AGU ADVANCING EARTH AND SPACE SCIENCES

Water Resources Research

RESEARCH ARTICLE
10.1029/2023WR036418

Predictor Importance for Hydrological Fluxes of Global Hydrological and Land Surface Models

João Paulo L. F. Brêda[1], Lieke A. Melsen[1], Ioannis Athanasiadis[2], Albert Van Dijk[3], Vinícius A. Siqueira[4], Anne Verhoef[5], Yijian Zeng[6], and Martine van der Ploeg[1]

[1]Hydrology and Quantitative Water Management, Wageningen University & Research, Wageningen, Netherlands, [2]Data Competence Centre, Wageningen University & Research, Wageningen, Netherlands, [3]The Fenner School of Environment & Society, Australia National University, Canberra, ACT, Australia, [4]Instituto de Pesquisas Hidráulicas, Universidade Federal do Rio Grande do Sul, Farroupilha, Brazil, [5]Department of Geography and Environmental Science, University of Reading, Reading, England, [6]Department of Water Resources, University of Twente, Enschede, Netherlands

Abstract Global Hydrological and Land Surface Models (GHM/LSMs) embody numerous interacting predictors and equations, complicating the understanding of primary hydrological relationships. We propose a model diagnostic approach based on Random Forest (RF) feature importance to detect the input variables that most influence simulated hydrological fluxes. We analyzed the JULES, ORCHIDEE, HTESSEL, SURFEX, and PCR-GLOBWB models for the relative importance of precipitation, climate, soil, land cover and topographic slope as predictors of simulated average evaporation, runoff, and surface and subsurface runoff. RF models functioned as a metamodel and could reproduce GHM/LSMs outputs with a coefficient of determination (R^2) over 0.85 in all cases and often considerably better. The GHM/LSMs agreed that precipitation, climate and land cover share equal importance for evaporation prediction, and mean precipitation is the most important predictor of runoff, while topographic slope and soil texture have no influence on the total variance of the water balance. However, the GHM/LSMs disagreed on which features determine surface and subsurface runoff processes, especially with regard to the relative importance of soil texture and topographic slope. Finally, the selection of soil maps was only important for target variables of which soil is a relevant predictor. We conclude that estimating feature importance is a useful diagnostic approach for model intercomparison projects.

Plain Language Summary Simulations of hydrological fluxes such as evaporation and runoff at a global scale are uncertain. This happens because the models that produce global simulations are different in terms of structure, parametrization and meteorological data. So, several model intercomparison projects (MIP) have tried to identify where the hydrological fluxes estimates are most discrepant. In order to make MIPs even more useful, we are proposing an additional method focusing on understanding why the models disagree. This method consists of replacing the original global model with a random forest model and then identifying which input variables are more relevant using the feature importance functionality. More specifically, we detected how important meteorological variables, soil properties, land cover and topography are for each global model. We observed that the models agree that precipitation, climate and land cover are equally important for evaporation and that precipitation is the most important feature for estimating runoff. When partitioning runoff into quick and slow flow, we observed that the models disagree on the importance of features, especially topographic slope and soil.

Key Points:

- Detecting the predictors importance can be an additional approach for Model Intercomparison Projects
- Global models agree about the features importance for water balance components but disagree for surface and subsurface runoff
- Selecting the soil database only matters when soil is a relevant predictor, which is not the case for all models and target variables

Supporting Information:

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

Correspondence to:

J. P. L. F. Brêda,
joaopaulolfb@gmail.com

Citation:

Brêda, J. P. L. F., Melsen, L. A., Athanasiadis, I., Van Dijk, A., Siqueira, V. A., Verhoef, A., et al. (2024). Predictor importance for hydrological fluxes of Global Hydrological and Land Surface Models. *Water Resources Research*, *60*, e2023WR036418. https://doi.org/10.1029/2023WR036418

Received 4 OCT 2023
Accepted 13 AUG 2024

1. Introduction

Global Hydrological Models (GHMs) and Land Surface Models (LSMs) embody the current state of knowledge in simulating the water cycle on land and its interactions with the atmosphere (Döll et al., 2016; Fisher & Koven, 2020). LSMs are often coupled with atmospheric and ocean models for numerical weather predictions (Pappenberger et al., 2010; Zhang et al., 2011) and climate projections (Collins et al., 2011; Dufresne et al., 2013), thus acting as key components in providing short-to long-term forecasts, as well as reanalysis (Hersbach et al., 2020). In addition, GHMs characterize the global water balance, quantifying the amount of freshwater that reaches the oceans, the anomalies of groundwater storage and anthropogenic water use (E. A. Clark et al., 2015; Müller Schmied et al., 2021).

However, global simulations present significant uncertainties. Global models oversimplify the hydrological cycle by reducing a complex environmental system to a limited set of equations calculated over a grid that has a

horizontal spatial resolution in the order of kilometers (10–100) (Bierkens et al., 2015; M. P. Clark et al., 2015). In addition, the uncertainty related to input parameters and driving data propagates to the model results. Consequently, different models frequently provide diverging or even conflicting predictions. Climate change impact assessments show that the GHM/LSMs model selection is a major source of uncertainty for projections of evaporation (Hagemann et al., 2013) and low discharge (Giuntoli et al., 2015; Krysanova et al., 2017), and that the ensemble spread of GHM/LSMs is considerably larger than the ensemble spread of catchment hydrological models for discharge (Gosling et al., 2017).

Since the 90's, Model Intercomparison Projects (MIPs) have been proposed to evaluate LSMs (Henderson-Sellers et al., 1993) usually by comparing model outputs to an observation database (Best et al., 2015). Throughout the years, MIPs have contributed to improved closure of the water and energy balance, and to improving soil wetness for climate predictions (Dirmeyer, 2011; van den Hurk et al., 2011). Recent MIPs have identified reduced performance of GHM/LSMs in snow and tropical regions (Giuntoli et al., 2015; Haddeland et al., 2011; Schellekens et al., 2017) and a general overestimation of runoff from GHMs (Beck, Van Dijk, De Roo, et al., 2017; Zaherpour et al., 2018). As such, conventional modeling comparisons have shown to be valuable approaches for identifying modeling weaknesses. However, it is complicated to address the detected issues when there is a limited understanding of the multitude of processes and variables' interactions within a GHM/LSM.

In that sense, a more in-depth analysis should be considered in addition to the conventional assessment of matching simulated outputs to observations (De Boer-Euser et al., 2017; Gleeson et al., 2021). For example, Wagener et al. (2022) suggested using global sensitivity analysis to evaluate models in terms of their ability to represent dominant processes and input-output responses. Similarly, M. P. Clark et al. (2011) discuss their concerns about the abundance of hydrological models and argue that models' scrutiny should be focused on their adequacy to multiple hypotheses in order to achieve greater physical representation. Following this path, Gnann et al. (2023) proposed an alternative approach for model intercomparisons that shifts from the fitness to historical observations to a process-oriented evaluation. The authors assessed the accuracy of GHM/LSMs in representing functional hydrological relationships, such as the Budyko framework and elasticities to streamflow (Chiew, 2006), and pointed out that the energy balance was poorly represented. In an effort to provide a better understanding of GHM/LSMs, Telteu et al. (2021) described and compared the hydrological structure of 16 models, from the number of water storage compartments to the human water use sectors. Their approach was limited due to the models' complexity and the authors recommended extended assessments to MIPs such as workshops for multi-model parameterization experiments. Thus, diagnostic evaluation approaches are currently being sought for advances in global modeling, providing more informative avenues for model improvements.

Progressively, data-driven techniques have been assuming a leading role in hydrological modeling (Nearing et al., 2021). Machine learning (ML) models have already been successful in predicting surface water and groundwater stores and flows at the catchment level (Shen, 2018; Zounemat-Kermani et al., 2021) and at global scales within a hybrid hydrological model (Kraft et al., 2022). Besides its primary purpose of prediction, ML models can provide important statistical information and process understanding (McGovern et al., 2019). Specifically, detecting features' importance is a secondary outcome that can indicate the most relevant input features of an ML model (Hastie et al., 2009). In the hydrological field, the ML input features are equivalent to predictors (Beck et al., 2015), attributes (Kratzert et al., 2019), and variables (Li et al., 2021), while feature importance has also been termed variable ranking (Laimighofer et al., 2022; Li et al., 2021). Since the early work of Beck et al. (2015), studies have used ML to identify the most important predictors for hydrological signatures (Addor et al., 2018), time series of discharge (Kratzert et al., 2019), flooding (Schmidt et al., 2020; Stein et al., 2021) and streamflow trends (Zeng et al., 2021).

Given the complexity of GHM/LSMs, the use of surrogate models (Razavi et al., 2012) might become an easier path for deciphering the global models' behavior. ML models have been successful in replicating GHM/LSMs (Gu et al., 2020; Sun et al., 2023) and favoring model interpretability (Cappelli et al., 2022; Wang et al., 2022). In particular, Random Forests are suitable alternatives for an explainable ML model (De la Fuente et al., 2023; Stein et al., 2021), due to the practicality and diversity of approaches for estimating feature importance (Huang et al., 2023). Antoniadis et al. (2021) concluded that, whenever RF can be used as a surrogate model, feature importance becomes an efficient alternative for global sensitivity analysis. In this paper, we are proposing to use RF as a surrogate model for GHM/LSMs and estimate the respective features' importance as a new diagnostic

approach for global model comparison. This analysis can indicate which variables and processes are being overlooked by the GHM/LSMs and provide guidance on further model development.

2. Methodology

The methodology is summarized as follows: (a) We selected five GHM/LSMs from the Earth2Observe (Schellekens et al., 2017) project for intercomparison. Time-dependent variables were averaged to create static maps. Thus, each grid cell in the global domain contains both input features and outputs of a given GHM/LSM. (b) We fed this information to a RF model, which would act as a surrogate model. The surrogate models were trained to reproduce average hydrological fluxes of the GHM/LSMs, namely evaporation, runoff, and surface and subsurface runoff. To enable comparison between the GHM/LSMs, the input features were grouped into climate, precipitation, soil, land cover and topographic slope. (c) We evaluated the fitness of the surrogate models to the original outputs of the GHM/LSMs. Finally, we estimated the importance of the feature groups for each GHM/LSM and analyzed how they differed.

In the following sections, we describe each step in more detail.

2.1. E2O Models Selection

Earth2Observe—E2O (Schellekens et al., 2017) was a European Union-funded project to integrate different Earth Observations techniques and obtain an extensive re-analysis of global water resources. The project legacy provides an organized data set with a common spatial-temporal resolution that facilitates comparisons and evaluations. We specifically used the Tier-2 data set from the E2O project consisting of 8 GHM/LSMs simulated using the same forcing data. For this study, we selected the GHM/LSMs that were not regionally calibrated (according to the model description) so that the ML model could capture the response of global features without spatial biases. The selected global models are JULES (Walters et al., 2014), ORCHIDEE (Krinner et al., 2005), HTESSEL (Balsamo et al., 2009), SURFEX (Le Moigne, 2018) and PCR-GLOBWB (Van Beek & Bierkens, 2008). How surface hydrology processes were represented by the models during the E2O project is briefly described below:

JULES: The soil column is 3 m deep and is discretized into 4 layers. The hydraulic relationships are determined according to Van Genuchten (1980). Surface runoff generation and the heterogeneity of soil moisture are based on the TOPMODEL (Beven & Kirkby, 1979) which is formulated according to hydrological concepts such as partial area contributions and saturation excess overland flow (Dunne & Black, 1970).

ORCHIDEE: The soil column is 2 m deep and is discretized into 11 layers. The hydraulic relationships are determined according to van Genuchten. Soil infiltration is inspired by the Green-Ampt model, which considers a wet front on the soil like a piston. The precipitated water that is not able to infiltrate during a time step generates surface runoff, therefore, runoff generation is based on an infiltration excess or Hortonian mechanism.

HTESSEL: The soil column is 3 m deep and it is discretized into 9 layers. The hydraulic relationships are determined according to van Genuchten. The surface runoff scheme is similar to the ARNO model (Todini, 1996) which estimates runoff based on a fraction of the catchment that is saturated (saturation excess mechanism).

SURFEX: The soil column is 12 m deep and it is discretized into 14 layers. The hydraulic relationships are determined according to Brooks and Corey (1966). The runoff generation is composed of infiltration excess and saturation excess mechanisms. The saturation-excess runoff is based on the TOPMODEL, while the infiltration-excess runoff depends on the maximum infiltration capacity which is computed via a Green-Ampt approach near the surface (10 cm).

PCR-GLOBWB: The soil is represented by a leak-bucket scheme with two vertically stacked layers and a third layer that accounts for the groundwater. The hydraulic relationships guide the vertical water exchanges between the layers and are based on Clapp and Hornberger (1978). The surface runoff scheme is similar to the ARNO model, thus a saturation excess mechanism.

In Tier-2, both forcing and model spatial resolution is 0.25°, with daily and monthly data available from 1980 to 2014. We downloaded GHM/LSMs monthly aggregated outputs and the respective meteorological data used to

force the models. The precipitation data was from MSWEP (Beck, Van Dijk, Levizzani, et al., 2017) and the remaining meteorological data was from the ERA-Interim data set (Dee et al., 2011).

2.2. Input and Output Data

In this study, both inputs and outputs correspond to static variables, most commonly long-term average values. The hydrological fluxes (outputs) we analyzed are long-term mean evaporation (*Evap*), runoff (*Q*), surface runoff (*Qs*), and subsurface runoff (*Qsb*) obtained from the E2O data sets. The E2O project defines *Q* as the average total liquid water draining from land (in a grid cell), *Qs* as the runoff from the land surface and/or subsurface stormflow, and *Qsb* as the gravity drainage and/or slow lateral response. We also calculated the long-term mean of the following meteorological variables: wind speed, temperature, specific humidity, air pressure at the surface, incident shortwave radiation, incident longwave radiation and precipitation. Because of its expected importance, we consider precipitation separately from the other meteorological variables, which we together term climate features. Our data domain is the common simulation domain among the GHM/LSMs, corresponding to 226,654 grid cells.

In addition to precipitation and climate features, there are input features that contribute to the spatial parametrization of a global model, such as soil properties, land cover and topographic slope. These input features were not provided by the E2O project, but were mentioned in the E2O report to some extent (Dutra et al., 2017). Therefore, we retrieved specific data sets used by each GHM/LSM individually. Since the E2O report was not conclusive on the employed parameter data sets used by the different GHM/LSMs, we had to search in published papers and contact modelers of the E2O project for confirmation. The land cover and soil properties features selected for this study are summarized in Table S1 in Supporting Information S1.

We used the same topographic slope for every model, as we assumed that the differences between topographic data sets at a global scale would be relatively small. We used 5-min Gridded Global Relief Data ETOPO5 (National Geophysical Data Center, 1993) to obtain a "slope proxy" (m), estimated as the standard deviation of the nine ETOPO5 cells within a 0.25° GHM/LSM cell.

Three land cover data sets and one soil texture data set needed to be resampled to be used as input features for the ML model (S1). We followed a hybrid aggregation method: most dominant class for higher resolutions (<0.025°) and class fractions for lower resolutions (≥0.025°). Due to computational limitations, maps with higher resolution were first upscaled to 0.025° using the mode of the sample (dominant class) after which the classes' fractions were calculated within the model grid resolution (0.25° × 0.25°). Note that the GHM/LSMs had their own approaches to treat subgrid variability.

2.3. Random Forest and Feature Importance

Random Forest (RF) is essentially an ensemble of decision trees trained with sub-samples of the training data and a subset of the input features (Breiman, 2001). In parameterizing the algorithm, we specified an unlimited maximum tree depth (i.e., the number of leaves and nodes was not limited); only 1/3 of the total input features would be considered in each node splitting; and that the RF would consist of 200 decision trees. The first two hyperparameters are standard (Hastie et al., 2009), while the number of decision trees was large enough to provide a stable performance on every new run. We randomly split the data (grid cells) into 70% for training and 30% for testing. Feature importance was estimated by the Mean Decrease in Impurity (MDI) algorithm, which gives higher importance to the input features selected for the nodes of the decision trees that decrease the model impurity, that is, the modeling errors, by the highest amount. The RF and Feature Importance algorithms were performed using the Python sklearn 1.2.1 library.

To increase confidence in our results, we performed a robustness test. We randomly split the whole data into training and testing data sets three times and subsequently ran the RF algorithm with three different initializations for bootstrapping and feature selection, using the same seed numbers for every GHM/LSMs. In total, the robustness test included nine models (3 × 3) for each combination of hydrological fluxes (four) and GHM/LSM (five). This approach allowed us to evaluate the sensitivity of the RF model performance and feature importance to randomization. In general, the feature importance estimates were very robust (see Text S3 in Supporting Information S1).

The RF models function as metamodels of the GHM/LSMs, replicating their outputs. Given that the metamodels have good performances, we assume that the feature importance estimated by the MDI algorithm also reproduces the original GHM/LSM sensitivity to the input variables. Therefore, we evaluated the RF models' performance in terms of coefficient of determination (R^2) and root mean squared error (RMSE) see Section 3.1.

2.4. Correlated Features

Feature importance assessment is impacted by correlations (Dormann et al., 2013; Gregorutti et al., 2017). Highly correlated features can each take a share of importance as both can reduce the "impurity" to a similar amount. Ideally, we would like to use completely independent variables as input for the RF model, but our analysis involves comparing climate and surface features that inevitably are naturally correlated (e.g., rainforest land cover and precipitation). Thus removing all significantly correlated features is not an option, since important features would be left out. For example, the average precipitation and wind speed correlation is −0.55, and most of the land cover classes of PCR-GLOBWB correlate with precipitation (>0.5). In addition, eliminating too many features would reduce the RF models' performance. Thus, we aimed to eliminate only excessive correlations between input features from the same GHM/LSM. First, we excluded one of each pair of variables that presented a Spearman correlation over 0.75 but belonged to different feature groups (e.g., elevation and air pressure). Second, we excluded one of each pair of variables that presented a Spearman correlation over 0.9 and was within the same feature group (e.g., incident longwave radiation and temperature). As our primary goal is to evaluate the importance at a feature group level, two variables in the same group sharing importance do not impact the conclusions. Thus, elevation and longwave radiation were removed from the analysis. The other input variables that were removed were either land cover or soil features which are specific for every GHM/LSM. Results from this input data selection are presented in Text S2 in Supporting Information S1.

2.5. Analysis

Our goal is to identify the influence of feature groups on different GHM/LSMs. So, we used RF models as a replacement to the original GHM/LSMs to obtain feature importance estimates. Thus, each RF model was trained with input and outputs of a single GHM/LSM. This is the usual approach, named here as the "Regular Case."

However, there remains a challenge in confirming that the differences in the estimated feature importances between GHM/LSMs are related to their structure and not to correlations. The elimination of highly correlated features mentioned in Section 2.4 only partially solved this problem, since there are still many features correlated. Thus, we conducted a cross-feature evaluation. This consisted of training RF with the input features of one GHM/LSM and hydrological fluxes from another. Only the land cover and soil maps were swapped between GHM/LSMs, since the remaining input features are exactly the same. The "Cross Case" does not necessarily eliminate correlations but helps to identify them. It helps to understand (a) how much the correlation between input features affects importance, and (b) to what extent different land cover and soil maps can still explain the variance of the GHM/LSM outputs. One example for each category is presented below:

1. If the JULES soil map is only important for estimating the JULES hydrological fluxes and not the other GHM/LSMs hydrological fluxes, it means that the soil importance is derived from the soil map exclusively. However, if JULES soil map is important to predict hydrological fluxes from other GHM/LSMs while their own soil maps are not that important, it means that the JULES soil map is correlated to some feature of greater relevance (e.g., precipitation) and it might be "stealing" some of its importance.
2. If the RF model can equally represent JULES *Evap* predictions using JULES soil map and/or HTESSEL soil map as input, it means that either the soil map is irrelevant for explaining JULES *Evap* spatial variance or the HTESSEL and JULES soil maps have similar spatial patterns.

A pre-assessment of the similarity between soil and land cover maps of GHM/LSMs was not processed for two reasons. First, the classification is different from model to model, thus it is not possible to calculate a direct correlation since there are no equivalent land and soil features between the models. Second, the spatial pattern refers to the feature group (e.g., land cover) which is a combination of individual features (e.g., each land cover class). Thus, a spatial similarity analysis between the land and soil inputs of different GHM/LSMs cannot be done by a simple feature-to-feature correlation.

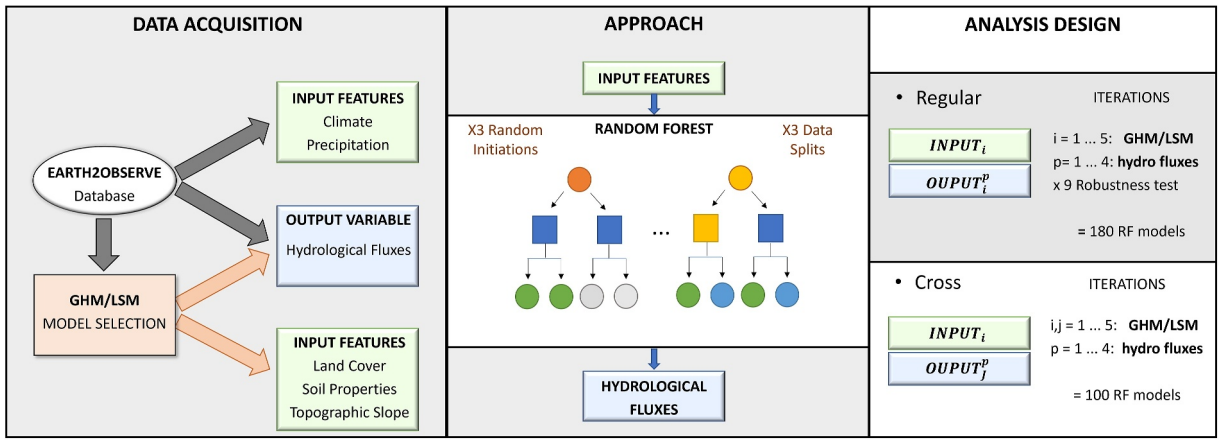

Figure 1. Schematic diagram to explain the methodology. Data Acquisition: obtaining input and output data from the Global Hydrological Model/Land Surface Models either from the E2O database (meteorological) or independently (soil, land cover and topography). Approach: using the input data as predictor and output data as target variables of a Random Forest model. Analysis Design: the Regular Case (with a robustness test) and Cross Evaluation for feature importance analysis.

Thus, we made two complementary analyses of feature importance: the "Regular Case," where the output and input features belong to the same model; and the "Cross Case," where the metamodel is trained with every possible combination of GHM/LSM input and output. In formula:

- Regular case:

$$FI_i = f_{ii}(Out_i, In_i) \quad i = 1\ldots5 : \text{models}$$

- Cross case:

$$FI_{ij} = f_{ij}(Out_i, In_j) \quad i, j = 1\ldots5 : models$$

Where *Out* and *In* are related to the outputs (hydrological fluxes) and input features, respectively. *f* represents the MDI algorithm that calculates the features' importance *FI*. The methodological scheme can be visualized in Figure 1.

3. Results and Interpretation

3.1. Random Forest Performance

The indirect sensitivity analysis through feature importance is only feasible if the surrogate models can adequately replicate the GHM/LSMs. In general, the performance of the RF in reproducing model-simulated fields was satisfactory, even if training the RF model with soil and land cover features from different GHM/LSMs (Figure 2). The RF models presented an R^2 higher than 0.85, even achieving an R^2 higher than 0.98 on the predictions of average evaporation (*Evap*) and average runoff (*Q*). In terms of RMSE, the error was lower than 50 mm for *Evap*, lower than 60 mm for *Q*, and lower than 70 mm for surface (*Qs*) and subsurface (*Qsb*) runoff, except when simulating the PCR-GLOBWB outputs. The RMSE seems high compared to the mean values of *Evap*, *Q*, *Qs*, and *Qsb* within the domain, that is, 490 mm, 300 mm, 110 mm, and 190 mm respectively, however, it is relatively smaller when compared to the spatial variability of these hydrological fluxes given by their standard deviation: 350 mm, 460 mm, 190 mm, 340 mm respectively.

Evap and *Q* are key components of the water balance, so they have been easily replicated by RF models using long-term averages of meteorological variables (Merz & Blöschl, 2009). On the other hand, runoff partitioning in quick (*Qs*) and slow flow (*Qsb*) is event-related and more dependent on temporal variability and previous moisture conditions. *Qs* is mostly formed in saturated areas (Dunne & Black, 1970; Saffarpour et al., 2016) which depends on antecedent rainfall events and meteorological conditions. This is reflected in a slightly lower RF

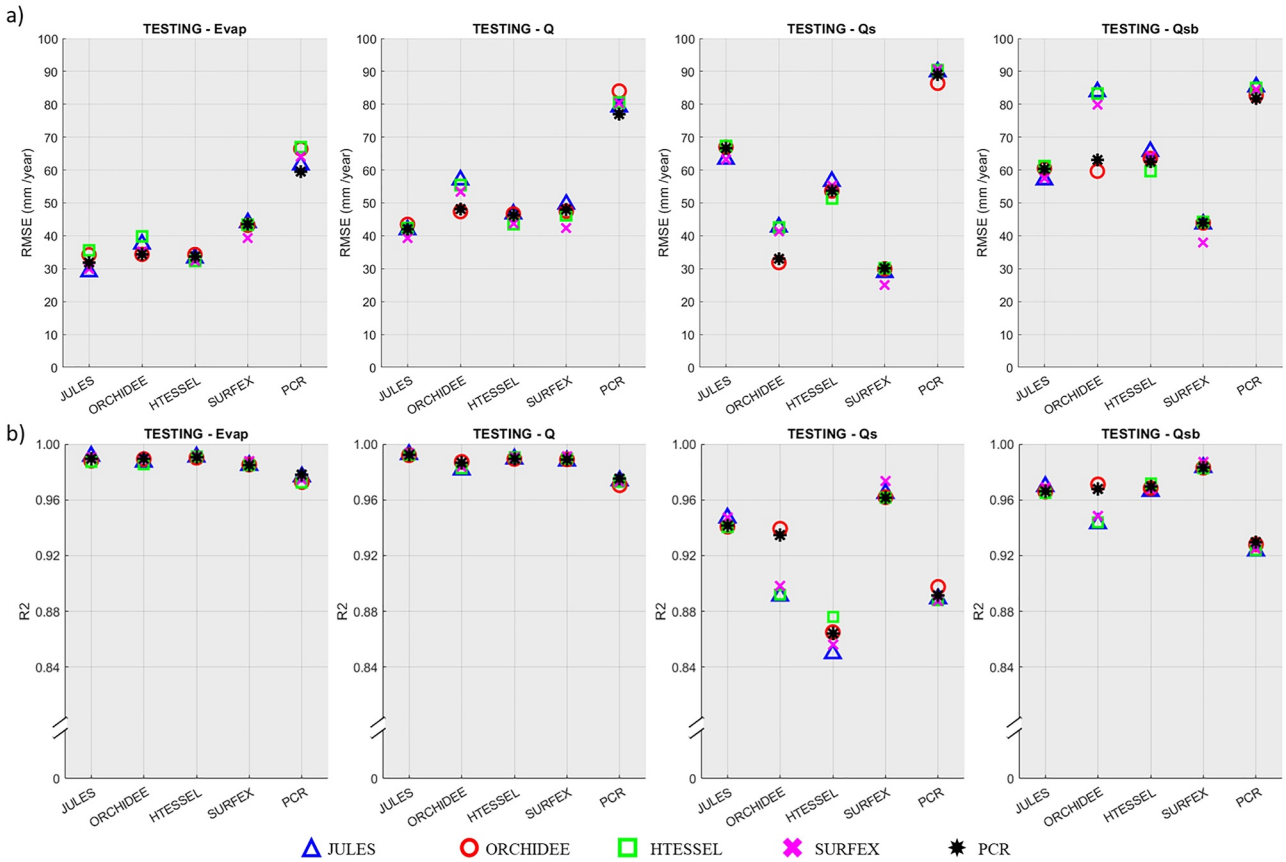

Figure 2. Performance of the Random Forest (RF) models in terms of root mean squared error (a) and R^2 (b) for the testing set. Each chart represents a different hydrological flux. Left to right: Evaporation, Runoff, Surface, and Subsurface Runoff. The hydrological fluxes were calculated from different Global Hydrological Model (GHM)/Land Surface Models (LSMs) outputs (*x*-axis). The symbols and colors indicate the GHM/LSM input features used to train the RF model.

performance for *Qs* and *Qsb*, although still showing quite acceptable R^2 values. Even though the RMSE for *Q* was just slightly lower than the RMSE for *Qs* and *Qsb*, it is important to remember that *Q* is given by the sum of *Qs* and *Qsb*, so the relative error for *Q* is significantly smaller. The same applies to *Evap*.

The RF model presented the best performance when both output and input features were from the same GHM/LSM (Regular Case), that is, *x*-axis label and symbol from the same GHM/LSM. Although this might seem obvious, the number of land cover and soil features varied from 17/18 (ORCHIDEE, SURFEX) to 33 (HTESSEL), and some spurious correlations could have interfered with the RF performance. Nevertheless, the Regular Case is only slightly better, presenting an RMSE 4.36 mm/year lower on average compared to the RF when provided with a different predictor source (Cross Case). This can be explained by two reasons: (a) precipitation and climate importance are generally higher than the land cover and soil importance (see next section), hence the RF model performs well anyway because the GHM/LSMs were simulated with the same meteorological data; (b) different soil and land cover databases still present similar spatial patterns, and the similarity can be sufficient to explain the spatial variability of the hydrological fluxes.

Some GHM/LSMs input features were more closely related than others. For example, land cover and soil features from PCR-GLOBWB and ORCHIDEE could reasonably explain the variance of *Qs* and *Qsb* calculated from ORCHIDEE, but performance was lower when using input features from JULES, HTESSEL or SURFEX. This happened because soil features of PCR-GLOBWB and ORCHIDEE both originate from the FAO Soil Map of the World (DSMW) top soil layer map (S1), and soil features have high importance in predicting ORCHIDEE's *Qs* and *Qsb* (see next section). We do not observe the same performance spread for PCR-GLOBWB, because for PCR-GLOBWB soil was less important in the prediction of *Qs* and *Qsb*. JULES and SURFEX input soil maps are

based on the Harmonized World Soil Database and the HTESSEL soil map is based on another soil layer (30–100 cm) of the DSMW.

The hydrological fluxes estimated by PCR-GLOBWB are the most difficult to predict. PCR-GLOBWB is the only GHM in this study and differs from LSMs in purpose and conceptualization (Beck, Van Dijk, De Roo, et al., 2017; Haddeland et al., 2011). GHMs are traditionally focused on providing accurate estimates of streamflow and surface/groundwater storage exchanges. As a result, hydrological processes are described in more detail and require spatial data on, for example, irrigation and hydrogeological maps, which are employed by PCR-GLOBWB but were not considered here. On the other hand, LSMs are traditionally most concerned with the vertical water balance and land-atmosphere interactions, which might be easier to replicate by the RF models with the given input feature groups.

3.2. Feature Importance Analysis

Figure 3 summarizes the main results of our study by showing the importance of each of the five feature groups for all combinations of GHM/LSMs outputs and inputs. Where RF performance has not changed significantly (see Figure 2), it means that the different maps can explain the variance of a hydrological flux from a specific GHM/LSM to the same amount. So we assume that there is no great loss in averaging importances, and thus the Mean Cross Case would be providing an approximately "unbiased" importance, since it eliminates an inflated importance that may happen in one of the soil/land cover maps due to correlation with a more important feature (like precipitation). On the other hand, when the RF performance has changed considerably, for example, ORCHIDEE *Qs/Qsb* predicted with inputs from JULES, HTESSEL or SURFEX, this assumption is not reasonable and the Regular Case (or an ensemble of the best performers) becomes the best representative of feature importance.

In general, land cover, precipitation and climate share the importance for evaporation estimate equally (Figure 3). By contrast, when estimating runoff more than 50% of the importance is associated with precipitation. Soil texture and topographic slope overall seemed weakly related to the simulated long-term water balance given by *Q* and *Evap*. This corresponds with results from ML studies based on observed data that already asserted a relatively minor influence of soil texture on mean discharge (Addor et al., 2018; Beck et al., 2015), and a high importance of land cover and climate/precipitation for the water balance components (Cheng et al., 2022). Beck et al. (2015) indicated climatological indices, such as the Aridity Index and mean precipitation, as the most important predictors of the flow duration curve, while the slope is shown in third with lower importance.

Besides identifying the general agreement between GHM/LSMs, we also want to evaluate their differences. In doing so, additional caution is required as feature importances may be biased. A noticeable bias example is the land cover importance of JULES. Evaluating the Regular Case alone, we are led to conclude that JULES *Evap* is highly influenced by land cover compared to other GHM/LSMs. However, when considering the Cross Case, Land Cover is predominant in each of the first of each group of columns (column J), which means that when using the land cover map of JULES to predict *Evap* from any GHM/LSM, land cover will always be assigned a higher importance. The JULES land cover bias can be visualized by the contrast between the Regular Case and the Mean Cross Case. In summary, the high importance of land cover for *Evap* in JULES is thus not the result of the JULES model structure, but of the choice for this particular land cover database that has a considerable spatial similarity with climatological features.

We detected substantial differences between the GHM/LSMs for runoff partitioning. Three out of five GHM/LSMs showed a significant influence of topographic slope on surface runoff (JULES, HTESSEL and PCR-GLOBWB). Theoretically, slope is directly related to surface runoff generation. On very steep terrains, hydrographs are dominated by subsurface stormflow (Dunne, 1983), which is defined as part of surface runoff in this article. In addition, hillslopes contribute to a convergent subsurface flow (Anderson & Burt, 1978) and consequently to a greater saturated zone for overland flow (Dunne & Black, 1970). JULES modifications to include slope as a predictor of surface runoff generation occurred during the E2O project as an improvement from Tier 1 to Tier 2 phases (Dutra et al., 2017; Martínez-De La Torre et al., 2019). HTESSEL already considered topographic slope indirectly through the *b* coefficient of the ARNO model (Balsamo et al., 2009; Todini, 1996) while PCR-GLOBWB considered slope explicitly through the representation of subsurface stormflow (Van Beek & Bierkens, 2008). ORCHIDEE and SURFEX do not seem to consider slope effects on surface runoff generation, at least for the E2O project. The spatial difference between JULES and SURFEX in terms of surface runoff and

10.1029/2023WR036418

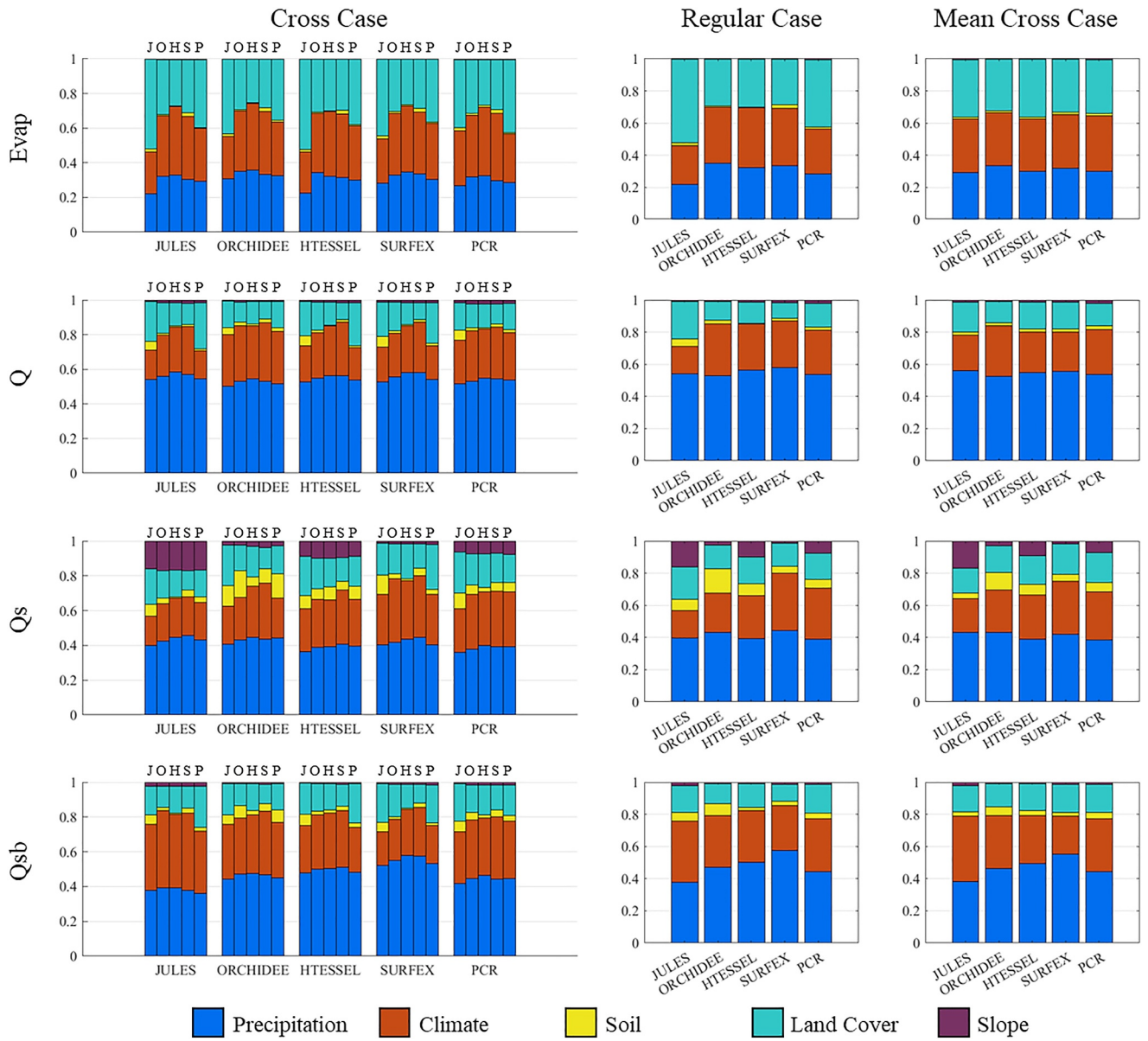

Figure 3. Feature importance of five feature groups (Precipitation, Climate, Soil, Land Cover, and Slope) for the prediction of four hydrological fluxes (*Evap*, *Q*, *Qs*, and *Qsb*) by Random Forest models given different Global Hydrological Model (GHM)/Land Surface Model (LSM) as the source of input (predictors) and output (predictand) data. The Cross Case considers all the possible combinations of GHM/LSM input and output data. The group columns indicated on the *x*-axis refer to the GHM/LSM that provided the output data and each minor column indicates which GHM/LSM provided the input data: J-JULES, O-ORCHIDEE, H-HTESSEL, S-SURFEX, and P-PCR-GLOBWB. The Regular Case indicates the inputs and outputs data from the same GHM/LSM. The Mean Cross Case is the average of the group columns of the Cross Case.

precipitation ratio (*Qs/P*) can be visualized in the left column of Figure 4. We can see how the *Qs/P* is outlined for JULES, presenting values over 0.5 in almost all mountain areas such as Western North America, the Andes Cordillera, Eastern Africa, the Caucasus, the Himalayas, New Guinea, the Altai mountains, etc., while we do not observe the same demarcation for SURFEX.

The GHM/LSMs also disagree about the importance of soil for runoff partitioning. Soil features seem more important to ORCHIDEE than for the other GHM/LSMs. For ORCHIDEE in particular, the feature importance shown by the Regular Case is more suitable since the RF performance considerably declines when using soil maps of other GHM/LSMs except for PCR-GLOBWB (Figure 2). Soil importance for ORCHIDEE was 15% for *Qs* and 7.5% for *Qsb*, which is twice the importance of the second highest soil importance estimated (*Qs*—HTESSEL,

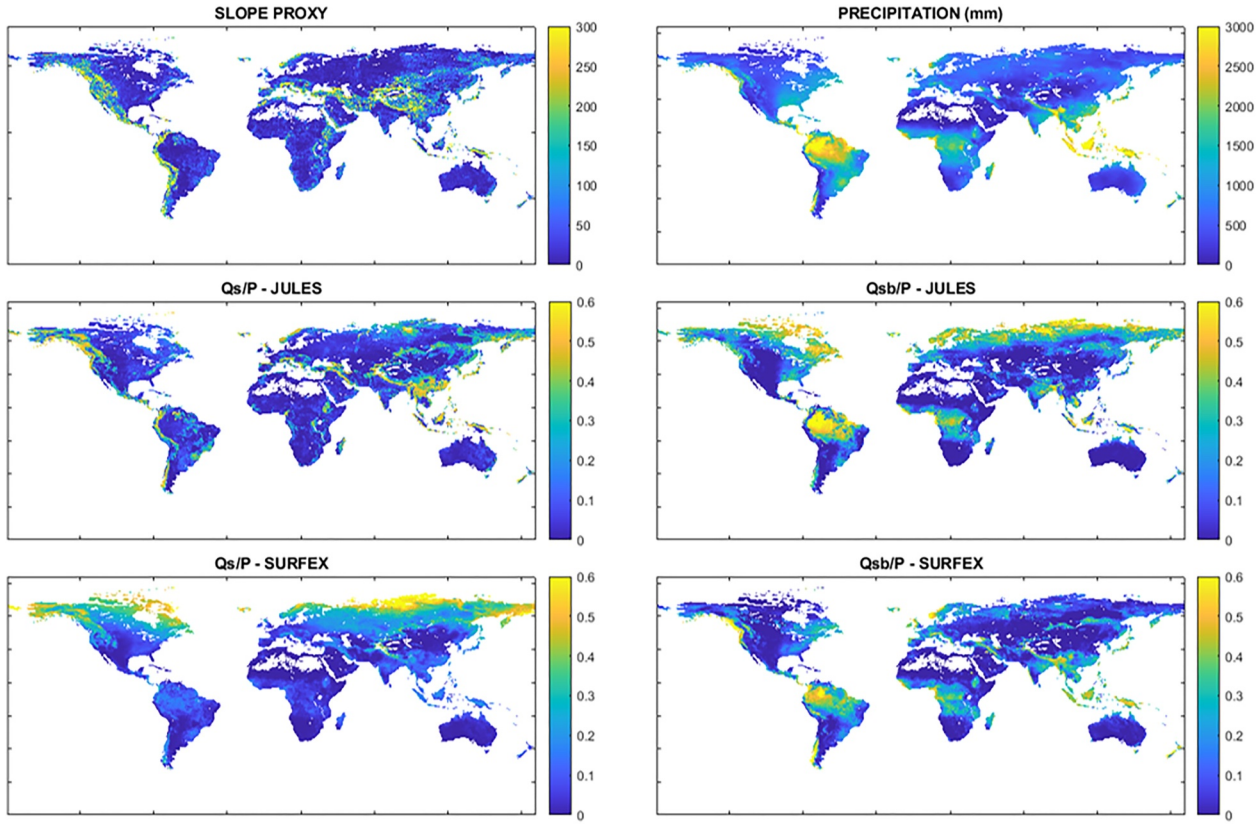

Figure 4. Global maps of the studied domain presenting the Slope Proxy, Annual Precipitation (mm), Surface and Subsurface Runoff Precipitation ratio (*Qs*/*P* and *Qsb*/*P*) estimated with outputs from JULES and SURFEX. Obs.: Some areas are not included in the domain due to very low values of P which is in the denominator of the hydrological indices (e.g., Sahara desert).

Qsb—PCR-GLOBWB/JULES). We speculate that this high importance given to soil texture can be explained by the infiltration-excess mechanism for surface runoff generation adopted by ORCHIDEE. In ORCHIDEE, the maximum infiltration capacity is equivalent to the soil hydraulic conductivity which is directly proportional to the saturated hydraulic conductivity, determined solely by soil texture. Although the saturation-excess mechanism adopted in the other GHM/LSMs also depends on soil texture, the connection between soil texture and surface runoff is not as straightforward. Tafasca et al. (2020) tested different soil texture maps in ORCHIDEE and observed a low sensitivity of the water balance but a considerable sensitivity of surface runoff and soil moisture, especially associated with soil clay percentage. Our findings seem in line with these conclusions. Previous ML studies based on observations have detected a weak but existing soil texture importance for streamflow properties with clay fraction ahead of sand and silt (Addor et al., 2018; Beck et al., 2015; Kratzert et al., 2019). Therefore, soil texture indeed appears to have some importance for runoff, but the real extent of soil importance is still in debate. For example, GHM/LSMs clearly represent it differently and there is a recently open discussion about hydrological models overestimating the soil importance (Gao et al., 2023). Nevertheless, there is still much room for improvements in soil process representation by global models (Vereecken et al., 2022), which may lead to greater consensus on soil importance in the future.

Finally, the GHM/LSMs disagreed on the importance of precipitation/climate for *Qsb*. SURFEX presented the highest precipitation importance (≈57%) and JULES the lowest (≈38%). Such a high influence of a single feature (mean precipitation) on *Qsb* from SURFEX explains why the RF performance ($R^2 > 0.98$) was so high even when using different soil and land cover databases (see Figure 2). Nevertheless, the visual differences between JULES and SURFEX related to the spatial influence of precipitation on *Qsb* are not obvious (Figure 4), except on high latitudes. This could also be related to the way these models treat frozen soils, and water flow within and over these permafrost surfaces.

4. Limitations

There are several limitations associated with employing RF feature importance as a diagnostic approach. For instance, the feature importance is not directly applied to GHM/LSMs but to surrogate models. Therefore, any analysis conducted is greatly dependent on the ability of the surrogate models to replicate the GHM/LSMs outputs. These constraints hinder the method's applicability and the outreach of this manuscript. However, all measures were taken to provide the most robust results and parsimonious conclusions given the existing limitations.

One relevant simplification relates to temporal aspects. Both predictors and target variables are long-term averages, so the feature importance did not reflect the GHM/LSMs responses to time variability such as meteorological events or seasonality. This might have affected the performance of the surrogate models, especially for the runoff partitioning since it depends on previous soil moisture conditions. Consequently, there could be an impact on the feature importance given that the temporal variability of the hydrological fluxes is ignored.

Furthermore, the number of predictors is constrained by the overlapping of GHM/LSMs input features. Only meteorological variables, soil texture, land cover, and topography were taken as predictors since they are the core and common inputs for hydrological simulations. However, each GHM/LSM has its own set of inputs; for instance, the PCR-GLOBWB simulations also involve water demand and groundwater data. The lack of enough predictors might explain why the surrogate models had the lowest performance for PCR-GLOBWB.

Despite the aforementioned limitations, the RF models could reasonably replicate the outputs of the GHM/LSMs. The coefficient of determination (R^2) was over 0.85 for all hydrological fluxes, which indicates that they were able to fairly represent the spatial variability of the global models. It also suggests that the respective input features were sufficient and that the MDI algorithm provided acceptable feature importance estimates.

However, the MDI algorithm is also susceptible to correlations among input features. Given that dependencies between environmental features are natural and inevitable, we proposed alternative solutions to reduce the correlation impact. We establish high correlation thresholds for feature elimination, thereby retaining 83% of the original input features designated for RF training. For instance, if we set the correlation threshold for different groups to 0.5 (actual value equal to 0.75), the percentage of original features would drop to 55%. Therefore we accepted highly correlated variables (>0.5) to the detriment of obtaining a larger number of representative features and, consequently, a better performance from the surrogate models. As a drawback, the feature groups remained highly correlated and the feature importance was compromised. This problem was exposed by the biased importance of JULES Land Cover (Figure 3). We introduced the “Cross Case” which was able to identify some of these problems, and the results were interpreted accordingly.

Lastly, it's worth mentioning that no set of observations was used. Thus, we only assessed how the models differ among themselves and avoided judging the merits of the modeling approaches. We tried to establish connections between the physical structure of the models and the respective feature importance. For instance, we identified possible explanations for slope and soil texture importance on simulated surface and subsurface runoff generation. However, defining which modeling approach is more accurate would demand additional analysis, deeper consensus on the underlying physical processes, and collaboration among modeling groups.

5. Conclusion

This paper proposed a novel model intercomparison approach that contributes to the structural understanding of GHM/LSMs. Random Forest (RF) was used to mimic the temporal average of hydrological fluxes of GHM/LSMs, enabling an indirect sensitivity analysis (Antoniadis et al., 2021). We presented a consistent set of approaches that increased the reliability of the results, such as considerably high RF performance, robustness test, correlation analysis and cross-feature evaluation.

Then we assessed the influence of five input feature groups (precipitation, climate, soil texture, land cover and topographic slope) on explaining the variance of mean evaporation, runoff, surface runoff and subsurface runoff on the global land domain. In general, GHM/LSMs agree on the importance of predictors for water balance but not for runoff partitioning in fast and slow flow. Soil texture and slope were irrelevant for simulated water balance

but relevant for surface and subsurface runoff, although GHM/LSMs disagreed on the strength of this importance. Specifically, the topographic slope had no influence on surface runoff predictions of ORCHIDEE and SURFEX.

We noticed that the soil map selection is relevant, but to a degree that depends on the hydrological variable and GHM/LSM analyzed. Tafasca et al. (2020) found a weak influence of soil mapping on the water balance for ORCHIDEE, which agrees with our conclusion. However, we found that, for surface and subsurface runoff calculated from ORCHIDEE itself, using different soil databases as predictors affected the surrogate model performance. On the other hand, we could not reach the same conclusion for other GHM/LSMs since the soil importance was lower compared to ORCHIDEE. Such findings are important for ongoing MIP projects such as the Soil Parameter MIP (Verhoef et al., 2022).

The present study documents the diagnostic potential of ML methods and shows that these or similar statistical/data-driven approaches can be valuable for MIPs. Combined with observation data sets, this method can provide an in-depth evaluation of GHM/LSMs. For example, feature importance can identify significant differences in model structures and point toward the best modeling approaches given that their performances are known. In addition, the sensitivity analysis can indicate whether the models simulate accurate feedback between meteorological drivers and hydrological fluxes on different landscapes. Nevertheless, this method was limited to mutual features of GHM/LSMs and average fluxes of surrogate models. With adequate planning involving the cooperation of modelers, it is possible to reach an even deeper analysis, considering a complete and selected set of input features and time-dependent relationships. Finally, this study introduces the applicability of ML techniques for intercomparison projects and adds to the existing literature on process-based evaluations of GHM/LSMs. Our work also highlights the great and enduring value of projects like E2O, which took care to standardize the model run specifications (e.g., simulation period and spatiotemporal resolution) and which greatly facilitates comparisons between models and analysis such as the one presented here.

Data Availability Statement

The original meteorological data and outputs from the GHM/LSMs were obtained from the Earth2Observe project (Dutra et al., 2017) and are available online. Processed data of all predictors and target variables, including the codes for making figures from the main text and Supplement Material are also available online (Brêda, 2023).

Acknowledgments
We thank Agnès Ducharne, Emanuel Dutra, and Alberto Martinez-de la Torre for leading to the right soil and land cover input data sets of ORCHIDEE, HTESSEL, and JULES respectively. And we thank the SURFEX and PCR-GLOBWB developers for making the "physiographic maps" easily accessible. We also thank Rafael Fontana for the discussion about the correlation impact on feature importance. Finally, we thank the anonymous reviewers for their essential contributions.

References

Addor, N., Nearing, G., Prieto, C., Newman, A. J., Le Vine, N., & Clark, M. P. (2018). A ranking of hydrological signatures based on their predictability in space. *Water Resources Research*, *54*(11), 8792–8812. https://doi.org/10.1029/2018WR022606

Anderson, M. G., & Burt, T. P. (1978). The role of topography in controlling throughflow generation. *Earth Surface Processes*, *3*(4), 331–344. https://doi.org/10.1002/esp.3290030402

Antoniadis, A., Lambert-Lacroix, S., & Poggi, J. M. (2021). Random forests for global sensitivity analysis: A selective review. *Reliability Engineering & System Safety*, *206*, 107312. https://doi.org/10.1016/j.ress.2020.107312

Balsamo, G., Viterbo, P., Beijaars, A., van den Hurk, B., Hirschi, M., Betts, A. K., & Scipal, K. (2009). A revised hydrology for the ECMWF model: Verification from field site to terrestrial water storage and impact in the integrated forecast system. *Journal of Hydrometeorology*, *10*(3), 623–643. https://doi.org/10.1175/2008JHM1068.1

Beck, H. E., de Roo, A., & van Dijk, A. I. J. M. (2015). Global maps of streamflow characteristics based on observations from several thousand catchments. *Journal of Hydrometeorology*, *16*(4), 1478–1501. https://doi.org/10.1175/JHM-D-14-0155.1

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

Beck, H. E., Van Dijk, A. I. J. M., Levizzani, V., Schellekens, J., Miralles, D. G., Martens, B., & De Roo, A. (2017). MSWEP: 3-hourly 0.25° global gridded precipitation (1979-2015) by merging gauge, satellite, and reanalysis data. *Hydrology and Earth System Sciences*, *21*(1), 589–615. https://doi.org/10.5194/hess-21-589-2017

Best, M. J., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., et al. (2015). The plumbing of land surface models: Benchmarking model performance. *Journal of Hydrometeorology*, *16*(3), 1425–1442. https://doi.org/10.1175/JHM-D-14-0158.1

Beven, K. J., & Kirkby, M. J. (1979). A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. *Hydrological Sciences Journal*, *24*(1), 43–69. https://doi.org/10.1080/02626667909491834

Bierkens, M. F. P., Bell, V. A., Burek, P., Chaney, N., Condon, L. E., David, C. H., et al. (2015). Hyper-resolution global hydrological modelling: What is next?: "Everywhere and locally relevant" M. F. P. Bierkens et al. Invited Commentary. *Hydrological Processes*, *29*(2), 310–320. https://doi.org/10.1002/hyp.10391

Brêda, J. P. L. F. (2023). Predictor importance for hydrological fluxes of Global Hydrological and Land Surface Models [Codes and Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.8379355

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

Brooks, R. H., & Corey, A. T. (1966). Properties of porous media affecting fluid flow. *Journal of the Irrigation and Drainage Division*, *92*(2), 61–88. https://doi.org/10.1061/jrcea4.0000425

Cappelli, F., Tauro, F., Apollonio, C., Petroselli, A., Borgonovo, E., & Grimaldi, S. (2022). Feature importance measures to dissect the role of sub-basins in shaping the catchment hydrological response: A proof of concept. *Stochastic Environmental Research and Risk Assessment*, *37*(4), 1247–1264. https://doi.org/10.1007/s00477-022-02332-w

Cheng, S., Cheng, L., Qin, S., Zhang, L., Liu, P., Liu, L., et al. (2022). Improved understanding of how catchment properties control hydrological partitioning through machine learning. *Water Resources Research*, *58*(4). https://doi.org/10.1029/2021WR031412

Chiew, F. H. S. (2006). Estimation of rainfall elasticity of streamflow in Australia. *Hydrological Sciences Journal*, *51*(4), 613–625. https://doi.org/10.1623/hysj.51.4.613

Clapp, R. B., & Hornberger, G. M. (1978). Empirical equations for some soil hydraulic properties. *Water Resources Research*, *14*(4), 601–604. https://doi.org/10.1029/wr014i004p00601

Clark, E. A., Sheffield, J., van Vliet, M. T. H., Nijssen, B., & Lettenmaier, D. P. (2015). Continental runoff into the oceans (1950–2008). *Journal of Hydrometeorology*, *16*(4), 1502–1520. https://doi.org/10.1175/JHM-D-14-0183.1

Clark, M. P., Fan, Y., Lawrence, D. M., Adam, J. C., Bolster, D., Gochis, D. J., et al. (2015). Improving the representation of hydrologic processes in Earth System Models. *Water Resources Research*, *51*(8), 5929–5956. https://doi.org/10.1002/2015WR017096

Clark, M. P., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, *47*(9). https://doi.org/10.1029/2010WR009827

Collins, W. J., Bellouin, N., Doutriaux-Boucher, M., Gedney, N., Halloran, P., Hinton, T., et al. (2011). Development and evaluation of an Earth-System model - HadGEM2. *Geoscientific Model Development*, *4*(4), 1051–1075. https://doi.org/10.5194/gmd-4-1051-2011

De Boer-Euser, T., Bouaziz, L., De Niel, J., Brauer, C., Dewals, B., Drogue, G., et al. (2017). Looking beyond general metrics for model comparison - Lessons from an international model intercomparison study. *Hydrology and Earth System Sciences*, *21*(1), 423–440. https://doi.org/10.5194/hess-21-423-2017

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

De la Fuente, L. A., Gupta, H. V., & Condon, L. E. (2023). Toward a multi-representational approach to prediction and understanding, in support of discovery in hydrology. *Water Resources Research*, *59*(1). https://doi.org/10.1029/2021WR031548

Dirmeyer, P. A. (2011). A history and review of the Global Soil Wetness Project (GSWP). *Journal of Hydrometeorology*, *12*(5), 729–749. https://doi.org/10.1175/JHM-D-10-05010.1

Döll, P., Douville, H., Güntner, A., Müller Schmied, H., & Wada, Y. (2016). Modelling freshwater resources at the global scale: Challenges and prospects. *Surveys in Geophysics*, *37*(2), 195–221. https://doi.org/10.1007/s10712-015-9343-1

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., et al. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Dufresne, J. L., Foujols, M. A., Denvil, S., Caubel, A., Marti, O., Aumont, O., et al. (2013). Climate change projections using the IPSL-CM5 Earth System Model: From CMIP3 to CMIP5. *Climate Dynamics*, *40*(9–10), 2123–2165. https://doi.org/10.1007/s00382-012-1636-1

Dunne, T. (1983). Relation of field studies and modeling in the prediction of storm runoff. *Journal of Hydrology*, *65*(1–3), 25–48. https://doi.org/10.1016/0022-1694(83)90209-3

Dunne, T., & Black, R. D. (1970). Partial area contributions to storm runoff in a small New England watershed. *Water Resources Research*, *6*(5), 1296–1311. https://doi.org/10.1029/WR006i005p01296

Dutra, E., Balsamo, G., Calvet, J.-C., Munier, S., Burke, S., Fink, G., et al. (2017). WP5-Task 5.1-D.5.2- Report on the improved water resources reanalysis. Retrieved from http://earth2observe.eu/files/Public%20Deliverables/D5.2%20-%20Report%20on%20the%20Improved%20Water%20Resources%20Reanalysis%20(WRR%20tier%202).pdf

Fisher, R. A., & Koven, C. D. (2020). Perspectives on the future of land surface models and the challenges of representing complex terrestrial systems. *Journal of Advances in Modeling Earth Systems*, *12*(4). https://doi.org/10.1029/2018MS001453

Gao, H., Fenicia, F., & Savenije, H. H. G. (2023). HESS Opinions: Are soils overrated in hydrology? *Hydrology and Earth System Sciences*, *27*(14), 2607–2620. https://doi.org/10.5194/hess-27-2607-2023

Giuntoli, I., Vidal, J. P., Prudhomme, C., & Hannah, D. M. (2015). Future hydrological extremes: The uncertainty from multiple global climate and global hydrological models. *Earth System Dynamics*, *6*(1), 267–285. https://doi.org/10.5194/esd-6-267-2015

Gleeson, T., Wagener, T., Döll, P., Zipper, S. C., West, C., Wada, Y., et al. (2021). GMD perspective: The quest to improve the evaluation of groundwater representation in continental-to global-scale models. *Geoscientific Model Development*, *14*(12), 7545–7571. https://doi.org/10.5194/gmd-14-7545-2021

Gnann, S., Reinecke, R., Stein, L., Wada, Y., Thiery, W., Müller Schmied, H., et al. (2023). Functional relationships reveal differences in the water cycle representation of global water models. *Nature Water*, *1*(12), 1079–1090. https://doi.org/10.1038/s44221-023-00160-y

Gosling, S. N., Zaherpour, J., Mount, N. J., Hattermann, F. F., Dankers, R., Arheimer, B., et al. (2017). A comparison of changes in river runoff from multiple global and catchment-scale hydrological models under global warming scenarios of 1°C, 2°C and 3°C. *Climatic Change*, *141*(3), 577–595. https://doi.org/10.1007/s10584-016-1773-3

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing*, *27*(3), 659–678. https://doi.org/10.1007/s11222-016-9646-1

Gu, H., Xu, Y. P., Ma, D., Xie, J., Liu, L., & Bai, Z. (2020). A surrogate model for the Variable Infiltration Capacity model using deep learning artificial neural network. *Journal of Hydrology*, *588*, 125019. https://doi.org/10.1016/j.jhydrol.2020.125019

Haddeland, I., Clark, D. B., Franssen, W., Ludwig, F., Voß, F., Arnell, N. W., et al. (2011). Multimodel estimate of the global terrestrial water balance: Setup and first results. *Journal of Hydrometeorology*, *12*(5), 869–884. https://doi.org/10.1175/2011JHM1324.1

Hagemann, S., Chen, C., Clark, D. B., Folwell, S., Gosling, S. N., Haddeland, I., et al. (2013). Climate change impact on available water resources obtained using multiple global climate and hydrology models. *Earth System Dynamics*, *4*(1), 129–144. https://doi.org/10.5194/esd-4-129-2013

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer.

Henderson-Sellers, A., Yang, Z.-L., & Dickinson, R. E. (1993). The project for intercomparison of land-surface parameterization schemes. *Bulletin of the American Meteorological Society*, *74*(7), 1335–1350. https://doi.org/10.1175/1520-0477(1993)074<1335:tpfiol>2.0.co;2

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Huang, F., Zhang, Y., Zhang, Y., Nourani, V., Li, Q., Li, L., & Shangguan, W. (2023). Towards interpreting machine learning models for predicting soil moisture droughts. *Environmental Research Letters*, *18*(7), 074002. https://doi.org/10.1088/1748-9326/acdbe0

Kraft, B., Jung, M., Körner, M., Koirala, S., & Reichstein, M. (2022). Towards hybrid modeling of the global hydrological cycle. *Hydrology and Earth System Sciences*, *26*(6), 1579–1614. https://doi.org/10.5194/hess-26-1579-2022

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*(12), 5089–5110. https://doi.org/10.5194/hess-23-5089-2019

Krinner, G., Viovy, N., de Noblet-Ducoudré, N., Ogée, J., Polcher, J., Friedlingstein, P., et al. (2005). A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system. *Global Biogeochemical Cycles*, *19*(1). https://doi.org/10.1029/2003GB002199

Krysanova, V., Vetter, T., Eisner, S., Huang, S., Pechlivanidis, I., Strauch, M., et al. (2017). Intercomparison of regional-scale hydrological models and climate change impacts projected for 12 large river basins worldwide - A synthesis. *Environmental Research Letters*, *12*(10), 105002. https://doi.org/10.1088/1748-9326/aa8359

Laimighofer, J., Melcher, M., & Laaha, G. (2022). Parsimonious statistical learning models for low-flow estimation. *Hydrology and Earth System Sciences*, *26*(1), 129–148. https://doi.org/10.5194/hess-26-129-2022

Le Moigne, P. (2018). Surfex scientific documentation.

Li, K., Huang, G., & Baetz, B. (2021). Development of a Wilks feature importance method with improved variable rankings for supporting hydrological inference and modelling. *Hydrology and Earth System Sciences*, *25*(9), 4947–4966. https://doi.org/10.5194/hess-25-4947-2021

Martínez-De La Torre, A., Blyth, E. M., & Weedon, G. P. (2019). Using observed river flow data to improve the hydrological functioning of the JULES land surface model (vn4.3) used for regional coupled modelling in Great Britain (UKC2). *Geoscientific Model Development*, *12*(2), 765–784. https://doi.org/10.5194/gmd-12-765-2019

McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, *100*(11), 2175–2199. https://doi.org/10.1175/BAMS-D-18-0195.1

Merz, R., & Blöschl, G. (2009). A regional analysis of event runoff coefficients with respect to climate and catchment characteristics in Austria. *Water Resources Research*, *45*(1). https://doi.org/10.1029/2008WR007163

Müller Schmied, H., Caceres, D., Eisner, S., Flörke, M., Herbert, C., Niemann, C., et al. (2021). The global water resources and use model WaterGAP v2.2d: Model description and evaluation. *Geoscientific Model Development*, *14*(2), 1037–1079. https://doi.org/10.5194/gmd-14-1037-2021

National Geophysical Data Center. (1993). *5-minute Gridded Global Relief Data (ETOPO5)*. National Geophysical Data Center, NOAA.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., et al. (2021). What role does hydrological science play in the age of machine learning? *Water Resources Research*, *57*(3). https://doi.org/10.1029/2020WR028091

Pappenberger, F., Cloke, H. L., Balsamo, G., Ngo-Duc, T., & Oki, T. (2010). Global runoff routing with the hydrological component of the ECMWF NWP system. *International Journal of Climatology*, *30*(14), 2155–2174. https://doi.org/10.1002/joc.2028

Razavi, S., Tolson, B. A., & Burn, D. H. (2012). Review of surrogate modeling in water resources. *Water Resources Research*, *48*(7). https://doi.org/10.1029/2011WR011527

Saffarpour, S., Western, A. W., Adams, R., & McDonnell, J. J. (2016). Multiple runoff processes and multiple thresholds control agricultural runoff generation. *Hydrology and Earth System Sciences*, *20*(11), 4525–4545. https://doi.org/10.5194/hess-20-4525-2016

Schellekens, J., Dutra, E., Martínez-De La Torre, A., Balsamo, G., Van Dijk, A., Sperna Weiland, F., et al. (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

Schmidt, L., Heße, F., Attinger, S., & Kumar, R. (2020). Challenges in applying machine learning models for hydrological inference: A case study for flooding events across Germany. *Water Resources Research*, *56*(5). https://doi.org/10.1029/2019WR025924

Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research*, *54*(11), 8558–8593. https://doi.org/10.1029/2018WR022643

Stein, L., Clark, M. P., Knoben, W. J. M., Pianosi, F., & Woods, R. A. (2021). How do climate and catchment attributes influence flood generating processes? A large-sample study for 671 catchments across the contiguous USA. *Water Resources Research*, *57*(4). https://doi.org/10.1029/2020WR028300

Sun, R., Pan, B., & Duan, Q. (2023). A surrogate modeling method for distributed land surface hydrological models based on deep learning. *Journal of Hydrology*, *624*, 129944. https://doi.org/10.1016/j.jhydrol.2023.129944

Tafasca, S., Ducharne, A., & Valentin, C. (2020). Weak sensitivity of the terrestrial water budget to global soil texture maps in the ORCHIDEE land surface model. *Hydrology and Earth System Sciences*, *24*(7), 3753–3774. https://doi.org/10.5194/hess-24-3753-2020

Telteu, C. E., Müller Schmied, H., Thiery, W., Leng, G., Burek, P., Liu, X., et al. (2021). Understanding each other's models An introduction and a standard representation of 16 global water models to support intercomparison, improvement, and communication. *Geoscientific Model Development*, *14*(6), 3843–3878. https://doi.org/10.5194/gmd-14-3843-2021

Todini, E. (1996). The ARNO rainfall—Runoff model. *Journal of Hydrology*, *175*(1–4), 339–382. https://doi.org/10.1016/S0022-1694(96)80016-3

Van Beek, R. L. P. H., & Bierkens, M. F. P. (2008). The global hydrological model PCR-GLOBWB: Conceptualization, parameterization and verification report. Retrieved from http://vanbeek.geo.uu.nl/suppinfo/vanbeekbierkens2009.pdf

van den Hurk, B., Best, M., Dirmeyer, P., Pitman, A., Polcher, J., & Santanello, J. (2011). Acceleration of land surface model development over a decade of glass. *Bulletin of the American Meteorological Society*, *92*(12), 1593–1600. https://doi.org/10.1175/BAMS-D-11-00007.1

Van Genuchten, M. T. (1980). A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal*, *44*(5), 892–898. https://doi.org/10.2136/sssaj1980.03615995004400050002x

Vereecken, H., Amelung, W., Bauke, S. L., Bogena, H., Brüggemann, N., Montzka, C., et al. (2022). Soil hydrology in the Earth system. *Nature Reviews Earth & Environment*, *3*(9), 573–587. https://doi.org/10.1038/s43017-022-00324-6

Verhoef, A., Zeng, Y., Cuntz, M., Gudmundsson, L., Thober, S., McGuire, P. C., et al. (2022). Assessing the variability of soil temperatures in Land Surface Models using outputs from the Soil Parameter Model Intercomparison Project (SP-MIP). *Copernicus Meetings*.

Wagener, T., Reinecke, R., & Pianosi, F. (2022). On the evaluation of climate change impact models. *Wiley Interdisciplinary Reviews: Climate Change*, *13*(3), e772. https://doi.org/10.1002/wcc.772

Walters, D. N., Williams, K. D., Boutle, I. A., Bushell, A. C., Edwards, J. M., Field, P. R., et al. (2014). The met office unified model global atmosphere 4.0 and JULES global land 4.0 configurations. *Geoscientific Model Development*, *7*(1), 361–386. https://doi.org/10.5194/gmd-7-361-2014

Wang, S., Peng, H., Hu, Q., & Jiang, M. (2022). Analysis of runoff generation driving factors based on hydrological model and interpretable machine learning method. *Journal of Hydrology: Regional Studies*, *42*, 101139. https://doi.org/10.1016/j.ejrh.2022.101139

Zaherpour, J., Gosling, S. N., Mount, N., Schmied, H. M., Veldkamp, T. I. E., Dankers, R., et al. (2018). Worldwide evaluation of mean and extreme runoff from six global-scale hydrological models that account for human impacts. *Environmental Research Letters*, *13*(6), 065015. https://doi.org/10.1088/1748-9326/aac547

Zeng, X., Schnier, S., & Cai, X. (2021). A data-driven analysis of frequent patterns and variable importance for streamflow trend attribution. *Advances in Water Resources*, *147*, 103799. https://doi.org/10.1016/j.advwatres.2020.103799

Zhang, L., Dirmeyer, P. A., Wei, J., Guo, Z., & Lu, C. H. (2011). Land-atmosphere coupling strength in the Global Forecast System. *Journal of Hydrometeorology*, *12*(1), 147–156. https://doi.org/10.1175/2010JHM1319.1

Zounemat-Kermani, M., Batelaan, O., Fadaee, M., & Hinkelmann, R. (2021). Ensemble machine learning paradigms in hydrology: A review. *Journal of Hydrology*, *598*, 126266. https://doi.org/10.1016/j.jhydrol.2021.126266