Reply to Reviewer 1

Thank you very much for taking the time to review our manuscript and for providing us with constructive feedback. We have carefully addressed all the points raised and will incorporate the necessary changes into the revised manuscript. Please find below the main points to be corrected:

1. **The title, the abstract, and the conclusions are not fully aligned. In the title authors mention to "prediction", in the abstract "interpretation", in the conclusion there are many points mixing the two topics.**

   We agree that the concepts in these sections are somewhat mixed and could be clarified. To address this issue, we have revised the abstract and conclusion to better align with the title, which emphasizes "prediction." We specifically highlighted how the study's results can contribute to improving prediction models.

   Here are the following changes we will make:

   Lines 13-14: *This study demonstrates the potential of causal inference techniques for predicting catchment responses by effectively representing the interconnected processes within hydrological systems in a more interpretable manner.* ~~*This study demonstrates the potential of causal inference techniques in representing the interconnected processes in hydrological systems in a more interpretable and effective manner.*~~

   From line 479:

   Three prediction models, BN, GAM, and RF, in five different settings, namely BN, GAM∼Par, GAM∼All, RF∼Par and RF∼All, are used to predict runoff signatures. These models are executed on the entire dataset as well as 22 clusters, with each configuration undergoing 100 random samplings of training and test sets, resulting in a total of 11,500 model executions.

   From line 485:

   - *The causal parents of the signatures identified by the PC algorithm do not always align with the most influential variables determined by correlation and variable importance analysis. This suggests that strong correlations may result from confounding variables, and causal relationships do not always coincide with high variable importance. This point can impact the robustness of prediction models, especially when the same set of predictor variables is used across diverse environments with varying characteristics.*
   - *BN shows the smallest decrease in accuracy between the training and test samples, demonstrating high transferability. The accuracy of the models is not sensitive to the training sample size and shift in the distribution of predictors. This indicates that $P(Effect \mid Cause)$ remains consistent across environments. Although BN's overall accuracy is lower than that of the nonlinear GAM and RF models, it outperforms RF in predicting mean daily runoff and high flows across different environments (clusters).*
   - *Using causal parents helps mitigate the overfitting problem and improve the robustness in prediction models, particularly in GAM, when the size of the training set is small.* ~~*Using causal parents helps reduce overfitting, particularly for GAM, when the training sample size is small.*~~
   - *The high accuracy of non-causal models, GAM∼All and RF∼All, in the baseline scenarios may be attributed to spurious relationships. This is supported by their reduced accuracy in environments with smaller training sets, highlighting a lack of robustness compared to causal models, which maintain higher reliability under such conditions.* ~~*The high accuracy of non-causal GAM∼All and RF∼All in the baseline models may be due to spurious relationships, as their accuracy decreases in environments with smaller training sets compared to the causal models.*~~
   - *In environments where the target signature is more difficult to predict, such as clusters of the geology category, using causal parents increases prediction accuracy.*

- *Independent variables identified through causal discovery can determine groups of catchments where prediction models exhibit consistent performance. For instance, topographic variables are among the independent variables in this context since all models perform consistently well in clusters 1, 2, and 3, and less effectively in cluster 4. This information helps identify environments where training models achieve higher accuracy, reduced uncertainty, and greater robustness.* ~~*The independent variables identified through causal discovery using DAGs can serve as reliable criteria for catchment classification. This is evident from the models performing consistently well in clusters 1, 2, and 3, while performing less effectively in cluster 4. This information improves model accuracy, reduces prediction uncertainty, and enhances consistency between training and test simulations.*~~
- *Causal inference methods contribute to improving prediction models'* ~~*model*~~ *parsimony, interoperability and robustness in hydrological systems* ~~*modelling*~~*.*

2. **Runoff signature is synonymous of "hydrological response" or "watershed response"? Maybe in the introduction this other common term could be mentioned just to better orient the reader.**

   Runoff signatures represent distinct characteristics of a catchment's response. While "catchment response" and "runoff signature" are sometimes used interchangeably, this could create confusion for readers. In the revised sections, we will carefully use "catchment response" to ensure clarity and avoid any ambiguity.

3. **In the lines 103-113 it should be clarified which is the innovative contribution or the advancement compared to the previous literature accurately listed by the authors**

   We appreciate this reviewer's remark. We have revised the section to clearly highlight the novelty of our work and provide more context. Specifically, we emphasized that the use of causal discovery to identify causal links between catchment attributes, climatic indices, and runoff signatures and integrating these findings into prediction models represents the key innovation of this research.

   To our current knowledge, the proposed study is the first analysis that connects the causal models and catchment attributes. Therefore, we will change this section to:

   This study introduces a novel approach for predicting runoff signatures by integrating causal information into predictive models. To the best of our knowledge, causal inference techniques have not yet been applied for this purpose. Unlike previous studies that primarily rely on correlated-based features for predicting a specific catchment response, we take a step beyond mere correlation by focusing on causally relevant variables, specifically, causal parents. By integrating causal information into predictive models (GAM and RF), we aim to investigate whether it can enhance the prediction models' robustness, interpretability, and parsimony compared to models that do not utilize causal insights. ~~*This study aims to represent the causal relationships between catchment attributes, climate characteristics and runoff signatures. We compare the performance of prediction models (GAM and RF) that incorporate causal information derived from causal discovery methods against models that do not.*~~ *We assume that a specific characteristic of catchment response is directly influenced by a subset of correlated variables, known as causal parents, rather than by all correlated variables.* ~~*runoff signatures are causally influenced by a subset of variables, known as causal parents, rather than by all available variables. We adopt the Peter and Clark (PC) causal discovery method (Spirtes et al., 2001), which is a constrained-based causal discovery algorithm, to identify these causal relationships and to structure the BNs. Our objective is to investigate whether incorporating causal information can provide new insight into hydrological systems modelling, enhance the prediction models' robustness, and improve their parsimony.*~~ *To achieve our objectives, we follow these steps: 1) select potential predictors for each runoff signature among the catchment and climate attributes, 2) identify causal relationships between catchment attributes, climate characteristics, and runoff signatures (network structure) using Peter and Clark (PC) causal discovery method (Spirtes et al., 2001),* ~~*2) identify causal parents and network structure for each signature,*~~ *3) execute models using both the causal parents (causal models) and all selected variables (non-causal models) for entire catchments and subset of catchments, 4) evaluate the robustness of the causal and non-causal models.*

4. **Section 3. Data are crucial for understanding the model application. In the Section 3 there is the attribute list but not the data characterization. A first question that**

could have the reader is "Did they authors select one number for each attribute and for each catchment?" or a time series?

Thank you for raising this point. We will improve Section 3 and include additional information about data characterization to enhance clarity. We will also mention that we do not use the time series data in our analysis. Each catchment in the dataset has five categories of attributes that are outlined in Table 1. We will use this information to categorize catchments (using cluster analysis) based on each category. Therefore, each catchment has been assigned five cluster IDs.

The changes will be as follows:

From line 219:

The clustering classifies the catchments according to the five categories. Time series data is not used for clustering analysis, and only catchment attributes available in the CAMELS dataset, as listed in Table 1, are utilized for this purpose. Table 3 shows the methods used for clustering, the optimum number of clusters according to the elbow and Silhouette scores, and the number of catchments in each cluster.

From line 220:

(a) ***Climate attributes:*** *Climate attributes in the CAMELS dataset are derived from area-weighted averaging of meteorological forcing time series from October 1, 1989, to September 30, 2009. The cluster analysis shows four distinct climate categories, which spread in the east (cluster 1), the Midwest (cluster 2), the west (cluster 3) and the northwest (cluster 4) (Fig. 2a). The largest group of catchments belong to cluster number one, with 334 in the north- and southeast of the US (Table 3). This cluster receives an average of 3.5 mm daily precipitation and has 2.8 mm daily evapotranspiration. Other clusters have the following average precipitation and evapotranspiration levels: Cluster 2 has 2.3 mm of precipitation and 2.7 mm of evapotranspiration, Cluster 3 has 5.5 mm of precipitation and 2.4 mm of evapotranspiration, and Cluster 4 has 2.0 mm of precipitation and 3.3 mm of evapotranspiration.*

(b) ***Soil attributes:*** *The soil properties data, derived from the State Soil Geographic Database (STATSGO), provides information about the top 2.5 meters of soil. Soil texture is represented in 16 classes, of which there are 12 classes based on the United States Department of Agriculture (USDA) and 4 non-soil classes. The saturated hydraulic conductivity and soil porosity are calculated based on the sand and clay fraction using multiple regression analysis. Cluster analysis identifies six groups of catchments. ~~This category is divided into 6 groups.~~ There is no distinctive spatial pattern among soil clusters. However, clusters 2 and 3 are mostly spread across the east and west coastlines (Fig. 2b). The maximum water content and porosity values are influenced by soil texture, which defines the proportion of sand, clay, silt, and other materials. For example, cluster 6 shows the highest soil porosity and maximum water content (Fig. 2b). This cluster has the highest percentage of clay (26%) and silt (47%) fractions among all clusters.*

(c) ***Topographic attributes:*** *The topographic information of catchments, namely catchments' contours, are determined using geospatial fabric (Viger and Bock, 2014) and Geospatial Attributes of Gages for Evaluating Streamflow (GAGES II) methods (Falcone, 2011). These methods are used to determine the area, and the Digital Elevation Model (DEM) is clipped for each catchment. This category is divided into 4 distinctive clusters ~~groups~~. Cluster 1 contains catchments located in the northeast, which are catchments with low elevation and slope (Fig. 2c). Cluster 2 consists of catchments along the west coast spread from the west to the northwest. The catchments with the lowest elevation and slope are in cluster 3, located in the southeast. Cluster 4 contains the highest elevation catchments in the Rocky Mountains (Fig. 2c).*

(d) ***Geological attributes:*** *The geological variables in the CAMELS datasets are derived from the Global Lithological Map (GLiM) (Hartmann and Moosdorf, 2012) and the Global HYdrogeology MaPS (GLHYMAPS) (Gleeson et al., 2014). From the GLiM dataset, sixteen lithological classes are identified, and their proportional areas are calculated for each catchment. The GLHYMAPS dataset is used to estimate subsurface permeability and porosity (Addor et al., 2017). This category is divided into 7 groups. Unlike the climate and topography categories, this category does not show a distinguishable spatial pattern (Fig. 2d). However, the catchments with the highest geological porosity are mainly concentrated in the southeast, and those with the lowest are located in the west (Fig. 2d).*

(e) ***Vegetation attributes:*** *Vegetation is represented using two indicators, vertical density, measured by the Leaf Area Index (LAI), and horizontal density, measured by the Green Vegetation Fraction (GVF). These measurements are derived from a 1-km resolution product of the Moderate Resolution Imaging Spectroradiometer (MODIS). The vegetation or land cover category is divided into 6 different groups (Fig. 2e). The spatial pattern of the vegetation is influenced by climate and topographic categories. According to Fig. 2e, the catchments with the highest forest fractions have the highest maximum leaf area index and are located in the northeast and east of the study area. This area has high precipitation and low evapotranspiration (Fig. 2a). The lowest vegetation cover belongs to the central and southern parts of the US, which are in clusters 4 and 6.*

5. **Figure 1 is not fully clear, Is the cluster analysis necessary? Is it an alternative way to analyze the entire data set? If yes it should be in a different level, like a starting option in the flow chart.**

   Regarding Figure 1, we prefer to use Figure 1 to highlight the core of the analysis, which presents the description of the lists of steps of the proposed comparative analysis on causal models and catchment runoff signatures, rather than a direct flowchart. In this figure, cluster analysis uses the entire dataset and assigns cluster IDs to each catchment. We will modify this figure to make it more understandable as follows:

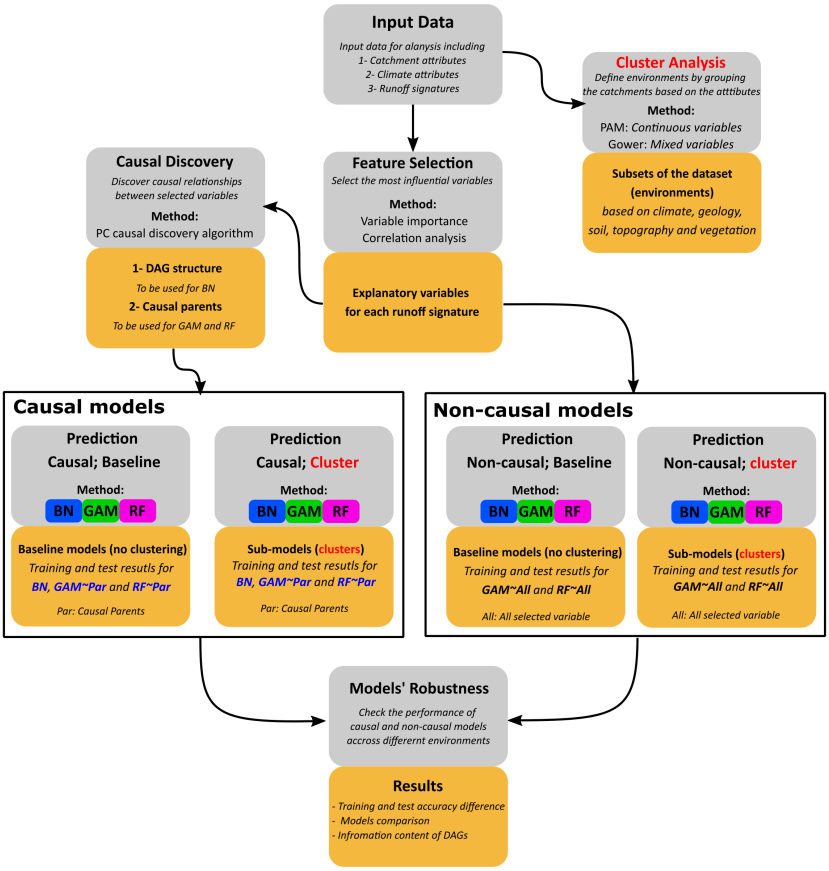

Figure 1: *Flowchart depicting the steps followed in this study. Grey boxes indicate the procedures, orange boxes present the results of these procedures,~~and~~ blue text highlights where information about causality is utilized , and the red text highlights the cluster analysis and indicates where the clustering results are applied.. PC refers to Peter and Clark's causal discovery algorithm, PAM stands for Partition Around the Centroid clustering algorithm, and DAG refers to Directed Acyclic Graph. BN refers to the Bayesian Network, GAM refers to the Generalized Additive Model, and RF refers to the Random Forest. GAMPar and RFPar are causal models (GAM and RF) using only causal parent variables for prediction, while GAMAll and RFAll are non-causal models that use all selected variables as predictors. Baseline models refer to models that use the entire dataset (all 671 catchments) for training and testing, while sub-models use only subsets of the dataset or clusters.*

Regarding the necessity of clustering, When using the entire dataset for training and testing the models, we work with a large, diverse dataset containing a variety of characteristics. However, the question arises: what happens when we focus on regions with fewer catchments that share similar attributes, such as climate, geology, vegetation, topography, or soil properties? To investigate this, we applied clustering analysis to group catchments with similar characteristics, creating categories of varying sizes.

This approach enables us to evaluate the performance of causal and non-causal models across different environments, identifying the situations where the models perform well and where they face challenges. For instance, if the models perform well in a cluster within the topographic category, we can infer that models are effective in regions with specific topographic characteristics and less sensitive to the sample size. Additionally, using causal parents as predictors establishes a causal mechanism for the runoff signature under the assumption that this mechanism remains consistent across different environments. Clustering helps test this assumption and assess the effectiveness of the PC algorithm in identifying the appropriate causal parents using the models' performance. In conclusion, clustering is a complementary rather than an alternative analysis to investigate the causal models further.

6. **The Section 2.2.1 seems incomplete and refers to the Supplementary materials, how-**

ever this step seems important in the whole procedure. More details on how the most important feature are ranked are necessary, indeed the "out-of-bag method" is vague and the sentence "variables are selected based on a combination of correlation analysis, variable importance assessment and consideration of the underlying physics of the runoff signatures." is too general.

We agree that the information provided in this section is not sufficient. Therefore, we have added more details to the feature selection section. We will explain that the out-of-bag method is used as a complementary method to the correlation analysis, which helps identify explanatory variables in categories with low correlation coefficients.

The changes will be as follows:

From line 145:

(a) *Correlation analysis: Pearson, Kendall, and Spearman correlation coefficients are computed to illustrate the potential explanatory variables. The correlation analysis reveals the most influential variables from each category, namely climate, geology, vegetation, topography and soil. In addition, the scatter plot of the data helped visually understand the relationship between variables.*

(b) *Variable importance: Since the results of the correlation analysis are not always consistent, another feature selection procedure is conducted using the random forest method to investigate the feature importance. ~~The same approach as correlation is repeated, using the random forest method to investigate the feature importance. Random forest is implemented using the R package randomForest (Liaw et al., 2015).~~ The variables are ranked using ~~according to~~ the out-of-bag method, which is quantified using the Mean Decreased Accuracy (IncMSE) score. The out-of-bag method ranks variables based on the increase in prediction error caused by removing each variable from the prediction process. Random forest is implemented using the R package randomForest (Liaw et al., 2015).*

With the information provided by the procedures mentioned above, variables are selected based on a combination of correlation analysis, variable importance assessment and consideration of the underlying physics of the runoff signatures. We tried to select the most influential variables from each category, including climate, geology, soil, topography, and vegetation. The number of selected variables varies across categories. Multiple variables are selected from categories where most variables exhibit high correlation. Conversely, only the highest-correlated variable is chosen for categories with a weak correlation to the runoff signature of interest. For example, climatic variables are often highly influential for runoff signatures, leading to the selection of multiple variables. In contrast, geological variables tend to show a weak correlation with some runoff signatures, so only the most influential variable from this category is selected. The results of feature selection are presented in the supplementary materials.

References

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, HYDROLOGY AND EARTH SYSTEM SCIENCES, 21, 5293–5313, https://doi.org/10.5194/hess-21-5293-2017, 2017.

Falcone, J. A.: GAGES-II: Geospatial attributes of gages for evaluating streamflow, Tech. rep., US Geological Survey, 2011.

Gleeson, T., Moosdorf, N., Hartmann, J., and van Beek, L. P. H.: A glimpse beneath earth's surface: GLobal HYdrogeology MaPS (GLHYMPS) of permeability and porosity, GEOPHYSICAL RESEARCH LETTERS, 41, 3891–3898, https://doi.org/10.1002/2014GL059856, 2014.

Hartmann, J. and Moosdorf, N.: The new global lithological map database GLiM: A representation of rock properties at the Earth surface, GEOCHEMISTRY GEOPHYSICS GEOSYSTEMS, 13, https://doi.org/10.1029/2012GC004370, 2012.

Liaw, A., Wiener, M., Breiman, L., and Cutler, A.: Package 'randomforest', 2015.

Spirtes, P., Glymour, C., and Scheines, R.: Causation, prediction, and search, MIT press, 2001.

Viger, R. and Bock, A.: GIS features of the geospatial fabric for national hydrologic modeling, US Geological Survey, 10, F7542KMD, 2014.