1 Comments on the authors' response:

1) First of all, I want to point out that the output of the PC algorithm is a CPDAG, i.e., the Markov equivalence class. A CPDAG is a graphical representation of a set of DAGs where the distribution satisfies the Markov property relative to every single DAG in that set. This means that given a distribution which satisfies the Markov property for one DAG, there could also be several other DAGs for which the Markov assumption is satisfied. In other words, given a distribution, there are several DAGs where the distribution fulfills the Markov assumptions with respect to those DAGs. When orienting undirected edges to obtain a DAG from this set of DAGs, you should be careful not to introduce new unshielded colliders / v-structures. Additionally, you can also introduce background knowledge before running the PC algorithm. See for example "Interpreting and using CPDAGs with background knowledge" (2017) or *"Constraint-based causal discovery with tiered background knowledge and latent variables in single or overlapping datasets"* (2025).

2) Depending on the specific algorithm and plot function you are using, you will obtain bi-directed edges. If you use R and the pcalg package, there are cases where the direction could not be determined (but not because of the Markov equivalence class). This leads to an invalid CPDAG, meaning that your output is not representing a Markov equivalence class. There are several reasons for this, but it is important to mention that undirected and bi-directed edges are not the same. There are at least three violations of assumptions which may lead to an invalid CPDAG: cycles, hidden common causes, and selection bias. It could be that you set a direction for an edge without realizing that you are violating something else (for example, introducing new unshielded colliders / v-structures or creating cycles).

3) You wrote: *"It is worth mentioning that we attempted to include all continuous variables in the causal discovery process without applying variable selection. This approach was tested to address the causal sufficiency assumption in the PC algorithm, which requires that all common causes of the target variables are accounted for. Despite this, we observed challenges such as the generation of disconnected DAGs with independent nodes or groups of nodes lacking causal relationships with runoff signatures."*

This should be a big red warning signal for your analysis and needs further investigation. The output of the PC algorithm heavily depends on the alpha value, this is the significance level for the tests. Have you optimized this value somehow? How does the results depends on alpha? Additionally, it seems strange to me to perform feature selection methods before applying the PC algorithm. The PC algorithm also estimates Pearson correlations, and the first step of the PC algorithm (learning the skeleton) is based purely on conditional independence tests. These tests are, for example, partial correlation for two variables conditional on a third one, or in the first step just a correlation between two variables. Why should you use an extra step of correlation analysis if the PC algorithm will do the same and additionally will save the information on potential unshielded colliders / v-structures?

4) Furthermore, the authors are mainly interested in discovering the parent set of the outcome Y. Why are you not considering methods which are designed for this task? I mean Invariant Causal Prediction (ICP). I would spend a bit more time on the limits of causal discovery and non-linear methods. I would recommend reading:

1. Model-Based Causal Feature Selection for General Response Types (2024)
2. Invariant Causal Prediction for Nonlinear Models (2018)
3. Causal inference by using invariant prediction: identification and confidence intervals (2016)

5) Furthermore, you are using GAMs without specifying which GAMs you are using. Which link function do you use? Which family of distribution is assumed for the outcome y? Generalized linear models are then useful if your outcome is discrete (Poisson, ...), binary (logistic, c-log-log, ...), or continuous but only positive valued (exponential, gamma, ...). I see why you are using the additive components (splines), but I miss some more information about the model specification. Without this information, reproducibility is not possible.

6) A further point to note is that while the PC algorithm basically only uses (partial) correlation, testing linear dependency, the subsequent use of additive models assumes a potentially non-linear association

between X and the mean of Y. This is not necessarily a flaw in your approach, but it is crucial to be aware that while the PC algorithm relies on (partial) correlations, which inherently assess linear relationships (keeping in mind the fundamental connection between linear regression and correlation), the application of GAMs (the additive part of it) implies that a non-linear relationship between the predictors and the mean of y is assumed.

What you could try is something like using regression modeling with cubic spline as a heuristic test of conditional independence. See for an data example: "Data-Driven Model Building for Life-Course Epidemiology (2021)".

7) While the paper makes a good effort to explore a new idea in prediction of hydrological responses, I think there are some areas where the authors could show a better awareness of the methods' limitations and assumptions they're using. It's important to acknowledge that this isn't entirely the authors' "fault"; causal discovery research is still developing, and it's not always straightforward to apply these methods to any kind of data. The PC algorithm, for instance, tends to work best with multivariate Gaussian data, perhaps with a few categorical variables, and when you have a large enough sample size and a well-chosen alpha value (probably only given in simulations). Applying it to more complex data with non-linear relationships and arbitrary data types can be very challenging. Therefore, I would not suggest using the PC algorithm, instead trying more task-specific methods, such as ICP mentioned above.