Adding machine learning to the MIP toolkit: 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], Martine van der Ploeg[1]

1 – Hydrology and Quantitative Water Management, Wageningen University & Research

2 – Data Competence Centre, Wageningen University & Research

3 – The Fenner School of Environment & Society, Australia National University

4 – Instituto de Pesquisas Hidráulicas, Universidade Federal do Rio Grande do Sul

5 – Department of Geography and Environmental Science, University of Reading

6 – Department of Water Resources, University of Twente

Abstract

Global Hydrological and Land Surface Models (GHM/LSMs) embody numerous interacting predictors and equations, complicating the diagnosis of primary hydrological relationships. We propose a model diagnostic approach based on Random Forest feature importance to detect the input variables that most influence simulated hydrological processes. 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 runoffs. The machine learning model could reproduce GHM/LSMs outputs with a coefficient of determination over 0.85 in all cases and often considerably better. The GHM/LSMs agreed precipitation, climate and land cover share equal importance for evaporation prediction, and mean precipitation is the most important predictor of runoff. However, the GHM/LSMs disagreed on which features determine surface and subsurface runoff processes, especially with regards to the relative importance of soil texture and topographic slope.

1. Introduction

Global Hydrological Models (GHM) and Land Surface Models (LSM) 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). In that sense, they provide valuable weather and climate forecasts for the short to long term, as well as historical re-analyses (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 level and the anthropogenic water use (Clark et al., 2015; Müller Schmied et al., 2021).

However, global simulations present significant uncertainties. Global models oversimplify the hydrological cycle by constraining 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; Telteu et al., 2021). 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 suggest that the GHM/LSMs model selection is a major source of uncertainty for evaporation (Hagemann et al., 2013) and low discharge (Giuntoli et al., 2015; Krysanova et al., 2017) projections, and the ensemble spread of GHM/LSMs is considerably larger than catchment hydrological models for discharge (Gosling et al., 2017).

Since the 90's, Model Intercomparison Projects (MIP) have been proposed to establish evaluation frameworks for LSMs (Henderson-Sellers et al., 1993) usually by comparing model outputs to an observation database (Best et al., 2015). Throught the year, 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 modelling weaknesses. However, it is complicated to address these issues when there is a limited understanding of the multitude of processes and variables interactions within a GHM/LSM.

Progressively, data-driven techniques have been assuming a leading role in hydrological modeling (Nearing et al., 2021). Machine learning (ML) has already been successful in predicting surface water and groundwater stores and flows at 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, ML are data-driven models that can provide important statistical information and process understanding. 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, attributes and variables, while feature importance has also been termed variable ranking. 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) and streamflow trends (Zeng et al., 2021).

If feature importance is understood better, modelers can direct effort toward improving the quality of the input data that has the greatest impact on the hydrological model performance. In addition, when an ML model is used to emulate a conceptual/physical-based model (Razavi et al., 2012), feature importance assessment helps to recognize which variables and processes are being overlooked by the physical model and give them due attention (Cappelli et al., 2022; Wang et al., 2022).

In this paper, we are proposing a new approach for global model comparison by using feature importance as a diagnostic tool in addition to the conventional assessment of the agreement of simulated outputs to observations. To achieve this goal, we trained random forest models to reproduce hydrological average fluxes from GHM/LSMs and infer about the importance of input data. In addition, we trained the ML models with swapped land cover and soil maps to understand if different input databases can explain equally the spatial variance of the hydrological fluxes and to identify biased importances. Finally, this study can provide guidance on further model development and help to indicate regions of high model disagreement in terms of process representation.

2. Methodology

We selected GHM/LSMs from the Earth2Observe project and downloaded the respective datasets. Time-dependent variables were averaged to create static maps. Thus, each grid cell in the global domain contains both input features and output of a given GHM/LSM. We fed this information to a Random Forest, which would act as a surrogate model or "metamodel" of the GHM/LSMs (Razavi et al., 2012). The advantage of this method is that the Random Forest algorithm is able to provide feature importance based on how features are selected for each node of the decision trees. To enable comparison between the GHM/LSMs, the input features were grouped into climate, precipitation, soil, land cover and topographic slope. In the following sections we describe the methodology 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 dataset with a common spatial-temporal resolution that facilitates comparisons and evaluations. We specifically used the Tier-2 dataset 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).

In Tier-2, both forcing and model horizontal resolution are 0.25°, with data available in daily or monthly time steps from 1980 to 2014. We downloaded GHM/LSMs monthly simulated results and the respective meteorological data. The precipitation data used was from MSWEP (Beck, Van Dijk, Levizzani, et al., 2017) and the remaining meteorological data was from the ERA-Interim dataset (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 datasets. 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 (Dutra et al., 2017). Therefore, we retrieved specific datasets used by each GHM/LSM individually. Since the E2O report was not conclusive on the employed parameter datasets 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 S1.

We assumed that the topographic slope used would be the same for each model, as the differences between topographic datasets at the model scale would be small. We used 5-minute 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.

Most original land cover and soil datasets needed to be resampled to be used as input feature for the ML model. We followed a hybrid aggregation method: most dominant class for higher resolutions and class fractions for lower resolutions. Due to computational limitations, maps with higher resolution (e.g. HWSM, soil map) 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.

In addition, we had to eliminate high correlations between input variables. Highly correlated variables can interfere on the estimate of feature importance, as both can reduce errors by a similar amount. More details about the features removal on S2.

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 that each tree could go as far as necessary (i.e., the number of leaves and nodes was not limited); that each tree would only be trained with 1/3 of the total input features; and that the Random Forest would consist of 200 decision trees. 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, i.e. the modeling errors, by the highest amount. The Random Forest and Feature Importance algorithms were available in the Python sklearn 1.2.1 library.

We split the data (grid cells) into 70% for training and 30% for testing. To increase confidence in our results, we performed a robustness test by splitting the data into three different training and testing datasets and subsequently running the Random Forest algorithm with three different initializations for bootstrapping and feature selection. In total, the robustness test therefore included nine models for each combination of hydrological fluxes and GHM/LSM. This approach allowed us to evaluate the sensitivity of the Random Forest model performance and Feature Importance to randomization (See S3).

2.4. Analysis

Our goal was to identify the importance of feature groups for different GHM/LSMs. However, given natural correlation between some input features (e.g. rainforest landcover and precipitation), there remains a challenge in confirming that the differences observed between GHM/LSMs are related to their structure and not to the correlation between input features. To tackle this, we conducted a cross-feature evaluation. This consisted of training Random Forest with the input features of one GHM/LSM and hydrological fluxes from another. More specifically, the land cover and soil maps were swapped between GHM/LSMs, since the remaining input features are exactly the same. More explanation about the purpose of the cross-feature evaluation in S4.

Thus, in addition to the 'Regular Case', where the output and input features belong to the same model, we analyzed a 'General Case' – which consists of every possible combination of GHM/LSM input and output, and the 'Cross Case' – where we averaged the features importance of different land cover and soil maps. In formula:

- Regular case:

$$FI_i = f_{ii}(\,Out_i\,, In_i) \qquad\qquad i = \;1 \ldots 5 \;:\; models$$

- General case:

$$FI_{ij} = f_{ij}\left(Out_i\,, In_j\right) \qquad\qquad i,j = \;1 \ldots 5 \;:\; models$$

- Cross case:

$$FI_i = \frac{1}{5}\sum_{j=1}^{5} f_{ij}\left(Out_i, In_j\right) \qquad\qquad i,j = \;1 \ldots 5 \;:\; models$$

Where Out and In are related to the outputs (hydrological fluxes) and input features, respectively. f is the function to calculate the features' importance – FI (from the Random Forest fitting). The methodological scheme can be visualized in Figure 1.

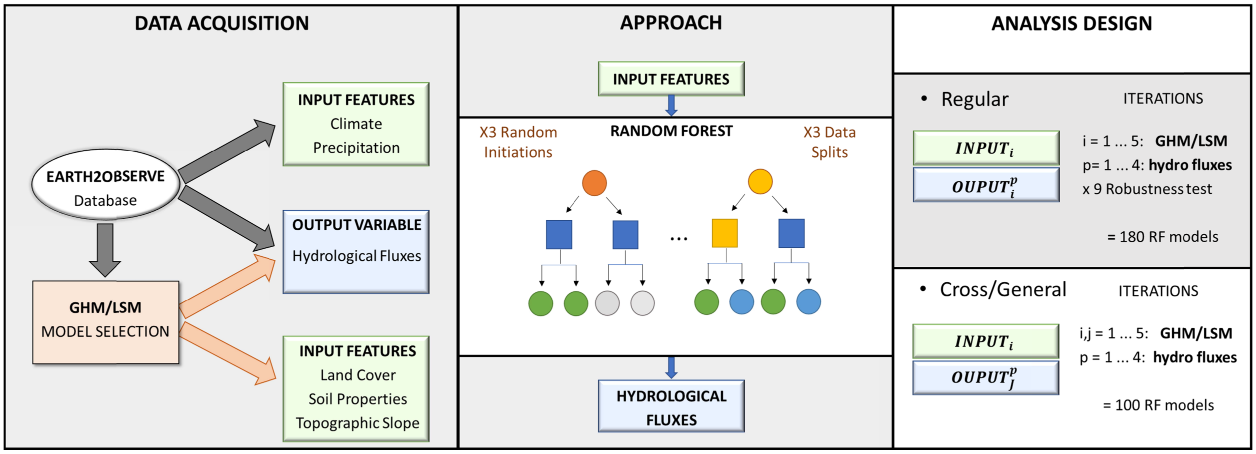

Figure 1. Schematic diagram to explain the methodology. Data Acquisition: obtaining input and output data from the GHM/LSMs either from the E2O database (meteorological) or independently (soil, land cover and topography). Approach: using the input data as predictor and output data as target variables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

3. Results and Interpretation

3.1. Random Forest Performance

In general, the performance of the Random Forest 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). $R^2 > 0.85$ for all input/output combinations and, most of the cases, $R^2 > 0.98$ on the predictions of average evaporation (*Evap*) and average runoff (*Q*). *Evap* and *Q* are key components of the water balance, so it could be easily replicated by RF models using long-term averages of meteorological variables. 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. This is reflected in a slightly lower RF performance, although still showing quite acceptable R^2 values.

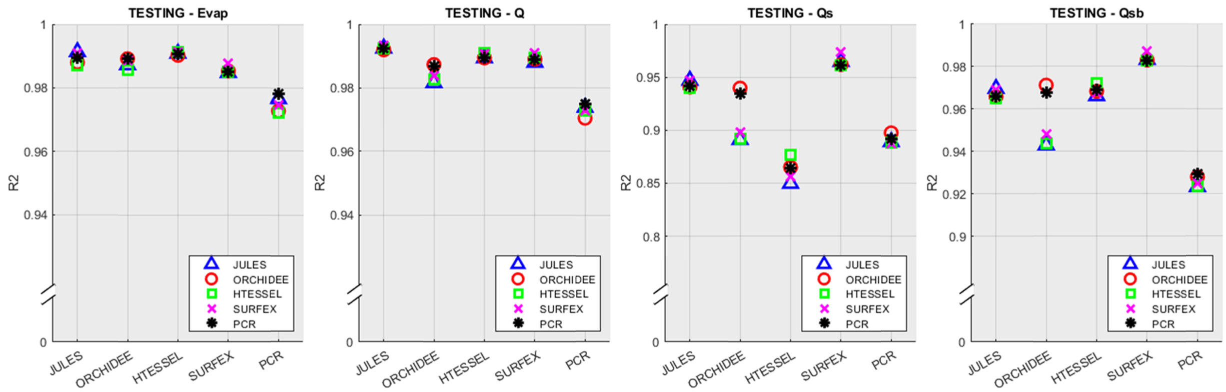

Figure 2. Performance of the Random Forest models in terms of R^2 for the testing set. Each chart represent a different hydrological flux. Left to right: Evaporation, Runoff, Surface and Subsurface Runoff. The hydrological fluxes were calculated from different GHM/LSMs outputs (x-axis). The symbols and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

The RF model presented the best performance when both output and input features were from the same GHM/LSM (Regular Case), i.e. x axis label and symbol from the same GHM/LSM, which confirms our input selection (S1). 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, and this can be explained by two reasons: 1) precipitation and climate importance are generally higher than 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; 2) different soil and land cover databases still present similar spatial patterns, so the overall RF performance decreases just in a small percentage when provided with a different predictor source.

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 top soil layer map (S1), and soil features have high importance in predicting ORCHIDEE's *Qs* and *Qsb* (see next section).

In addition, it seems that 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 details 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

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, General Case, and 2 other cases to guide the analysis. The Regular Case represents the ideal case where the RF was trained with input features and outputs from the same GHM/LSM. We also calculated the average importances from different soil and land cover maps, named Cross Case. 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 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).

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 to 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).

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 General Case, Land Cover is predominant in each of the first of each group of columns, 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 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 is correlated with other feature groups.

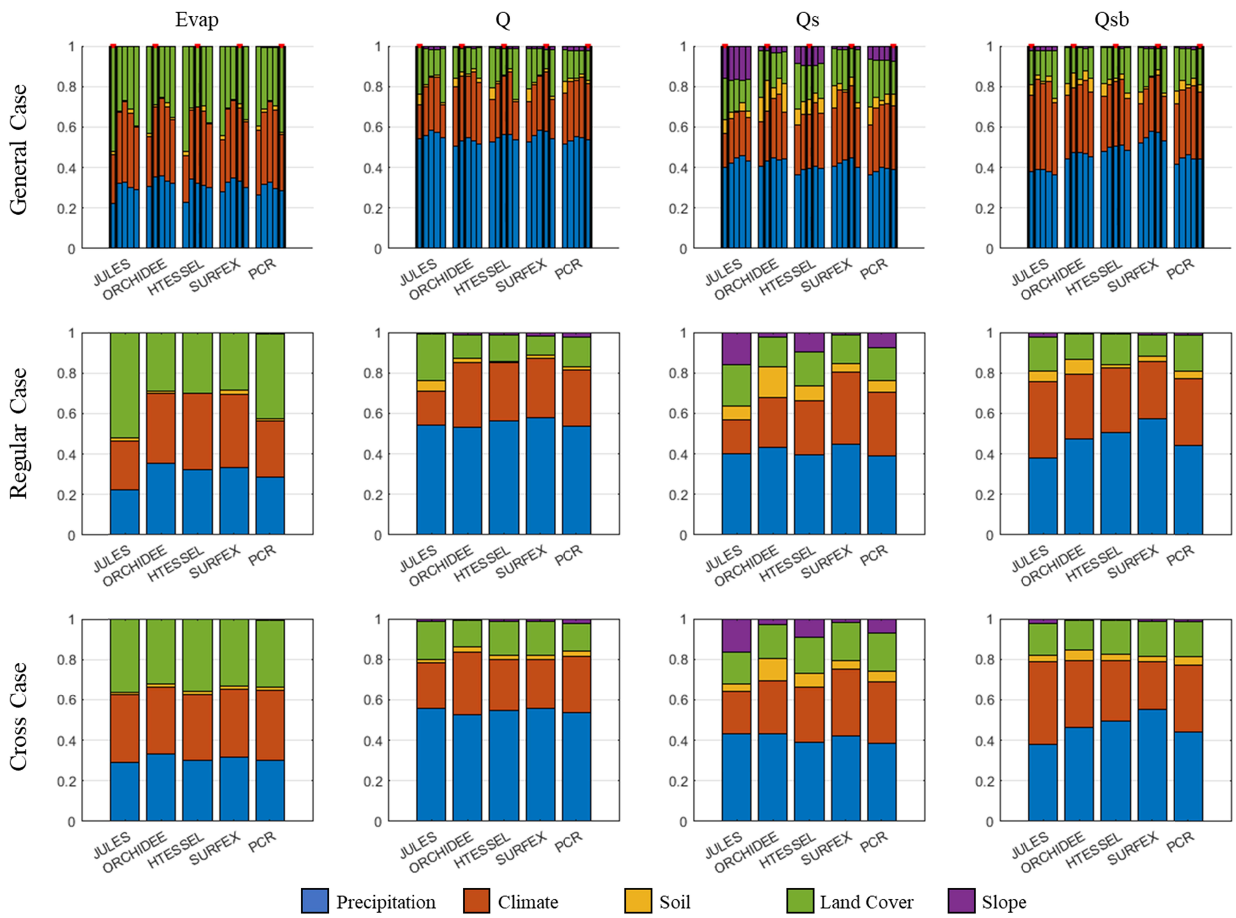

Figure 3. Feature importance of five feature groups (Precipitation, Climate, Soil, Land Cover and Slope) for the prediction of four hydrological fluxes (Evap, Q, Qs and Qsb) by RF models given different GHM/LSM as the source of input (predictors) and output (predictand) data. The General Case consider all the possible combination of GHM/LSM input and output data. The group columns indicated on the x-axis refers to the GHM/LSM that provided the output and each single column indicate the GHM/LSM that provided the input always in the following order: JULES, ORCHIDEE, HTESSEL, SURFEX and PCR-GLOBWB. The Regular Case is represented by the column marked in red on the top which indicates the inputs and outputs data from the same GHM/LSM. The Cross Case is the average of the group columns of the General Case.

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). Indeed, slope is directly related to surface runoff generation. A hilly terrain contributes to a convergent subsurface flow (Anderson & Burt, 1978) and consequently to a greater saturated zone for overland flow (Dunne & Black, 1970). Consequently, slope has been proven to be a significant predictor as (Addor et al., 2018) showed that it is highly correlated with the runoff ratio and (Beck et al., 2015) presented slope as the third most important predictor for the flow duration curve (Aridity Index and mean precipitation were first and second, respectively). In fact, 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 termed "interflow" (Van Beek & Bierkens, 2008). ORCHIDEE and SURFEX do not seem to consider slope effects on surface runoff generation, at least for the E2O project. As an example the spatial difference between JULES and SURFEX in terms of surface runoff and precipitation ratio (*Qs/P*) is shown in Figure 4 (e.g. Andes Cordillera).

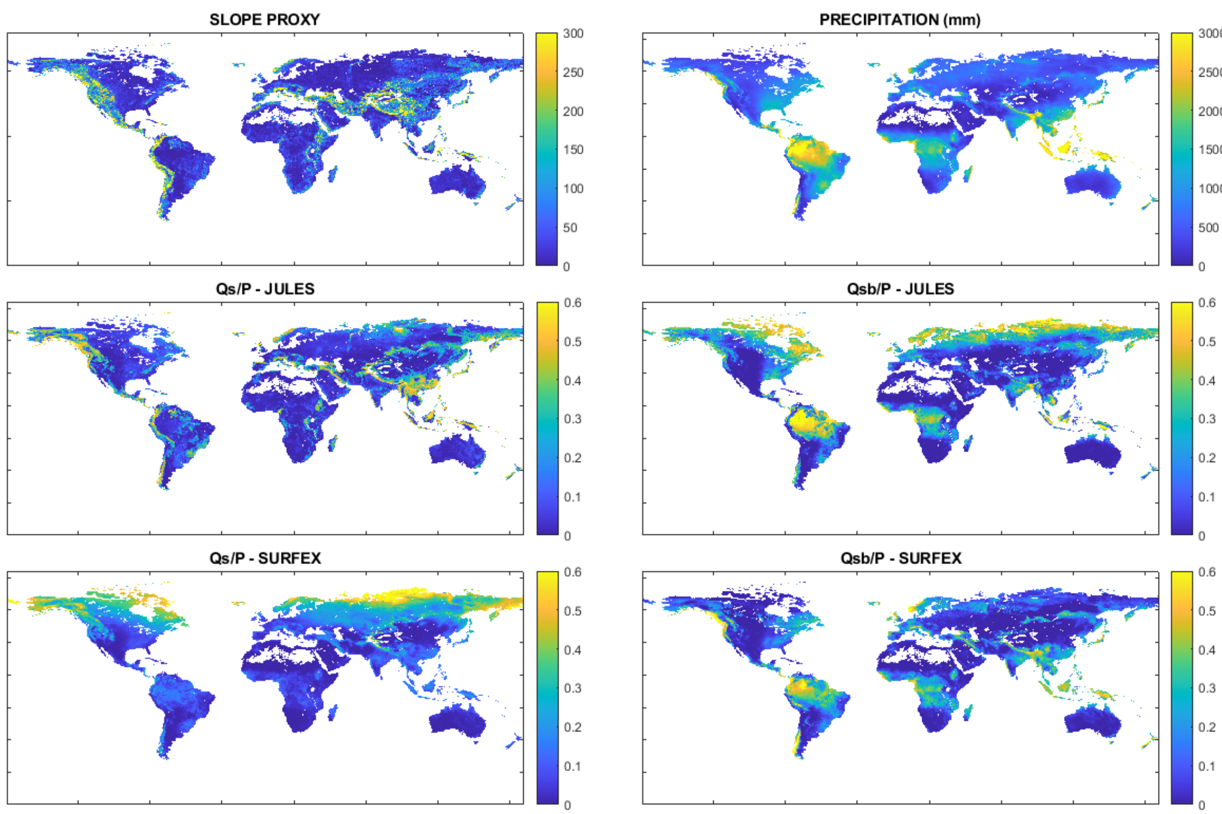

Figure 4. Global maps of the studied domain presenting the Slope Proxy, Annual Precipitation (mm), Surface and Subsurface Runoff Precipitation ratio (Qs/P and Qsb/P) estimated with outputs from JULES and SURFEX.

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. Soil importance for ORCHIDEE was 15% for *Qs* and 7.5% for *Qsb*, double the importance for the second-ranked GHM/LSM (*Qs* – HTESSEL, *Qsb* – PCR-GLOBWB). For ORCHIDEE in particular, the feature importance shown by the Regular Case is more suitable since the RF performance considerably declined when using soil maps of other GHM/LSMs (Figure 2). (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. 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 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. Conclusion

This paper proposed a novel model intercomparison study to quantify and visualize differences between GHM/LSMs regarding the importance of different inputs on hydrological simulations, many of which could be interpreted in the context of model structure. We presented a practical method of comparing global models with 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 feature groups (precipitation, climate, soil texture, land cover and topographic slope) on explaining the variance of mean evaporation, runoff, surface runoff and subsurface runoff worldwide. In general, GHM/LSMs agree on the importance of features 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 degree of that importance.

We noticed that soil maps are 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 RF 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. Our analysis 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 analyses such as the one presented here.

5. Acknowledgements

We thank Agnès Ducharne, Emanuel Dutra, and Alberto Martinez-de la Torre for leading to the right soil and land cover input datasets of ORCHIDEE, HTESSEL and JULES respectively. And we thank the SURFEX developers for making the “physiographic maps” easily available online. We also thank Rafael Fontana for the discussion about the correlation impact on feature importance.

6. Open Research

The original meteorological data and outputs from the GHM/LSMs were obtained from the Earth2Observe project (Dutra et al., 2017). Processed data of all predictors and target variables, including the codes for making figures from the main text and Supplement Material are available at https://zenodo.org/record/8379355.

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

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

Van Beek, R. L. P. H., & Bierkens, M. F. P. (2008). *The Global Hydrological Model PCR-GLOBWB: Conceptualization, Parameterization and Verification Report*. Utrecht, the Netherlands. Retrieved from http://vanbeek.geo.uu.nl/suppinfo/vanbeekbierkens2009.pdf

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

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

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

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*. 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

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

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

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

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, March 1). Modelling Freshwater Resources at the Global Scale: Challenges and Prospects. *Surveys in Geophysics*. Springer Netherlands. https://doi.org/10.1007/s10712-015-9343-1

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., & 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*.

Fisher, R. A., & Koven, C. D. (2020, April 1). Perspectives on the Future of Land Surface Models and the Challenges of Representing Complex Terrestrial Systems. *Journal of Advances in Modeling Earth Systems*. Blackwell Publishing Ltd. 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

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

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.

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 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

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, March). A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system. *Global Biogeochemical Cycles*. 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). https://doi.org/10.1088/1748-9326/aa8359

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

Le Moigne, P. (2018). *SURFEX SCIENTIFIC DOCUMENTATION*.

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, March 1). What Role Does Hydrological Science Play in the Age of Machine Learning? *Water Resources Research*. Blackwell Publishing Ltd. 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*. https://doi.org/10.1029/2011WR011527

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, November 1). A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists. *Water Resources Research*. Blackwell Publishing Ltd. https://doi.org/10.1029/2018WR022643

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, June 24). 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*. Copernicus GmbH. 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

Vereecken, H., Amelung, W., Bauke, S. L., Bogena, H., Brüggemann, N., Montzka, C., et al. (2022, September 1). Soil hydrology in the Earth system. *Nature Reviews Earth and Environment*. Springer Nature. 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.

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*. 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). 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*. 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, July 1). Ensemble machine learning paradigms in hydrology: A review. *Journal of Hydrology*. Elsevier B.V. https://doi.org/10.1016/j.jhydrol.2021.126266