Local Causal Discovery and Prediction of Spring Flood Events in Swiss Catchments

Georg Khella, Applied Statistics, Project 3

March 2025

1 Introduction

Understanding the hydrological and climatological conditions that lead to flood events is crucial for risk management in alpine regions. In this study, we focus on extreme discharge events during the spring season (March–April–May), based on hydrologically simulated data for Swiss catchments.

Our objective is to investigate which meteorological and hydrological variables are most closely associated with the occurrence of flood events. For this purpose, we define a binary target variable T, which equals 1 whenever the daily discharge exceeds the empirical 90th percentile threshold over the observation period, and 0 otherwise. This implicitly assumes stationarity in the distribution of discharge, an assumption we briefly examine during the analysis.

Rather than attempting to learn the full causal structure of the system, we apply *local causal discovery* methods to identify the Markov blanket of the target variable T. This consists of its direct causes, direct effects, and the parents of its children, forming the minimal set of variables that renders all others conditionally independent of T in a causal graph.

To this end, we apply both standard variable selection techniques, specifically an ℓ_1-penalised logistic regression (LASSO), and a constraint-based local discovery algorithm, namely the PC algorithm with a high significance level. These are based on statistical tests of conditional independence and are tailored to recover the Markov blanket of a specific variable of interest.

Finally, we evaluate the predictive performance of the selected variable sets in a classification task using logistic regression. The comparison is carried out for two contrasting catchments—one located at low elevation and one at high elevation—to account for the heterogeneity in flood dynamics across different topographical settings.

2 Data Overview and Analysis

This section provides an overview of the dataset and an exploratory analysis of its distribution, including some relevant plots.

2.1 Dataset Description

The data used in this study consists of daily hydrological and meteorological variables simulated with the PREVAH model over the period 1981–2016, for a total of 13,149 consecutive days. The original dataset covers 307 Swiss catchments, each with the same temporal resolution and set of variables.

For the purposes of this project, we extracted the two catchments assigned to us: catchment 207, representing a low-elevation basin, and catchment 106, representing a high-elevation basin. For each catchment, we constructed a dataset containing the following continuous variables:

- **Precipitation** $[mm]$
- **Evapotranspiration** $[mm]$
- **Radiation** $[W/m^2]$
- **Snowmelt** $[mm]$
- **Soil moisture** $[mm]$

- **Temperature** $[°C]$
- **Discharge** $[mm]$

Each observation is indexed by date, starting from January 1, 1981. These datasets serve as the foundation for the subsequent modelling and classification tasks. The response variable will be derived from the discharge series, while all other variables will be used as predictors, including lagged versions up to three days prior.

2.2 Exploratory Data Analysis

This section provides a visual and statistical exploration of the discharge variable, which plays a central role in the construction of the binary response T_t, indicating flood events. The goal is to assess the distributional characteristics of the discharge, its seasonal behavior, and how these features differ between the two catchments under study.

Discharge Distribution and Threshold Definition To define the binary target variable T_t, we consider the empirical 90th percentile of the daily discharge over the entire period. Figure 1 shows the histogram of daily discharge values for both catchments, with the threshold indicated by a red dashed line.

The distribution is markedly right-skewed in both catchments, with most values concentrated at low levels. The 90th percentile threshold lies around 8.5 mm for catchment 207 and 5.2 mm for catchment 106, reflecting the difference in elevation and hydrological response. These histograms justify the choice of a high quantile to define flood events and reveal the rarity and extremeness of such occurrences.

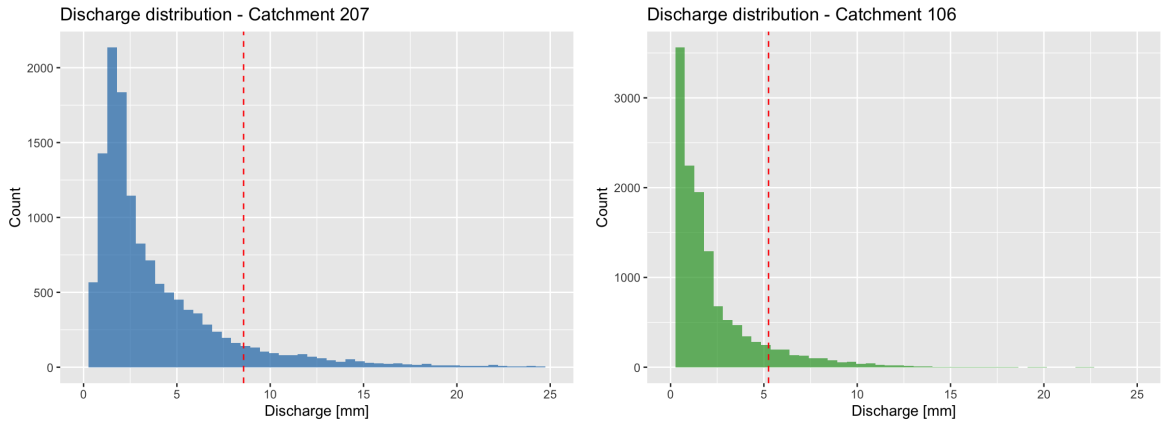

Figure 1: Daily discharge distribution for catchments 207 and 106. The red dashed line indicates the 90th percentile threshold used to define flood events.

Seasonal Patterns and Intra-Annual Discharge Dynamics To investigate the temporal behavior of discharge and the relevance of seasonal effects, we first analyze the distribution of daily discharge values by calendar month (Figure 2). The boxplots reveal that both the median and variability of discharge increase markedly during the spring and early summer months, particularly from April to June. Catchment 207 displays a wider spread and more frequent extreme outliers than catchment 106, whose discharge remains more concentrated but still exhibits seasonal patterns.

To gain additional insight into intra-seasonal dynamics, we focus on the spring months (March–May), during which flood events are most frequent. Figure 3 presents the daily spring discharge time series for both catchments, overlaid with the empirical 90th percentile threshold (dashed red line). This visualization confirms that high discharge values are concentrated in this window, with catchment 207 exhibiting more sporadic yet higher peaks, while catchment 106 displays more regular, moderate fluctuations.

The binary target variable T_t, indicating flood events, is defined using a fixed threshold corresponding to the empirical 90th percentile of daily discharge over the full observation period. This implicitly assumes stationarity in the discharge distribution. To assess this assumption, we compute a rolling 90th percentile using a one-year moving window (Figure 4). In catchment 207, substantial interannual

variation is observed, suggesting sensitivity to long-term climatic factors. Catchment 106 shows more temporal stability, though some deviations from the global threshold are still apparent.

These findings justify the use of a springtime analysis window and highlight the trade-off between interpretability and temporal precision in threshold selection. While the fixed threshold may not fully capture local trends in extreme discharge, it enables consistent comparisons across years and between catchments.

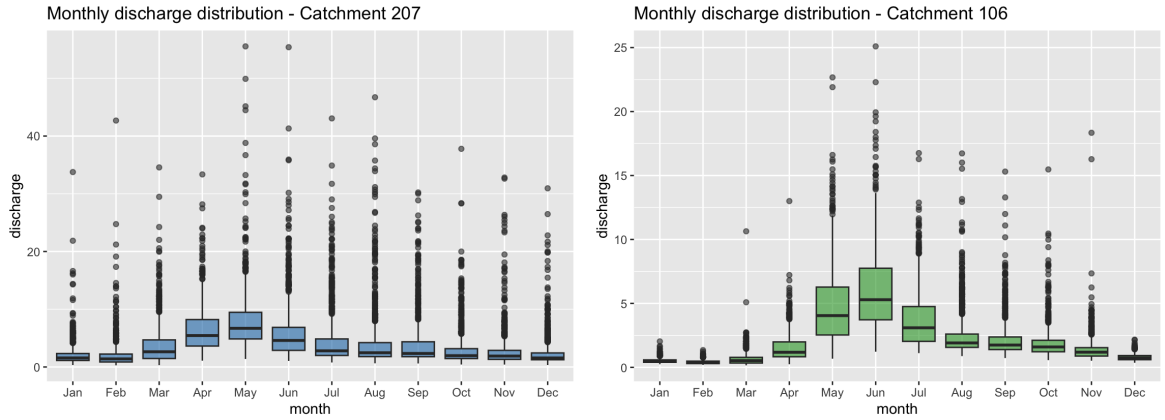

Figure 2: Monthly discharge distribution across the year for catchment 207 and 106. The spring months exhibit significantly higher discharge variability.

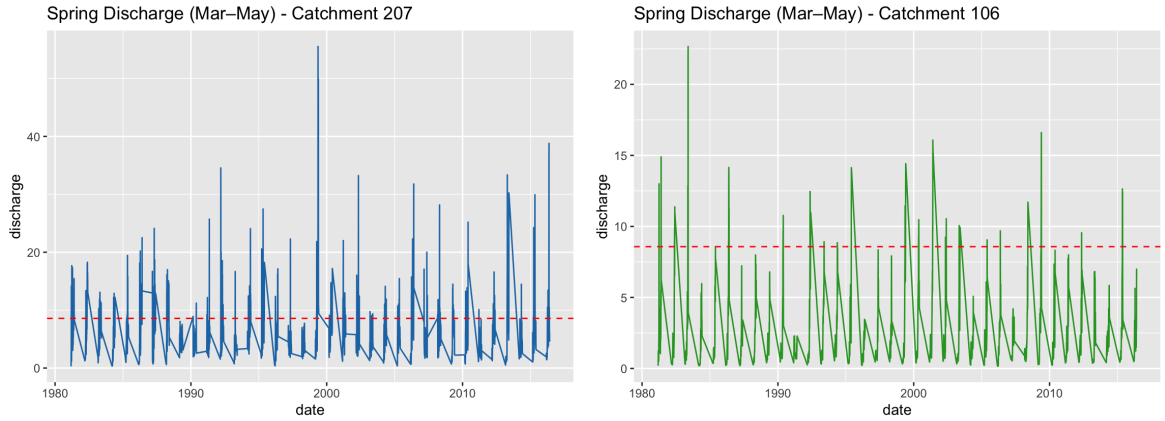

Figure 3: Spring discharge time series (March–May) for catchments 207 and 106, with the 90th percentile threshold shown as a dashed red line.

Temporal Stability of the Discharge Threshold The binary response variable T_t is defined based on a fixed discharge threshold corresponding to the empirical 90th percentile over the entire period. This implicitly assumes stationarity in the distribution of discharge values. To empirically evaluate this assumption, we compute a rolling 90th percentile using a 1-year moving window.

Figure 4 shows the temporal evolution of this rolling threshold for both catchments, overlaid with the global fixed threshold (dashed red line). The plots reveal notable fluctuations over time. In catchment 207, the rolling threshold exhibits marked interannual variability, suggesting that discharge extremes are influenced by decadal-scale climatic variability. Catchment 106, while more stable, still shows deviations from the fixed threshold.

Although some degree of non-stationarity is present—especially in catchment 207—we adopt a fixed threshold definition of T_t for interpretability and comparability purposes. The rolling analysis highlights the limitations of this simplifying assumption and helps contextualize the robustness of the derived conclusions.

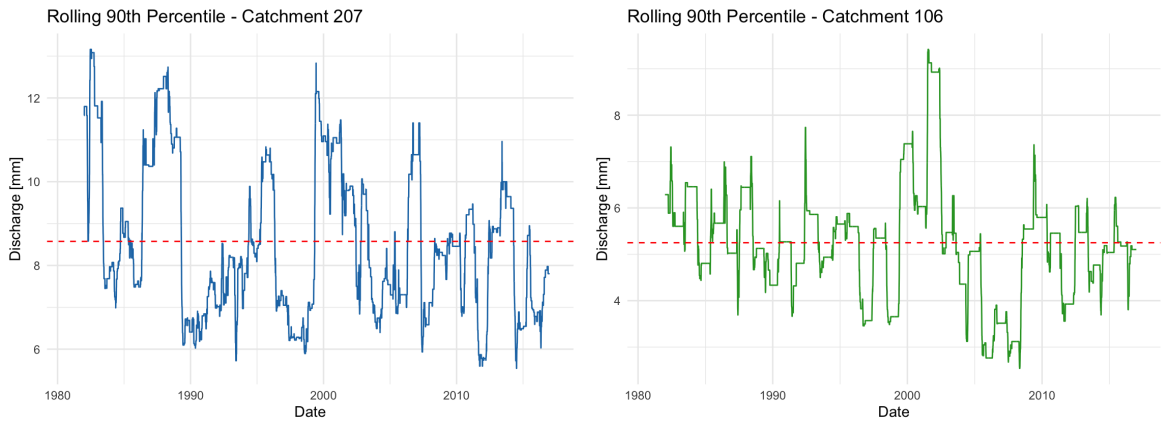

Figure 4: Rolling 90th percentile of discharge using a 1-year window for catchment 207 and catchment 106, with the 90th percentile threshold shown as a dashed red line.

Marginal Distributions of Covariates To gain insight into the distributional characteristics of the meteorological and hydrological covariates, we examine their marginal densities separately for the two catchments. Figure 5 displays kernel density estimates for the six covariates under consideration: evapotranspiration (et), precipitation (precip), radiation, snowmelt, soil moisture, and temperature.

The plots highlight notable differences in the distributions across variables and between catchments. Several covariates related to evapotranspiration, precipitation, and snow-related processes exhibit strong right-skewness, with a large concentration of values close to zero. In contrast, the variable associated with air temperature shows an approximately symmetric distribution. Differences in skewness patterns are also evident in variables linked to soil and radiation inputs. In particular, soil moisture levels in Catchment 207 tend to cluster near their upper bound, suggesting consistently high saturation, whereas in Catchment 106 the values appear more dispersed.

These findings inform our later modelling choices, suggesting that variable transformations (e.g., logarithmic) may be appropriate for skewed predictors. Moreover, the differences across catchments underline the importance of treating them separately throughout the analysis.

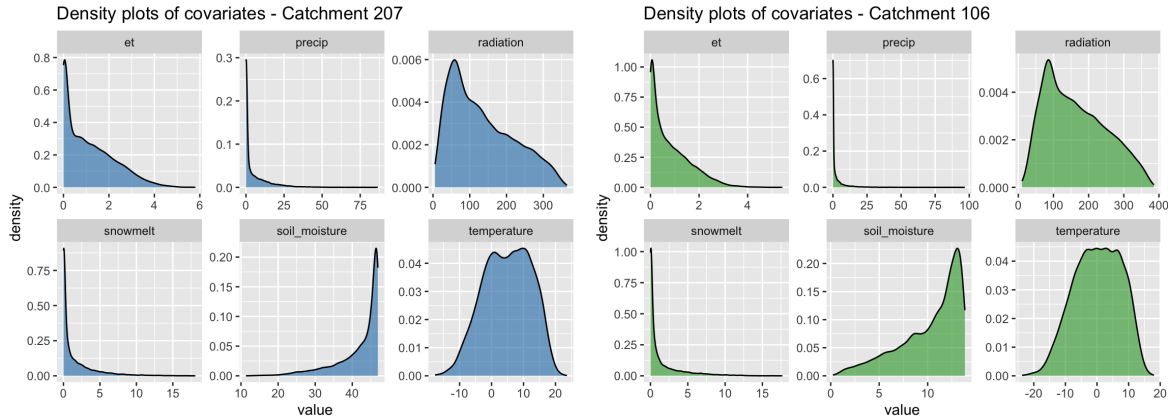

Figure 5: Kernel density estimates of meteorological and hydrological covariates in catchment 207 and catchment 106.

Flood vs. No Flood – Covariate Distributions We examine how the distributions of the meteorological and hydrological covariates differ between flood and non-flood days, as defined by the binary indicator T_t. Figure 6 presents kernel density estimates for each covariate, stratified by flood status.

Several variables exhibit clear distributional shifts between the two regimes. For example, in both catchments, flood days are associated with higher snowmelt and evapotranspiration, while precipitation is sharply right-skewed and concentrated near zero in both cases. Soil moisture in catchment 207 appears to be nearly saturated during flood events, in contrast to catchment 106. These patterns highlight the

relevance of conditional distributions for understanding the mechanisms underlying flood events.

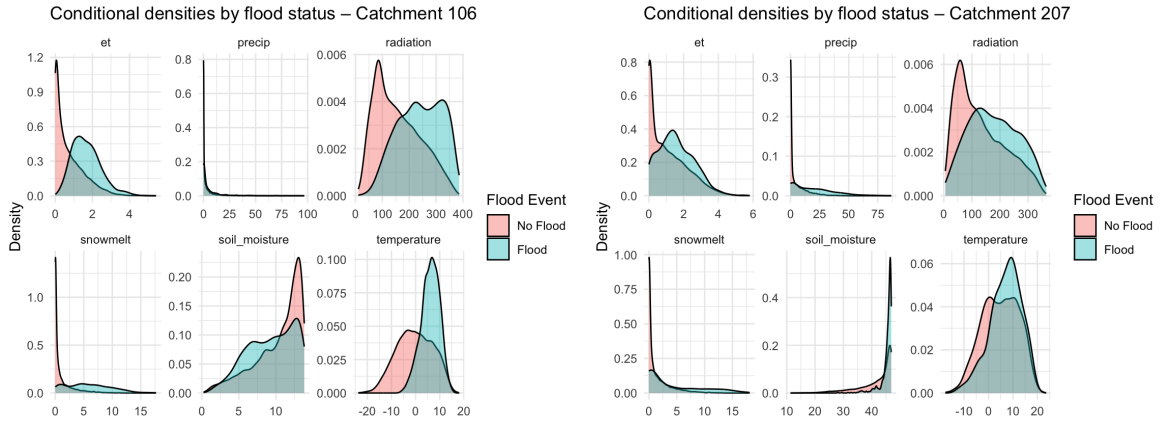

Figure 6: Conditional density estimates of the six covariates by flood status, for catchment 106 and catchment 207.

Autocorrelation Structure of the Covariates To evaluate the temporal dependence of the covariates and motivate the inclusion of lagged predictors in the modeling phase, we analyze their autocorrelation functions (ACFs) up to 30 days. Figure 7 displays the ACF plots for each covariate and catchment.

All variables exhibit significant autocorrelation, especially snowmelt, evapotranspiration, and soil moisture, which retain memory over several days. This justifies the use of lagged versions of the covariates in the logistic regression and causal feature selection stages.

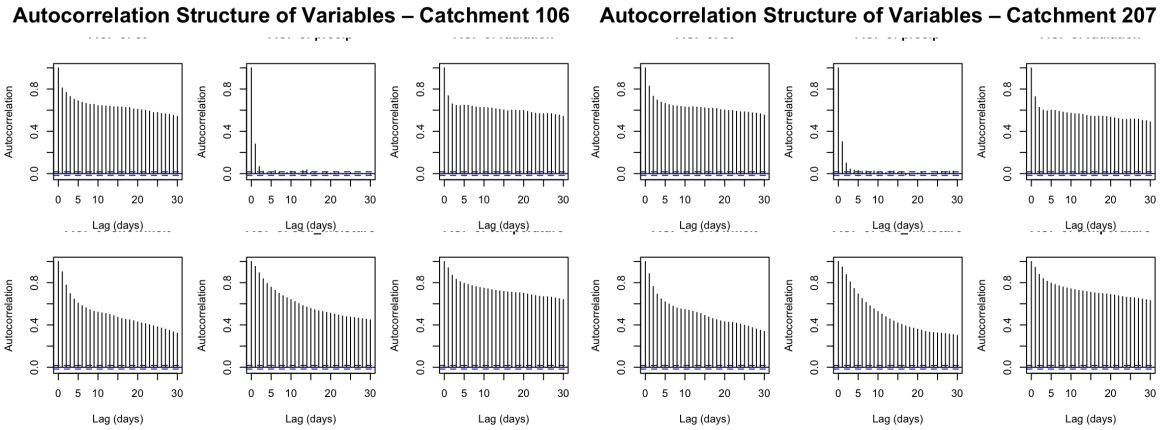

Figure 7: ACF of the covariates for catchment 106 and catchment 207, up to 30 lags.

3 Statistical Methods

This section outlines the methodology employed to identify hydrological and climatological drivers of spring flood events in Switzerland. The analysis aims to predict whether the daily discharge of a given catchment exceeds a high threshold and to evaluate the performance of both standard and causal variable selection techniques.

3.1 Problem Formulation

Let $Y_t \in \{0, 1\}$ denote a binary target variable indicating whether the discharge Q_t on day t exceeds the empirical 90^{th} percentile of its historical distribution. The predictor vector $\mathbf{X}_t \in \mathbb{R}^p$ includes daily meteorological and hydrological variables (e.g., precipitation, temperature, radiation, snowmelt, evapotranspiration, soil moisture) and their lagged values (e.g., at time $t-1$, $t-2$, and $t-3$).

To model the flood occurrence probability, we fit a logistic regression of the form:

$$\mathbb{P}(Y_t = 1 \mid \mathbf{X}_t) = \frac{1}{1 + \exp(-\beta_0 - \boldsymbol{\beta}^\top \mathbf{X}_t)},$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ are the model coefficients.

Due to the high dimensionality and collinearity of the covariates, we consider three different strategies for variable selection, as detailed below.

3.2 Standard Variable Selection: LASSO

The LASSO (Least Absolute Shrinkage and Selection Operator) is a regularized regression method that performs simultaneous coefficient estimation and variable selection. It solves the optimization problem:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \left\{ \ell(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

where $\ell(\boldsymbol{\beta})$ is the logistic log-likelihood and λ is a non-negative regularization parameter. Large values of λ induce sparsity in $\boldsymbol{\beta}$, effectively selecting a subset of covariates. The optimal λ is selected via k-fold cross-validation.

3.3 Causal Feature Selection via the PC Algorithm

To select variables based on causal relevance, we use the PC (Peter-Clark) algorithm, a constraint-based method for causal structure learning [1]. The algorithm aims to recover the skeleton and v-structures of the underlying causal directed acyclic graph (DAG) by testing conditional independencies among variables.

Under the standard assumptions of *causal sufficiency* (i.e., no unmeasured confounders), the *Markov condition*, and *faithfulness*, the PC algorithm can asymptotically recover the equivalence class of the true DAG from observational data. In particular, it leverages the fact that two variables are d-separated in the causal graph if and only if they are conditionally independent in the data-generating distribution.

In the multivariate Gaussian setting, conditional independence reduces to testing whether the partial correlation $\rho(X_i, X_j \mid S)$ between variables X_i and X_j, given a conditioning set S, is zero. The PC algorithm performs these tests iteratively, beginning with marginal independencies and progressively increasing the size of the conditioning set.

Once the full partially directed acyclic graph is estimated, the *Markov blanket* of the target variable Y_t is extracted as the union of its parents, children, and parents of its children. This set of variables is theoretically sufficient for optimal prediction of Y_t under the faithfulness assumption.

We then fit a logistic regression model using only the covariates in the Markov blanket, thereby restricting the classifier to variables with a direct or indirect causal link to the target. This approach reduces dimensionality, promotes interpretability, and serves as a baseline for comparing purely predictive and causally informed selection strategies.

3.4 Local Causal Discovery: Approximation to HITON-PC

As a complementary approach to global structure learning, we aimed to apply the HITON-PC algorithm [2], a local causal discovery method designed to estimate the Markov blanket of a target variable. HITON-PC combines a forward selection phase, where variables strongly associated with Y_t are iteratively added, and a backward elimination phase, where variables that become conditionally independent of Y_t given the current subset are removed. This procedure typically requires conditional independence testing and access to specialized implementations.

Due to technical issues with the required software, we approximated this procedure by ranking covariates according to their mutual information (MI) with the binary target variable Y_t. Mutual information is a non-parametric measure of dependency defined as:

$$\mathrm{MI}(X, Y) = \sum_{x,y} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right),$$

where $p(x, y)$ is the empirical joint distribution of the discretized covariate X and the response Y, and $p(x), p(y)$ are the corresponding marginals.

Each covariate was discretized prior to MI computation, and the top $k = 8$ covariates with the highest MI were retained. This approach corresponds to selecting the variables with the strongest marginal dependence to the response. The selected variables were then used as predictors in a logistic regression model. While this method does not incorporate conditional independence testing, it remains computationally efficient and preserves a causal motivation rooted in variable relevance to Y_t.

This simplified strategy provides an interpretable and scalable alternative to HITON-PC, and serves as a reference point for comparing local causal selection with global structure discovery via the PC algorithm.

3.5 Classification Metrics

To evaluate and compare the predictive performance of the classification models derived from different variable selection strategies, we consider several standard metrics in binary classification. Let us denote:

- TP — True Positives: correctly predicted flood events,
- TN — True Negatives: correctly predicted non-flood events,
- FP — False Positives: non-flood events incorrectly predicted as floods,
- FN — False Negatives: flood events incorrectly predicted as non-floods.

From these quantities, the following metrics are computed:

Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy measures the proportion of correct predictions over the total number of predictions. Although intuitive, it may be misleading in the presence of class imbalance, as is often the case in flood prediction tasks.

Precision

$$\text{Precision} = \frac{TP}{TP + FP}$$

Precision (also called positive predictive value) quantifies how many of the predicted flood events are actually correct. High precision implies a low false positive rate.

Recall

$$\text{Recall} = \frac{TP}{TP + FN}$$

Recall (also known as sensitivity or true positive rate) measures the ability of the model to correctly identify flood events. It is critical in applications where missing a flood is particularly costly.

F1 Score

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1 score is the harmonic mean of precision and recall, providing a balanced measure when both false positives and false negatives are relevant.

Area Under the ROC Curve (AUC) The Receiver Operating Characteristic curve plots the true positive rate (recall) against the false positive rate (FPR = $FP/(FP+TN)$) as the classification threshold varies. The AUC measures the area under this curve, providing an aggregate measure of performance across all classification thresholds:

$$\text{AUC} = \int_0^1 \text{TPR}(t)\, d(\text{FPR}(t)).$$

AUC values range from 0.5 (no discrimination) to 1.0 (perfect discrimination). It is threshold-independent and widely used in imbalanced classification scenarios.

Remarks All metrics are computed on a test set consisting of the last two years of data, ensuring temporal separation from the training set. Particular attention is paid to precision and recall, given the asymmetric costs of false negatives (missed floods) and false positives (false alarms) in hydrological forecasting.

4 Results

4.1 Overview

This section presents the results obtained from the application of four modelling strategies to two hydrological catchments of different elevation. Catchment 207 corresponds to a low elevation area, while Catchment 106 represents a high elevation basin. For each catchment, we modelled the binary target variable $Y = \mathbb{1}\{\text{discharge} > Q_{0.9}\}$, indicating flood events exceeding the 90^{th} percentile of discharge.

The same workflow was applied to both catchments: covariates were lagged up to three days and standardized, and the time series was split into training data (before 2015) and testing data (from 2015 onwards). The four modelling strategies employed are:

1. **Full Logistic Regression**: a classical model including all lagged covariates.
2. **LASSO Logistic Regression**: regularized regression with ℓ_1 penalty to induce sparsity.
3. **PC Algorithm**: structure learning based on conditional independence tests, followed by regression on the Markov blanket of the target.
4. **Simulated MMPC**: a proxy for local causal discovery based on mutual information ranking (due to limitations in installing the `MXM` package).

Each method was evaluated using five metrics on the held-out test set: accuracy, precision, recall, F1 score, and AUC. Results are reported separately for each catchment, followed by a comparative discussion. Only the most informative plots (e.g., Markov blanket graphs) are included to support the interpretation of results.

4.2 Catchment 207 (Low Elevation)

We applied all four modelling strategies to Catchment 207, a low-elevation basin. The binary response variable Y indicates whether the discharge exceeds its empirical 90^{th} percentile, and lagged covariates up to three days were constructed for all meteorological inputs. The dataset was split temporally, with training data before 2015 and test data from 2015 onwards.

The **full logistic regression model**, which includes all lagged covariates without penalization, achieved an accuracy of 0.970 and an AUC of 0.991, suggesting excellent discriminatory power. Interestingly, the **LASSO-regularized logistic regression**, which selected only 19 covariates out of the 24 used in the full model using the $\lambda_{1\text{se}}$ criterion, obtained *identical performance* across all metrics. This indicates that the predictive signal is concentrated in a small subset of features and highlights the model's potential for interpretability and dimensionality reduction.

The **PC algorithm**, used to estimate the Markov blanket of Y via conditional independence tests, resulted in a compact structure involving primarily lagged values of meteorological drivers related to precipitation, radiation, and snowmelt. A logistic regression model restricted to these variables achieved slightly lower performance, with an F1 score of 0.683 and a recall of 0.571, but maintained a high precision (0.848) and AUC (0.982). The associated graph is shown in Figure 8.

Finally, the **simulated MMPC approach** based on mutual information ranking selected the top 8 variables and achieved results comparable to the PC algorithm (accuracy 0.962, AUC 0.966). Although both causal discovery approaches yielded lower recall compared to the full and LASSO models, they offer valuable insights into the most relevant predictors of extreme discharge events.

4.3 Catchment 106 (High Elevation)

We repeated the modelling pipeline on Catchment 106, located at higher elevation. As with the previous basin, the target variable Y indicates whether the discharge exceeds the 90^{th} percentile, and the same lagged covariates and temporal train/test split were employed.

Table 1: Evaluation metrics for Catchment 207 (test set)

Model	Accuracy	Precision	Recall	F1 Score	AUC
Full Logistic Regression	0.970	0.829	0.694	0.756	0.991
LASSO (λ_{1se})	0.970	0.829	0.694	0.756	0.991
PC Algorithm (Markov Blanket)	0.964	0.848	0.571	0.683	0.982
Simulated MMPC (MI ranking)	0.962	0.800	0.571	0.667	0.966

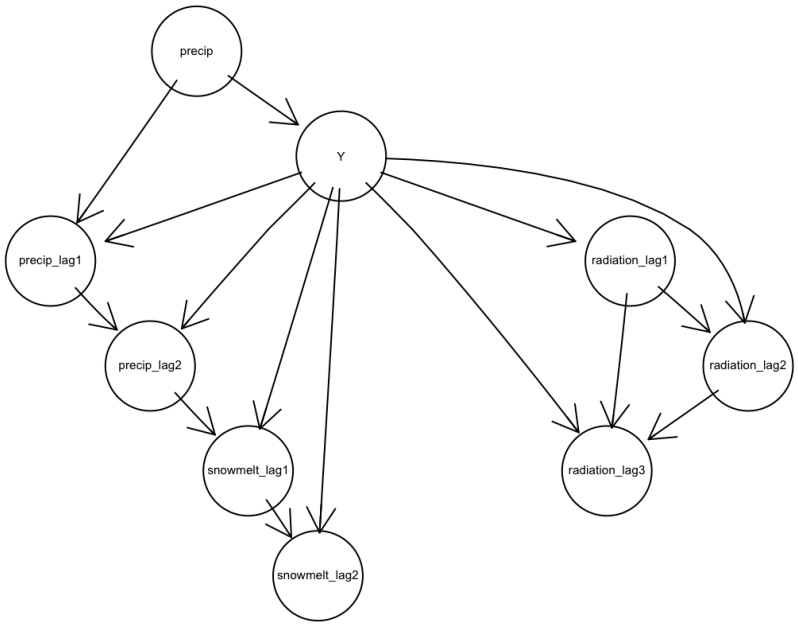

Figure 8: Markov blanket of Y for Catchment 207 estimated via the PC algorithm.

The **full logistic regression model**, based on all 24 covariates, achieved strong performance with an accuracy of 0.960 and an AUC of 0.980. Precision was particularly high (0.957), although the recall was more moderate (0.625), suggesting some difficulty in capturing all extreme discharge events.

The **LASSO-regularized model**, using the λ_{1se} rule, selected only 15 covariates yet yielded nearly identical results (accuracy 0.958, AUC 0.981). This again demonstrates that much of the predictive power is retained by a reduced set of lagged predictors.

The **PC algorithm** identified a concise Markov blanket around Y, primarily involving lagged values of hydrometeorological inputs associated with atmospheric demand, liquid precipitation, and snow-related processes. Restricting the logistic regression to these variables led to a slight drop in performance, with an F1 score of 0.661 and a recall of 0.514. However, AUC remained very high at 0.981. The corresponding causal graph is displayed in Figure 9.

Lastly, the **simulated MMPC model** based on mutual information ranking selected the 8 most informative covariates. This model showed lower recall (0.444), indicating a tendency to miss some true positives, but maintained a competitive AUC of 0.960 and very high precision (0.941).

Table 2: Evaluation metrics for Catchment 106 (test set)

Model	Accuracy	Precision	Recall	F1 Score	AUC
Full Logistic Regression	0.960	0.957	0.625	0.756	0.980
LASSO (λ_{1se})	0.958	0.936	0.611	0.739	0.981
PC Algorithm (Markov Blanket)	0.948	0.925	0.514	0.661	0.981
Simulated MMPC (MI ranking)	0.943	0.941	0.444	0.604	0.960

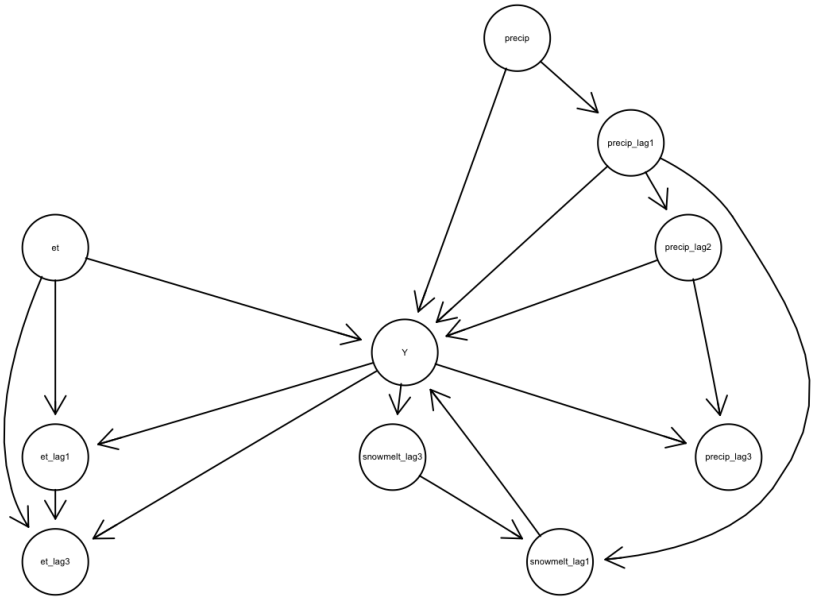

Figure 9: Markov blanket of Y for Catchment 106 estimated via the PC algorithm.

5 Discussion and Conclusion

Model Comparison across Catchments

Across both catchments, the full logistic regression and the LASSO-regularized model yielded very similar predictive performance. In Catchment 207, both models reached an AUC of 0.991 with an F1 score around 0.76, while in Catchment 106 the AUC remained high (0.980–0.981), with slightly lower F1 scores. LASSO achieved these results using only 15 covariates out of the 24 available, confirming its ability to provide a sparser yet competitive alternative to the full model.

The causal feature selection strategies, based on the PC algorithm and on mutual information ranking (approximating HITON-PC), showed lower recall in both catchments. This behavior is expected, as these methods aim to identify the most causally relevant covariates, rather than to maximize purely predictive metrics. In Catchment 207, the PC-based model still maintained good balance (F1 score 0.683; AUC 0.982), whereas in Catchment 106 the F1 score dropped to 0.661. The mutual information model, although computationally efficient, suffered from lower recall (0.444 in Catchment 106) and should be interpreted more as a baseline method.

Interestingly, model performance was consistently better in Catchment 207 across all approaches. This may be due to clearer separation between extreme and non-extreme discharge values in the low-elevation basin, or to stronger signal in covariates like soil moisture and snowmelt. In contrast, Catchment 106 displayed more variability in predictors and target values, as noted in the exploratory analysis, which may have limited the models' recall despite high AUC values.

Model Selection and Interpretation

From a predictive standpoint, the full logistic regression and the LASSO model provide the highest overall performance, with AUC values close to 1 and well-balanced F1 scores. Given that LASSO achieves comparable accuracy while discarding approximately 40% of the covariates, it offers a strong trade-off between performance and parsimony. This sparsity enhances interpretability and reduces overfitting, making LASSO an attractive default choice when interpretability and dimensionality reduction are desired.

However, when prioritizing causal relevance over predictive power, the PC-based selection offers valuable insights. By restricting the input space to the Markov blanket of Y_t, this approach ensures that only variables with a direct or indirect causal pathway to the target are retained. Although the resulting models exhibit lower recall, a common consequence of variable exclusion, they maintain high precision

and AUC, indicating robustness in predicting true positives.

The mutual information ranking method, while lacking the conditional independence refinement of HITON-PC, serves as a lightweight approximation to causal feature selection. It performs adequately as a baseline, especially when computational constraints or lack of prior causal structure prevent full causal discovery.

We also observed that the PC algorithm selected different subsets of covariates across the two catchments. This reflects underlying differences in their hydrological and climatic behavior, and highlights how causal discovery methods can adapt to the specific structure of the data. Such variability reinforces the value of causally motivated selection when generalization across contexts is not guaranteed.

Ultimately, the choice of model depends on the analytical objective. If the goal is accurate forecasting of extreme discharge events, LASSO or the full logistic model are preferable. If the goal is to identify and interpret the most influential drivers of extremes, particularly under potential interventions, then causal selection via PC is more appropriate, despite slightly lower predictive metrics.

Limitations and Final Remarks

Several limitations must be acknowledged. First, the HITON-PC algorithm could not be implemented due to software constraints. While the mutual information ranking provides a computationally efficient proxy, it does not perform conditional independence testing and may retain spurious associations. Therefore, it should not be considered a rigorous substitute for HITON-PC, but rather a crude approximation. In particular, since mutual information is based on marginal dependencies, the method fails to detect whether the dependence between a covariate and the response variable vanishes once conditioning on other variables. This distinction is crucial in causal discovery, where the goal is to isolate direct effects from mere associations.

Second, all models are based on logistic regression, which assumes linear relationships on the log-odds scale and additive effects. While this facilitates interpretability and estimation, more flexible models (e.g., tree-based methods or additive models) may capture non-linear dependencies and interactions that are not accounted for here.

Finally, the evaluation of models is based solely on predictive performance over a single test period. A more comprehensive assessment would involve cross-validation or temporal holdout schemes across multiple years, as well as the incorporation of uncertainty quantification for estimated parameters and predictions.

Despite these limitations, the analysis provides clear evidence that causally informed feature selection can identify compact, interpretable models with reasonable predictive power. Meanwhile, LASSO regularization emerges as a competitive approach that balances predictive accuracy and model simplicity. Overall, the framework developed in this study illustrates the trade-offs between predictive performance and causal interpretability, and highlights the importance of aligning model choice with the ultimate objective of the analysis.

References

[1] Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, Prediction, and Search*. MIT Press.

[2] Aliferis, C. F., Tsamardinos, I., & Statnikov, A. (2003). HITON: A novel Markov blanket algorithm for feature selection. *Proceedings of the AMIA Annual Symposium*, 21–25.

[3] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, 58(1), 267–288.

[4] Pearl, J. (2009). *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press.