. (1985). Applied Linear Regression Models. *Biometrics*, *41*(2), 596. https://doi.org/10.2307/2530893

Menard, S. (2002). *Applied Logistic Regression Analysis*. SAGE Publications, Inc. https://doi.org/10.4135/9781412983433

Milly, P. C. D., Dunne, K. A., & Vecchia, A. V. (2005). Global pattern of trends in streamflow and water availability in a changing climate. *Nature*, *438*(7066), 347–350. https://doi.org/10.1038/nature04312

Mohammadi, B., Moazenzadeh, R., Christian, K., & Duan, Z. (2021). Improving streamflow simulation by combining hydrological process-driven and artificial intelligence-based models. *Environmental Science and Pollution Research*, *28*(46), 65752–65768. https://doi.org/10.1007/s11356-021-15563-1

Schellekens, J., Dutra, E., 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

Shamseldin, A. Y., O'Connor, K. M., & Liang, G. C. (1997). Methods for combining the outputs of different rainfall–runoff models. *Journal of Hydrology*, *197*(1–4), 203–229. https://doi.org/10.1016/S0022-1694(96)03259-3

Sharma, P., & Machiwal, D. (2021). Streamflow forecasting. In *Advances in Streamflow Forecasting* (pp. 1–50). Elsevier. https://doi.org/10.1016/B978-0-12-820673-7.00013-5

Shen, Y., Ruijsch, J., Lu, M., Sutanudjaja, E. H., & Karssenberg, D. (2022). Random forests-based error-correction of streamflow from a large-scale hydrological model: Using model state variables to estimate error terms. *Computers & Geosciences*, *159*, 105019. https://doi.org/10.1016/j.cageo.2021.105019

Sutanudjaja, E. H., van Beek, R., Wanders, N., Wada, Y., Bosmans, J. H. C., Drost, N., van der Ent, R. J., de Graaf, I. E. M., Hoch, J. M., de Jong, K., Karssenberg, D., López López, P., Peßenteiner, S., Schmitz, O., Straatsma, M. W., Vannametee, E., Wisser, D., & Bierkens, M. F. P. (2018). PCR-GLOBWB 2: a 5 arcmin global hydrological and water resources model. *Geoscientific Model Development*, *11*(6), 2429–2453. https://doi.org/10.5194/gmd-11-2429-2018

Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J., & Viterbo, P. (2014). The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data. *Water Resources Research*, *50*(9), 7505–7514. https://doi.org/10.1002/2014WR015638

Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. https://doi.org/10.18637/jss.v077.i01

Xiong, L., Shamseldin, A. Y., & O'Connor, K. M. (2001). A non-linear combination of the forecasts of rainfall-runoff models by the first-order Takagi–Sugeno fuzzy system. *Journal of Hydrology*, *245*(1–4), 196–217. https://doi.org/10.1016/S0022-1694(01)00349-3

Yang, S., Yang, D., Chen, J., Santisirisomboon, J., Lu, W., & Zhao, B. (2020). A physical process and machine learning combined hydrological model for daily streamflow simulations of large watersheds with limited observation data. *Journal of Hydrology*, *590*, 125206. https://doi.org/10.1016/j.jhydrol.2020.125206

Yaseen, Z. M., Ebtehaj, I., Bonakdari, H., Deo, R. C., Danandeh Mehr, A., Mohtar, W. H. M. W., Diop, L., El-shafie, A., & Singh, V. P. (2017). Novel approach for streamflow forecasting using a hybrid ANFIS-FFA model. *Journal of Hydrology*, *554*, 263–276. https://doi.org/10.1016/j.jhydrol.2017.09.007

Zaherpour, J., Mount, N., Gosling, S. N., Dankers, R., Eisner, S., Gerten, D., Liu, X., Masaki, Y., Müller Schmied, H., Tang, Q., & Wada, Y. (2019). Exploring the value of machine learning for weighted multi-model combination of an ensemble of global hydrological models. *Environmental Modelling & Software*, *114*, 112–128. https://doi.org/10.1016/j.envsoft.2019.01.003