Risk Analytics
Practical 1

Frédéric Aviolat and Juraj Bodik

Winter semester 2025-2026, HEC, UNIL

Hydrological data in Lake Neuchâtel region

Figure 1: View of Neuchâtel. The city and surrounding region are subject to hydrological variability influenced by both climatic and topographic factors. Source: Wikipedia.

Understanding the statistical properties of hydrological data is crucial for modeling and predicting extreme events such as floods or droughts. In regions like Lake Neuchâtel, precipitation and river discharge play a key role in water management, early warning systems, and climate resilience planning.

In this practical, we begin by exploring daily measurements of river discharge and precipitation in the Lake Neuchâtel region. The data reflect complex interactions between climate and terrain, and analyzing their statistical features allows us to understand both typical behavior and rare extremes.

Dataset: Daily river discharge of River Thielle (river merging into Lake Neuchâtel) and daily precipitation in the region (only summer months). Dataset is available on Moodle. Note that we slightly modified the dataset as the original dataset is not public.

Part 0 (do not include in the report)

(a) **Data upload.** Load the dataset containing daily river discharge (in m^3/s) and daily precipitation (in mm) in Lake Neuchâtel region. Get familiar with the data.

(b) **Visual inspection.** Plot the time series of river discharge and precipitation. Are there some NA values, periods of high variability or sharp spikes or different anomalies?

Part 1: Statistical assumptions for modeling extremes

Many statistical models used in hydrology rely on assumptions such as stationarity and approximate normality (especially after transformations). In this part, you will test these assumptions for the Lake Neuchâtel region dataset.

(a) **Visual assessment of distribution.** Plot histogram and QQ-plot of the river discharge series. Comment on the shapes of the distributions and any signs of skewness, kurtosis, or heavy tails.

(b) **Formal assessment of distribution.** Use the Anderson-Darling test to evaluate whether the river discharge series follow a Normal distribution. Can we say they follow a Gaussian distribution? If not, which well-known distribution does it resemble? (there is more than one reasonable answer)

(c) **Fit a distribution.** Fit the distribution of your choice from part (b) (e.g. using `fitdistr()`). Use QQ-plots to compare the fit of the distribution to that of the Normal distribution. Which one seems to better capture the empirical distribution?

(d) **Tail comparison and interpretation.** Plot the fitted density curves of the Normal and your chosen distributions over the histogram of the river discharge series. Compare their tails: which distribution assigns more probability mass to extreme events?
Based on this, discuss the implications for modeling extreme hydrological events such as floods using Gaussian distribution.

Part 2: Correlation versus causation

We are interested in understanding the relationship between the two hydrological variables in Lake Neuchâtel region: daily river discharge and daily precipitation. Are they statistically dependent? Is there a lagged influence — for instance, does rainfall today help predict river discharge tomorrow?

(a) **Are river discharge and precipitation dependent?** Compute the correlation using the `cor.test()` function. Can we conclude that the time series are independent?

(b) **Lagged dependence: Cross-correlation function (CCF).** Calculate the cross-correlation function between precipitation and river discharge using the `ccf()` function. What patterns do you observe? Are there lags at which the correlation is stronger?

(c) **Extremograms: Cross- and auto-dependence of extreme events.** The extremogram is a tool used to assess the temporal dependence of extreme events in a time series. Unlike the classical CCF, which focuses on average behavior, the extremogram quantifies how extreme values propagate over time or across variables.

Using a high threshold (e.g., the 95th percentile), compute:

- **Univariate extremograms** (autocorrelation of extremes) for Precipitation and River discharge datasets. Assess whether extreme events (e.g., heavy rainfall, high discharge) tend to cluster over time within the same variable.
- **Cross-extremogram** between Precipitation and River discharge.

Which dataset exhibits stronger clustering of extreme values?

Hint: You may use the `extremogram1(),extremogram2()` functions from the `extremogram` package.

(d) **Predictive relationships.** Assess whether one variable helps to predict the other using the `grangertest()` function from the `lmtest` package. Discuss whether one time series Granger-causes[1] the other. Repeat the analysis using the `Extreme_causality_test()` function from `JuroExtremes.R`, available on Moodle. (This function is also available from `https://github.com/jurobodik/Granger-causality-in-extremes.git`.)

(e) **Extreme events and predictive insight.** Based on your answer in (d), discuss the following:

(a) "We observe an extreme spike in precipitation. What should we expect for river discharge in the following days?"

(b) "We observe an extreme surge in river discharge. What can we infer about future precipitation?"

[1] C. Granger was a Nobel Prize winner (2003), recognized for his concept of Granger causality. Unlike the philosophical or counterfactual notion of causality—which seeks to identify the effect of an intervention or change—Granger causality is a predictive concept: if knowing the past of variable X improves the prediction of variable Y beyond what can be predicted using the past of Y alone, then X is said to Granger-cause Y. However, this does not imply that X is a true causal driver of Y, as the observed dependence may be due to hidden common causes or confounding. `https://calculatedcontent.com/2013/05/27/causation-vs-correlation-granger-causality/`

Part 3: Time series modeling, heteroscedasticity, and weather-driven volatility

Understanding the variability and predictability of environmental processes is crucial in hydrology and climate-aware planning. In particular, many hydrological and behavioral time series exhibit heteroscedasticity — periods of high and low variability — which complicates forecasting. In this part, we focus on modeling river discharge and investigate the assumptions of constant variance and random fluctuations over time.

(a) **Autocorrelation patterns.** Plot the autocorrelation function (ACF) of the raw river discharge series. Then difference the series (e.g., using lag-1 differencing) and plot the ACF of the differenced series. Which of the two seems easier to model?

(b) **Serial dependence testing.** Use the Ljung-Box test with $lag = 1$ to formally assess whether there is serial dependence in both the raw and differenced river discharge series. What do you conclude?

(c) **ARIMA modeling.** Based on visual tools (like the ACF/PACF), suggest a few plausible ARIMA models for the differenced discharge series. Then use `auto.arima()` from the `forecast` package to select an optimal model. Are the residuals independent and have constant variance? Do the residuals look Gaussian, or do they have heavier tail? Do you think that some transformation (exponential, logarithmic, Fourier...) can help here?

(d) **Modeling volatility: GARCH.** Fit a GARCH(1,1) model to the differenced discharge series using both the Normal and Student-t conditional distributions. Use the `garchFit()` function from the `fGarch` package. Again, comment on the residual diagnostics.

(e) **Two-step modeling approach.** An ARIMA model captures trends and autocorrelation in the mean but assumes constant variance. On the other hand, GARCH models are designed to capture time-varying volatility but require uncorrelated residuals and no trend. A two-step approach, where we first model the mean with ARIMA, and then model the residual volatility with GARCH, allows us to decouple these two aspects. However, too complex models can easily lead to overfitting.

- Fit an ARIMA model to the differenced river discharge series.
- Fit a GARCH(1,1) model to the residuals from the ARIMA model.

Again, comment on the residual diagnostics.

(f) **Model comparison and conclusion.** Compare the performance (using for example AIC) of the ARIMA-only model, the GARCH-only model, and the two-step ARIMA+GARCH model, possibly after an appropriate transformation of the original series. Which of these best captures the temporal structure and changing variability? Are the residuals in some of the models Gaussian (or at least close to Gaussian)? Which model would you pick for future modeling of the river discharge?