Project

1 Introduction

Temperature plays a central role in the climate system, influencing atmospheric processes, ecosystems, and human activities. Variations and extremes in temperature are key indicators of climate variability and change, and their statistical characterisation provides important insights for impact assessments and adaptation planning.

The objective of this project is to estimate the 10- and 100-year return levels of summer temperatures (June, July, and August), based on approximately 44 years of hourly measurements from monitoring stations across Switzerland provided by the MétéoSuisse platform (see Table 1). The measured variable corresponds to air temperature at 2 m above ground, expressed in degrees Celsius.

site	latitude	longitude	variable	start_date	end_date	filename
Bern	46.9479	7.4446	Temperature	1981-01-01	2025-12-31	bern_temperature_1980_latest.Rdata
Luzern	47.0300	8.3000	Temperature	1981-01-01	2025-12-31	luzern_temperature_1980_latest.Rdata
Lugano	46.0037	8.9511	Temperature	1981-01-01	2025-12-31	lugano_temperature_1980_latest.Rdata
Payerne	46.8100	6.9500	Temperature	1981-01-01	2025-12-31	payerne_temperature_1980_latest.Rdata
Pully	46.5100	6.6700	Temperature	1981-01-01	2025-12-31	pully_temperature_1980_latest.Rdata
Sion	46.2333	7.3500	Temperature	1981-01-01	2025-12-31	sion_temperature_1980_latest.Rdata
Zermatt	46.01998	7.74863	Temperature	1981-01-01	2025-12-31	zermatt_temperature_1980_latest.Rdata

Table 1: Metadata of the 7 monitoring sites in Switzerland, with RData filenames and observation periods.

You will work in pairs, so find a partner and tell us who you are working with. You will then be assigned a time series of temperature observations from a particular site, which you can load into `R` by (for example) `data2025 <- load(file="bern_temperature_1980_latest.Rdata")`.

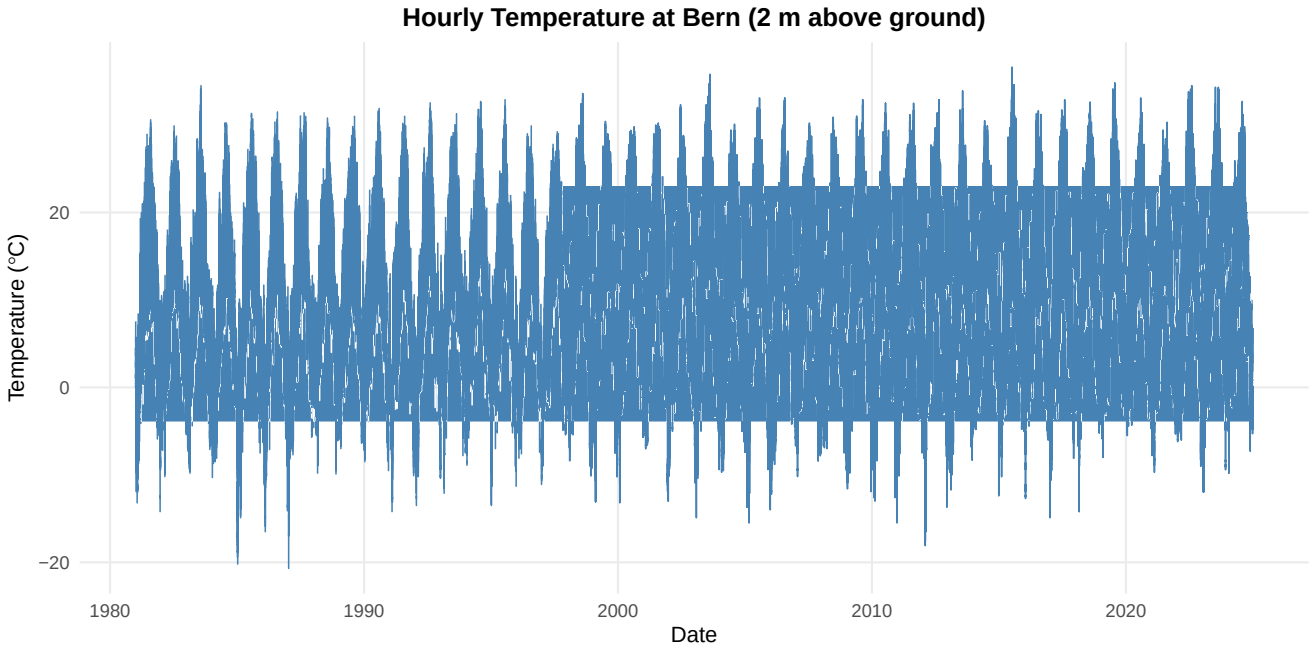

Figure 1: Hourly temperature (2 m above ground) recorded at the Bern/Zollikofen station.

Final deadline

By 17:00 on January 9th, 2026 *at the latest*, you should send files `SurnameA-SurnameB-Report.pdf` containing your report and `SurnameA-SurnameB-Code.txt` containing the R code to `ignacio.gonzalezperez@epfl.ch`. The code should be ready to run, and will be used to check that your results are honest, and that there has been no sharing of code.

The maximum length of the report should be **15 sides of A4 in 12pt text**, including all tables and figures. You can include an appendix with additional graphs and tables if you find this necessary, but there is no need to do so (and it may not be read). The report should not contain code or output directly copied from `R` or any other package.

You may discuss your work with other students, but the report and the code should be all your own work. Any signs of plagiarism or code-sharing will be heavily penalised.

2 Data

The `R` object `data2025` contains a time series of hourly measurements of air temperature measured at 2 m above ground at different sites in Switzerland. Each pair of students will be assigned data from one of the locations. After loading the file, you can see the structure of the object by typing `head(data2025)`. Some of the sites have missing data, but this is beyond the scope of the project, so we will simply ignore those.

We provide some examples of code that may be useful in processing the data; for instance, the following can be used to create a data-frame object with the original data:

```
# install.packages("lubridate")
library(lubridate)
data.tmp <- data.frame("date"= data2025$datetime, "value"= data2025$tre200h0,
"year"= year(data2025$datetime), "month"= month(data2025$datetime),
"day"= day(data2025$datetime))
```

The following code can be used to compute daily maximum of hourly measurements, ignoring any missing data:

```
# install.packages("dplyr")
library(dplyr)
daily_maxima <-data.tmp %>%
group_by(year, month, day) %>%
slice_max(order_by = value, n = 1, with_ties = FALSE) %>%
drop_na() #drop missing observations
```

You can adapt this code accordingly if you decide to analyse monthly or weekly maxima.
The following code can be used to keep only summer (June, July, and August) measurements:

```
data.tmp <- data.tmp %>% filter(month %in% c(6, 7, 8))
```

3 Goal and guidelines

The goal of your project is to use ideas from the course to estimate the 10- and 100-year return levels of the daily maximum of hourly temperature measurements.

Although **we only focus on summer measurements** and can thus safely assume the absence of seasonality in the measurements, the assumption of stationarity might be unrealistic due to the clear trend induced by climate change. In a first step, non-stationarity can be ignored by fitting stationary models for extremes and estimating return levels, which might be adjusted in the presence of clustering. In a second step, models taking into account time variation in the parameters should be considered to capture trend in the data. Return levels in the non-stationary setup should be computed using ideas from the stationary approach; see Section 3.4.

3.1 Stationary analysis

For comparison with later results, we first apply standard methods that ignore the non-stationarity.

Try fitting the standard three-parameter GEV to annual maxima (e.g., using the `fgev` function) and use your fit to estimate the 10- and 100-year return levels and to give confidence intervals for them. Then repeat this for monthly maxima.

Don't forget to check whether your models fit the data, using suitable residual or other plots. In this context it would be wise to plot suitable residuals against time, in case they show a trend.

Also try POT methods with a constant threshold (e.g., using the `fpot` function), also attempting to estimate the 10- and 100-year return levels, and again checking the fit.

Comment on your results, in particular discussing the fit of your model(s) and saying whether you think the return level estimates are realistic. Do you think they will over- or under-estimate values you might get from taking the non-stationarity into account? Is it sensible to estimate 100-year return levels in this context?

3.2 Extremal index

Extreme temperatures tend to occur in episodes, so it may be wise to estimate the extremal index θ, to check whether the hourly observations exhibit dependence at high levels. You might use the function `exiplot` from the `evd` package in `R`. Does this plot look stable? What can you conclude about the asymptotic dependence of extreme events? How would this affect the return levels computed above? Justify by adjusting the return levels for any clustering you think is present in the data.

3.3 Modelling non-stationarity

Two approaches to handle non-stationarity are suggested in this section and you need to choose only one to perform the data analysis. The selected method should also be used to compute the return levels as described in Section 3.4.

To model non-stationarity, we fit the GEV model to *monthly* or *yearly* maxima of hourly observations or apply POT methods to the *daily* exceedances of hourly observations. The two approaches differ and below we provide some suggestions about how you might attempt an analysis using either the GEV (with annual or monthly maxima) or POT models.

Non-stationarity via the GEV model: Figure 1 indicates the presence of an upward trend in the temperature measurements. One could model non-stationarity in the monthly maxima using a GEV distribution in which the parameters η, τ, and ξ depend on time. Figure 1 suggests that it may be reasonable to model the location parameter by taking (for instance)

$$\eta(t) = \eta_0 + \eta_1(t - t_0)/(100 \times 3), \tag{1}$$

where t denotes the month, starting from some suitable date. For example, if t_0 corresponds to 1 June 1981, then $\eta(t_0) = \eta_0$ and the location parameter would increase by η_1 every 100 years.

Such models can be fitted using the function `gev.fit` (in the `R` package `ismev`) using appropriate covariates, e.g., only time in this case (see the options `ydat`, `mul`, `mulink`, `muinit`, etc.). You can get help on such functions, including examples of code, by typing `?gev.fit` at the prompt. For instance, in case of monthly maxima, the following code models the location parameter η as in (1):

```
gev.fit(xdat=monthly_maxima$value,
  ydat=matrix((month+3*(year-1981))/(3*100), ncol=1, byrow=F), mul=c(1))
```

You might also want to model trend in the scale parameter.

You can compare the fits of the stationary and non-stationary models using their deviances and likelihood ratio statistics. Discuss whether there is a clear improvement over the stationary models fitted above.

Provide a discussion on the model fits and on the uncertainty of the parameter estimates; see for example the Venice data example in the lecture notes.

Non-stationarity and POT modelling The usual approach to POT fitting is to first fit a time-varying threshold u, for example with trend analogous to (1). This can be implemented via quantile regression, which lets the quantiles of a response variable to depend on other quantities. For instance, when the response variables are the daily maxima and we want to model the 90% quantile via covariates, we can use the function `gpd.fit` (in the library `ismev`) and code such as

```
(gpd_daily <- gpd.fit(daily_maxima$value,
 threshold= rq(val ~ matrix((day of year+92*(year-1981))/(100*92)), ncol=1, byrow=F),
 tau=.9, data=daily_maxima)$fitted.values, npy=92))
```

Exceedances of the varying threshold can sometimes be modelled by a stationary GPD, but (if you are less lucky because the stationary model is inadequate) you may need to model the scale parameter of the GDP using the same covariate. This can be implemented by using the covariates as input `ydat` in the function `gpd.fit` setting the argument `sigl=c(1)`, similar to argument `mul=c(1)` we saw in `gev.fit`. Do you notice any changes from the previous fit?

As with the GEV model, discuss the fitted model and interpret its fit using the diagnostic plots. How do its results compare with those from your stationary approach?

Note on quantile regression

In a non-stationary (covariate-dependent) setting, the quantile of level p of a random variable $Z \sim F_Z(\cdot; x)$ is defined as

$$Q_x(p) = \inf\{z \in \mathbb{R} : F_Z(z; x) \geq p\}, \tag{2}$$

where $F_Z(z; x) = \mathrm{P}(Z \leq z; x)$, i.e., the distribution of Z and hence its quantiles depend on the known covariates x. In `R`, an estimate of the p-th quantile function of a response variable for fixed covariates is obtained using the function `rq` in the package `quantreg`.

3.4 Return levels under non-stationarity

GEV: We now switch to the non-stationary framework, and look at the return levels from the GEV model. You may assume

$$\mathrm{P}(Y_t \leq y) = \exp\left\{-\left(1 + \xi\frac{y - \widehat{\eta}_t}{\tau}\right)^{-1/\xi}\right\}, \tag{3}$$

where $\widehat{\eta}_t$ is the fitted location parameter using the covariate x. Now, if $Y_t \sim \mathrm{GEV}(\widehat{\eta}_t, \tau, \xi)$, then $Y_t - \widehat{\eta}_t \sim \mathrm{GEV}(0, \tau, \xi)$, i.e., the 'standardised' variables $\tilde{Y}_t = Y_t - \widehat{\eta}_t$ are approximately stationary. You may now estimate the return levels of $\tilde{Y}_t$. Can you think of a way to transform these 'standardised return levels' to the original scale? What about the underlying uncertainty? Note that by working with $\tilde{Y}_t$ we are essentially ignoring the uncertainty of the estimated coefficients on the right-hand side of equations such as (1). How do these return levels compare to those for the background (hourly) observations?

GPD: The procedure for obtaining return levels in a non-stationary setup and via the POT approach mimics the steps followed for the GEV model. In particular, we assume that for $x > \widehat{u}_t$, we have the following model

$$\mathrm{P}(X_t \leq x) = 1 - p_{u_t}\left(1 + \xi\frac{x - \widehat{u}_t}{\sigma}\right)^{-1/\xi}, \tag{4}$$

where p_{u_t} is the probability of exceeding the threshold u_t at time t (is this constant?) and the threshold u_t, corresponding to a high quantile of the series of daily measurements X_t, is modelled via quantile regression.

Therefore, if $X_t - \widehat{u}_t \mid X_t > \widehat{u}_t \sim \mathrm{GPD}(\sigma, \xi)$, we can take $\tilde{X}_t = X_t - \widehat{u}_t$ and write $\tilde{X}_t \mid \tilde{X}_t > 0 \sim \mathrm{GPD}(\sigma, \xi)$, i.e., the threshold of the transformed variables $\tilde{X}_t$ can be set to zero.

If you think that modelling both u_t and σ_t via covariates gives a better model, you may then assume that $\tilde{X}_t = (X_t - \widehat{u}_t)/\widehat{\sigma}_t \sim \text{GPD}(1, \xi)$, and proceed as above to compute the return levels in a stationary setup. Once you compute the 'standardised' return levels, transform them to the original scale. Discuss their uncertainty and how they compare to the return levels estimated from the GEV. Do you find these estimates realistic relative to the observed hourly temperature measurements?