Università degli Studi di Padova

Dipartimento di Fisica e Astronomia Galileo Galilei

Bayesian Analysis of ARPAV time series on temperatures

arpav

Advanced statistics for physics analysis

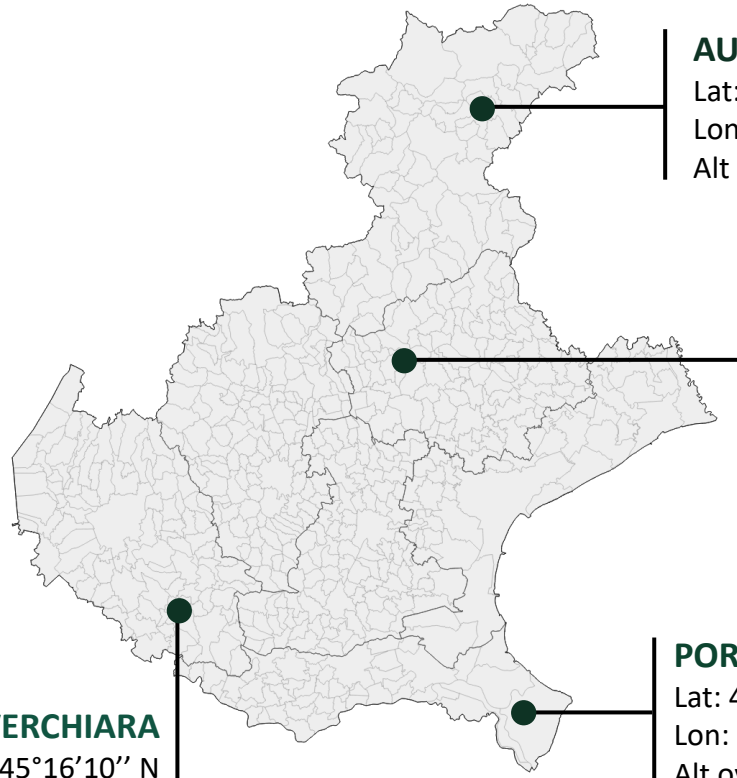
Prof. Alberto Garfagnini

Riccardo Corte
Alessandro Miotto
Lorenzo Rizzi

Introduction

This project analyzes ARPAV's temperature data from four distinct stations across the Veneto region.

Our analysis is based on Bayesian regression to uncover historical trends and an ARIMA based forecasting to predict future temperatures



AURONZO DI CADORE
Lat: 46°33'33'' N
Lon: 12°25'28'' E
Alt over see level: 849 m

CASTELFRANCO VENETO
Lat: 45°40'00'' N
Lon: 11°55'00'' E
Alt over see level: 49 m

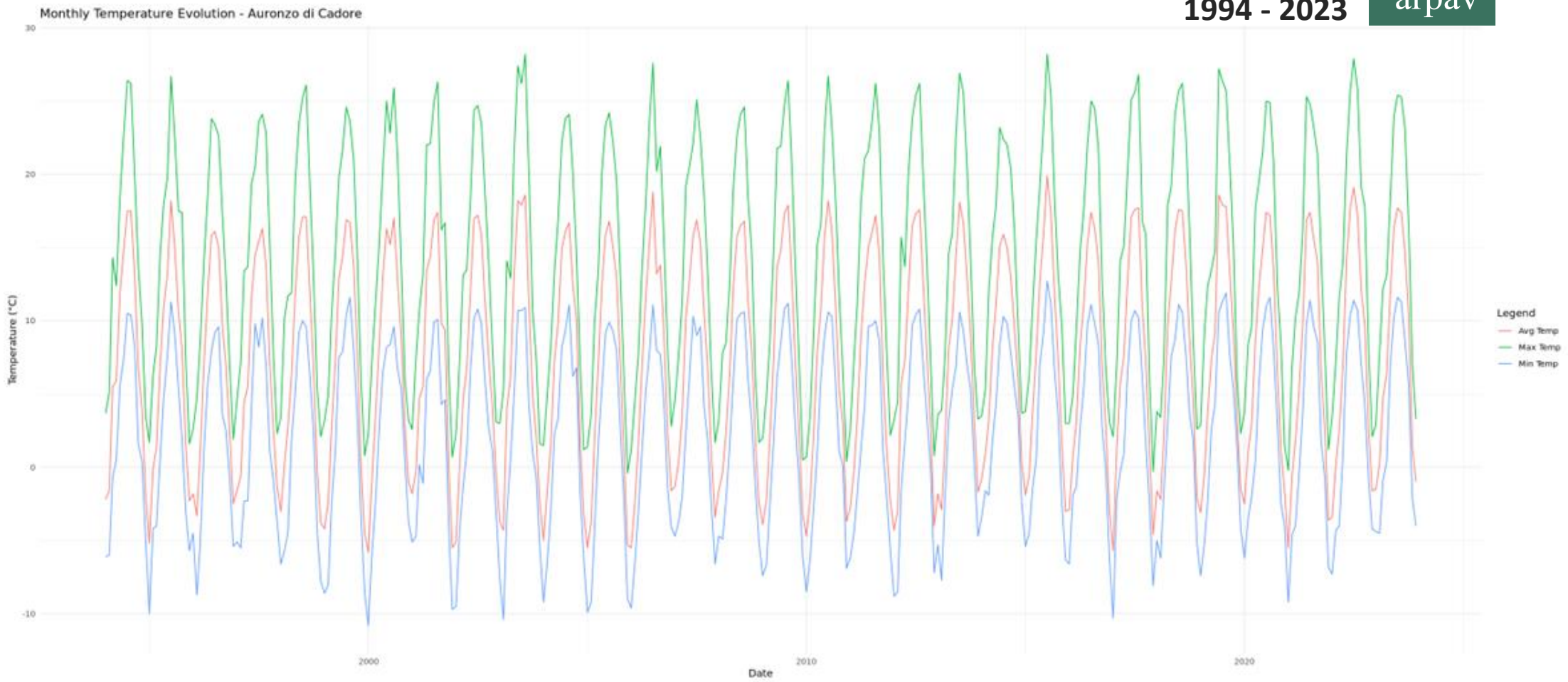
PORTO TOLLE
Lat: 44°56'58'' N
Lon: 12°19'28'' E
Alt over see level: -3 m

ROVERCHIARA
Lat: 45°16'10'' N
Lon: 11°14'41'' E
Alt over see level: 17 m

Introduction

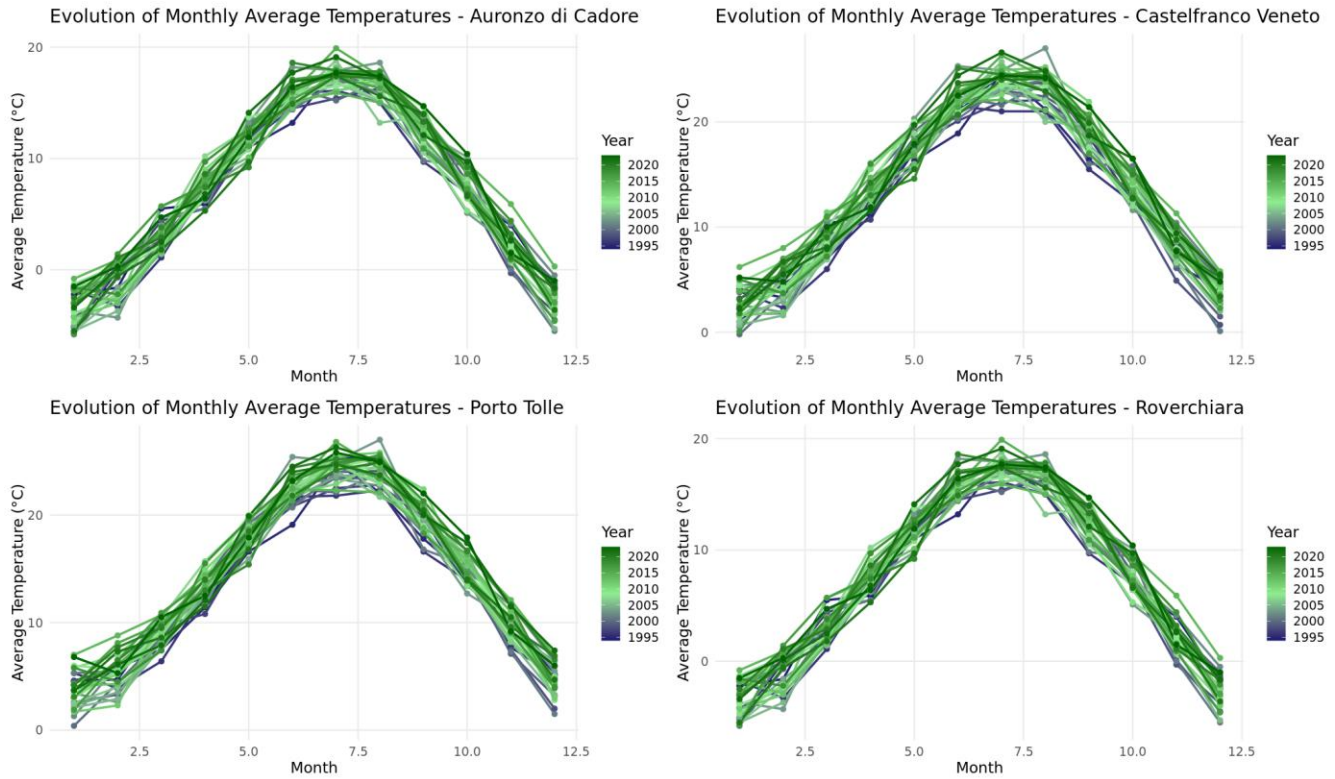
Monthly weather data from four different stations

1994 - 2023



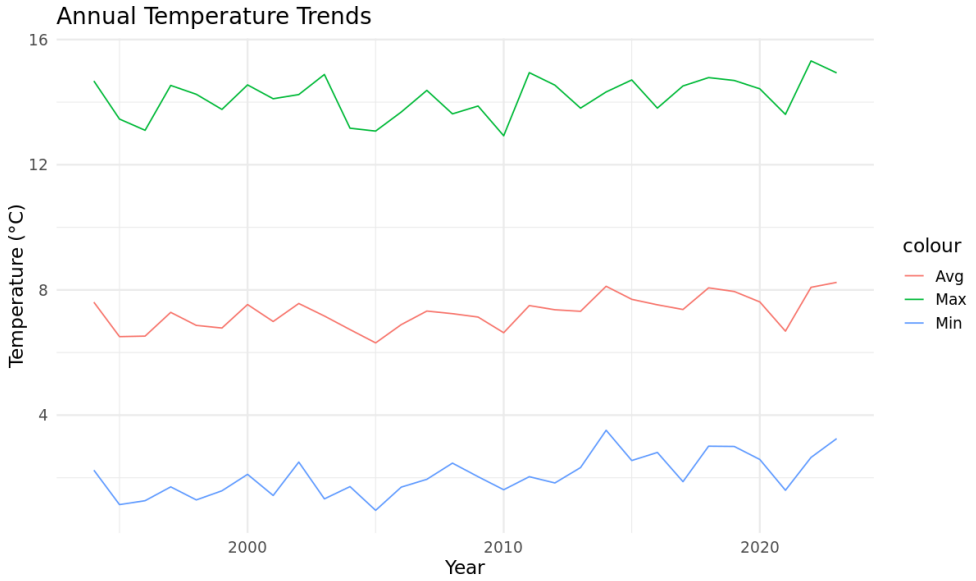
https://www.arpa.veneto.it/dati-ambientali/open-data/clima/principali-variabili-meteorologiche

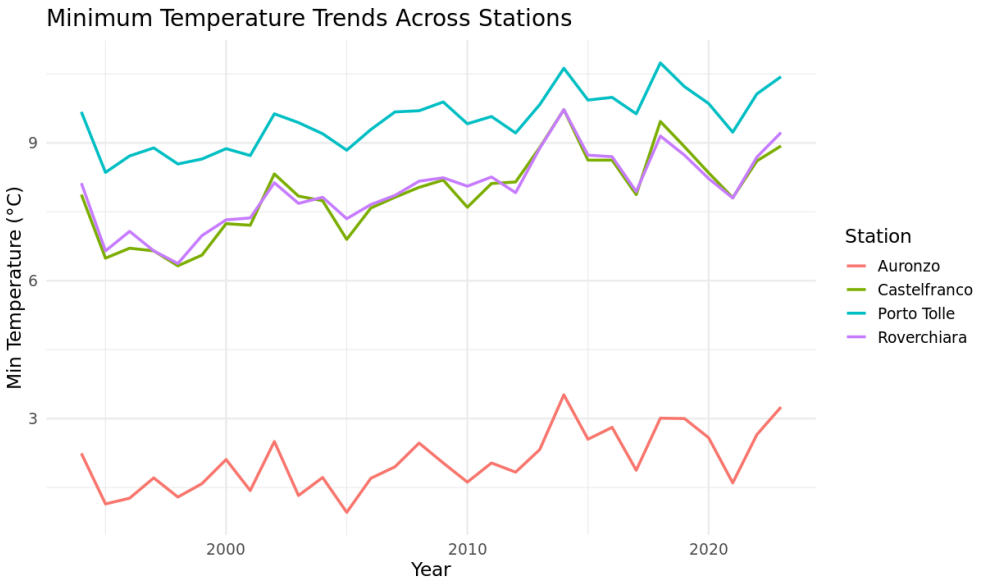
Part 0: Evolution of the temperatures over a month

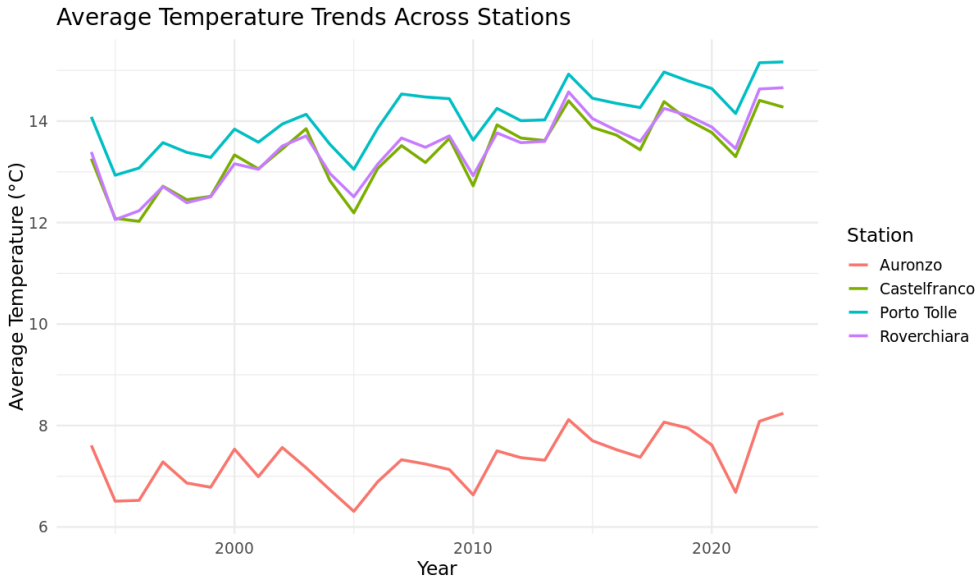


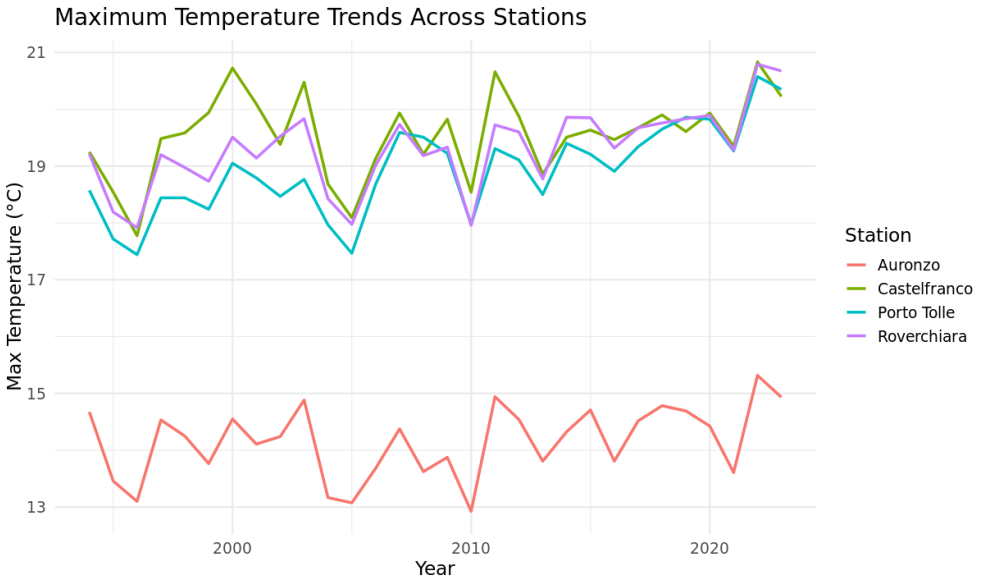
Mean Values (°C)	Auronzo di Cadore	Castelfranco Veneto	Porto Tolle	Roverchiara
Maximum	14.1	19.5	18.9	19.3
Average	7.3	13.3	14.1	13.4
Minimum	2.0	7.9	9.5	7.9

Part 1: Evolution of the temperatures over the year

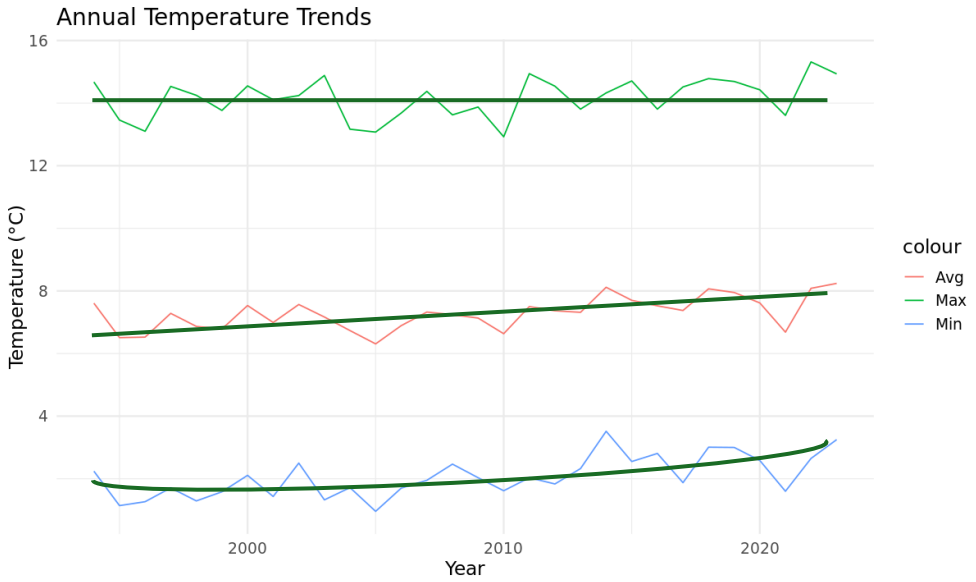


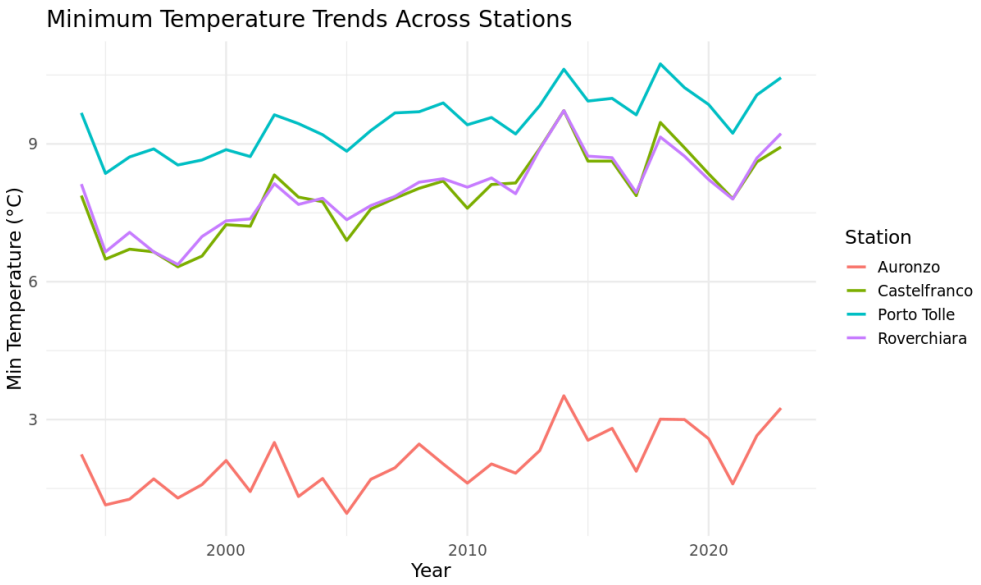


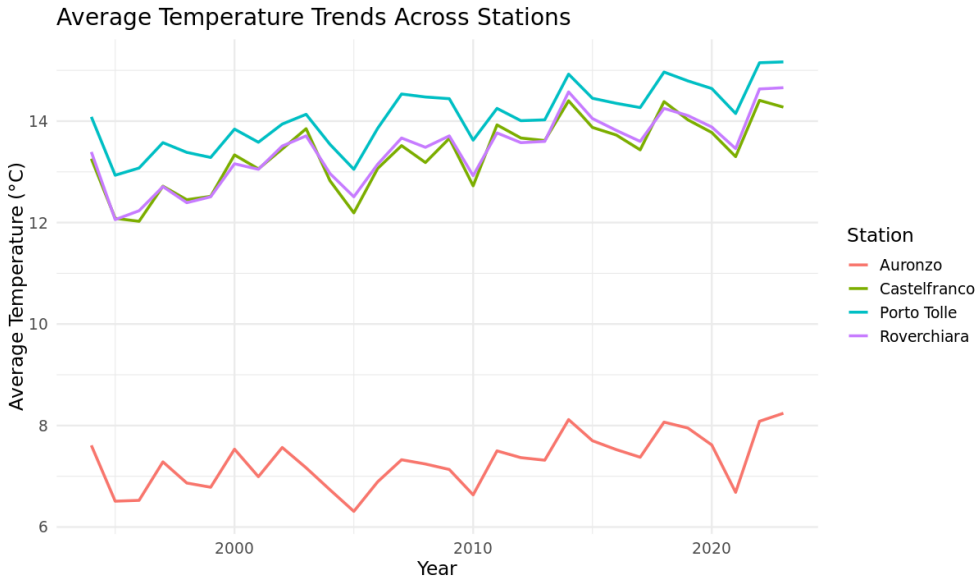


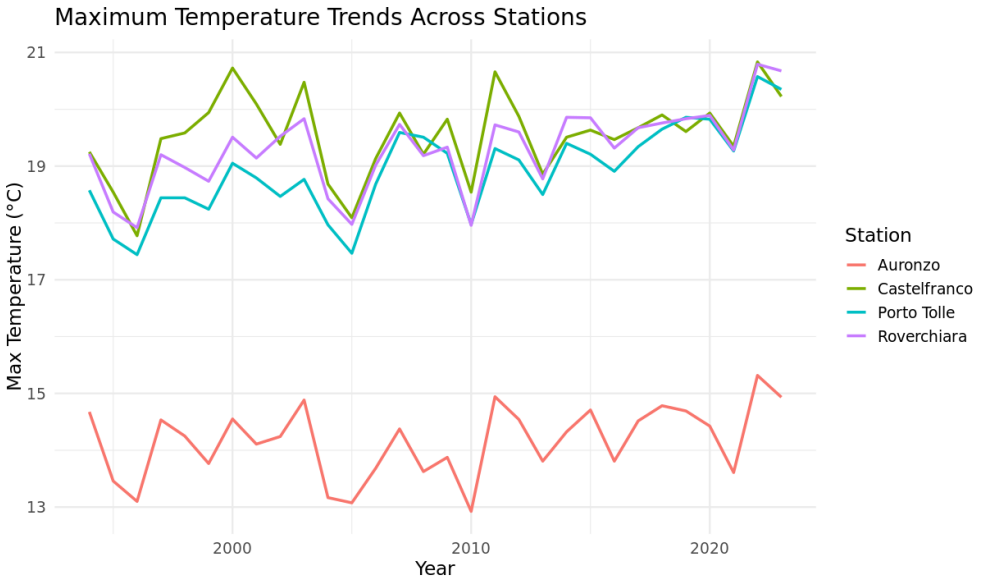


Part 1: Evolution of the temperatures over the year

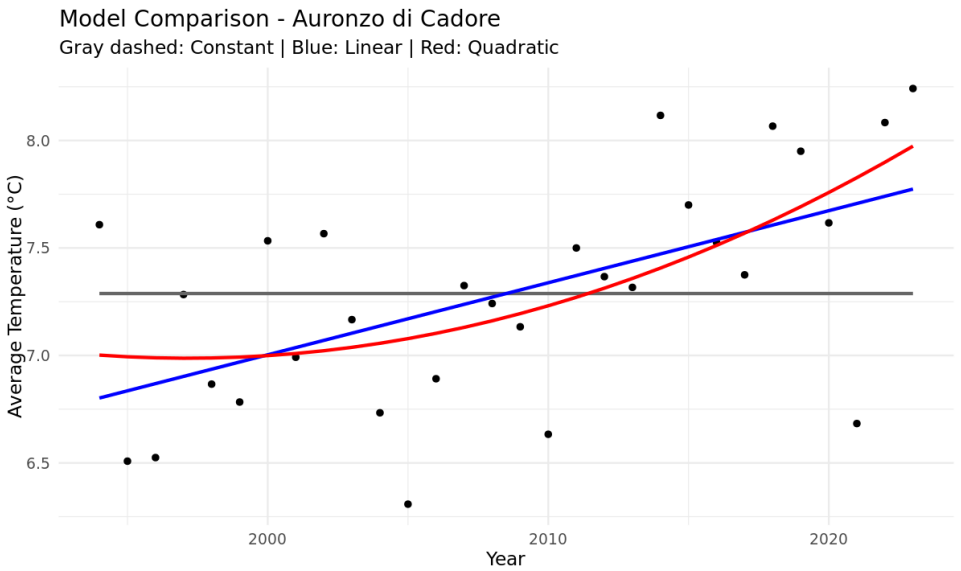


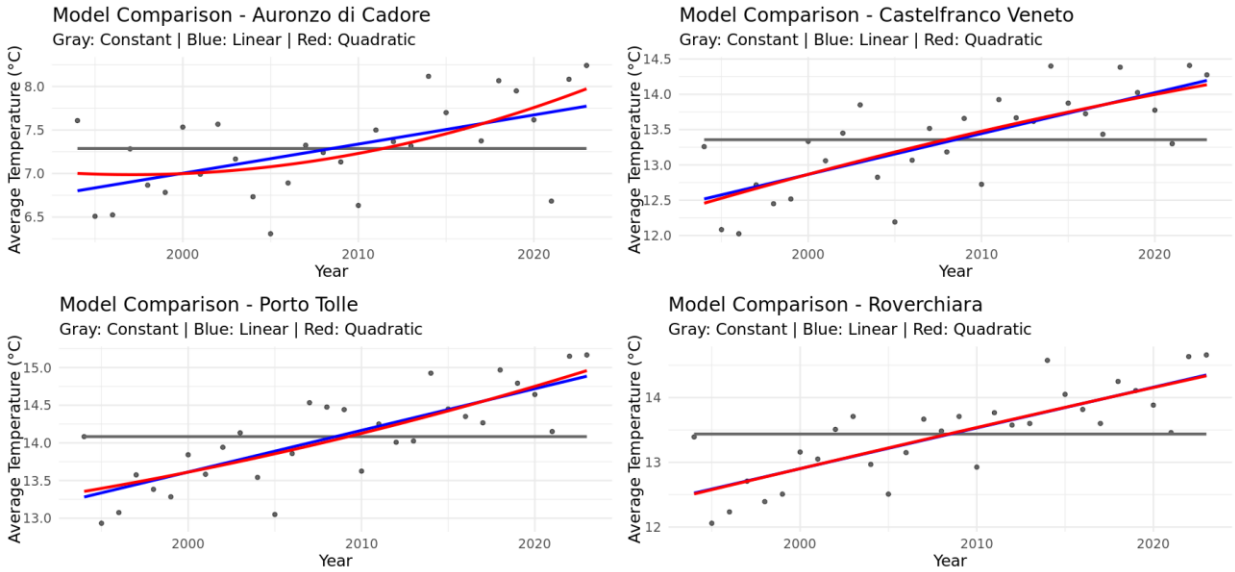






Part 1: Frequentist regression





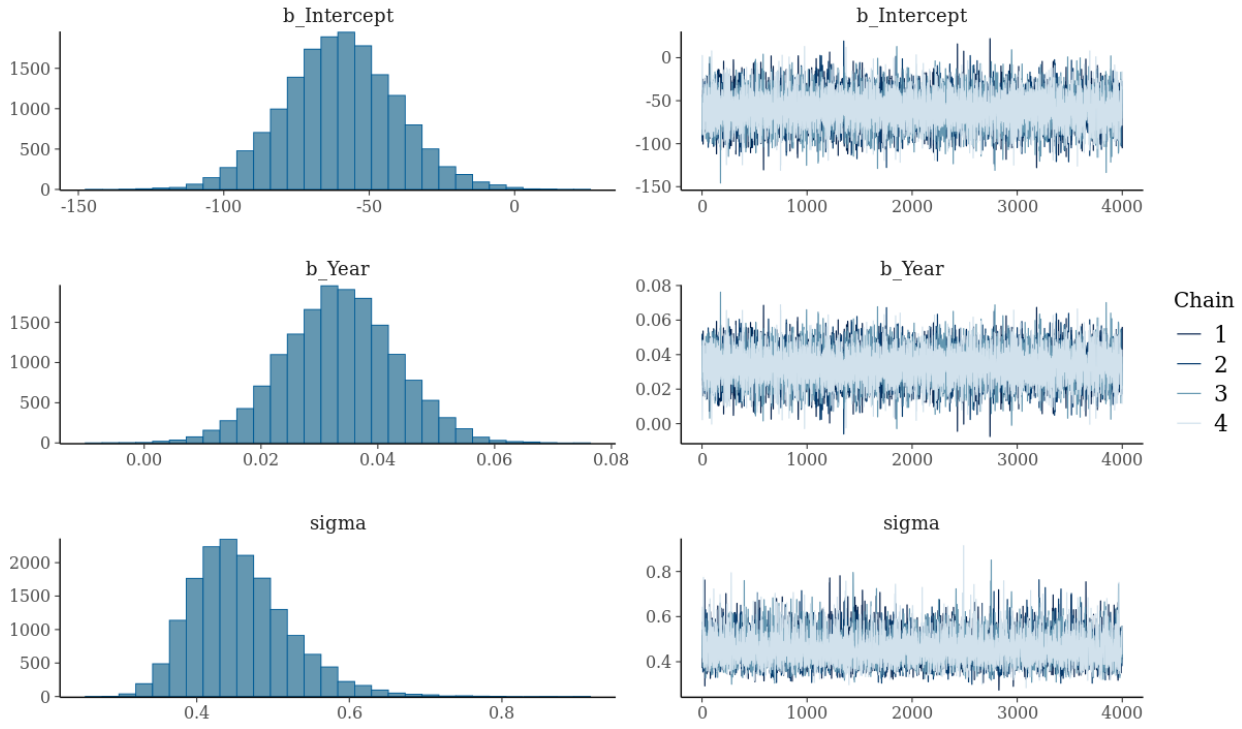
$$T(t) = a + \varepsilon$$

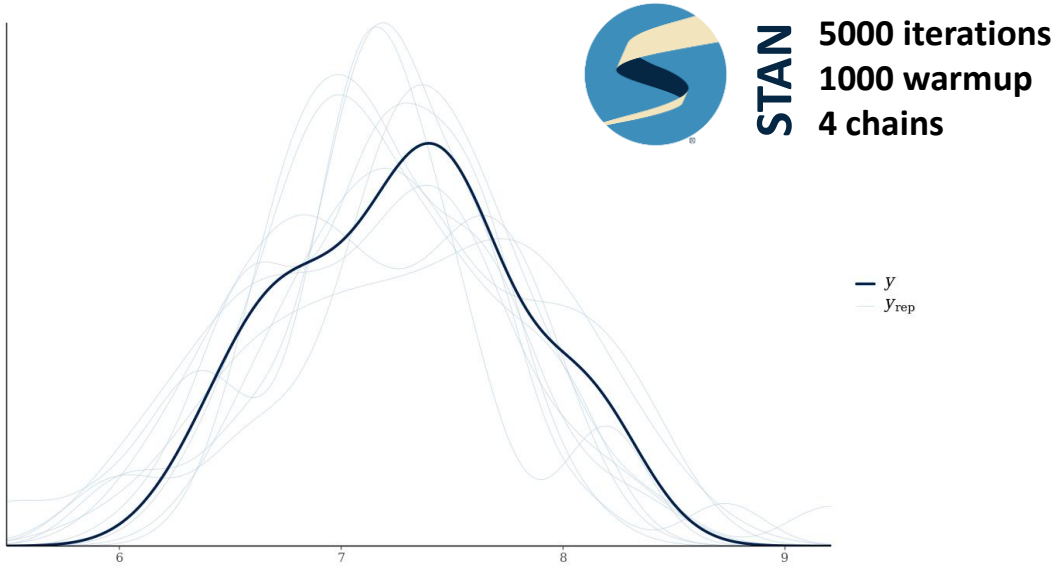
$$T(t) = b\,t + a + \varepsilon$$

$$T(t) = ct^2 + b\,t + a + \varepsilon$$

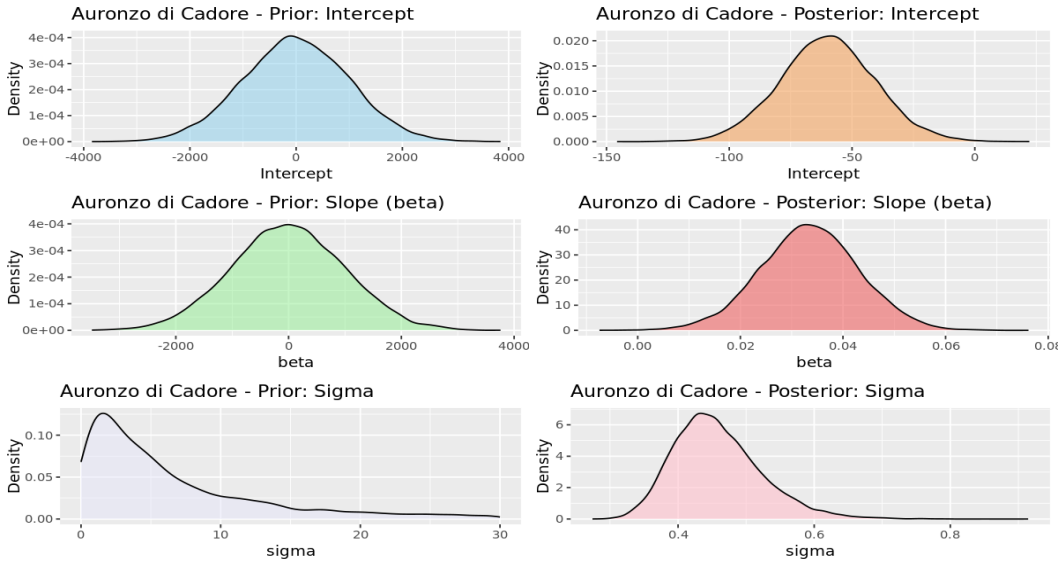
ANOVA RSS $(°C)^2$:	Auronzo di Cadore	Castelfranco Veneto	Porto Tolle	Roverchiara
Constant Model	7.87	13.5	11.0	13.9
Linear Model	5.35	6.03	4.18	5.02
Quadratic Model	5.06	6.01	4.14	5.02

Part 1: Bayesian Linear Regression using STAN



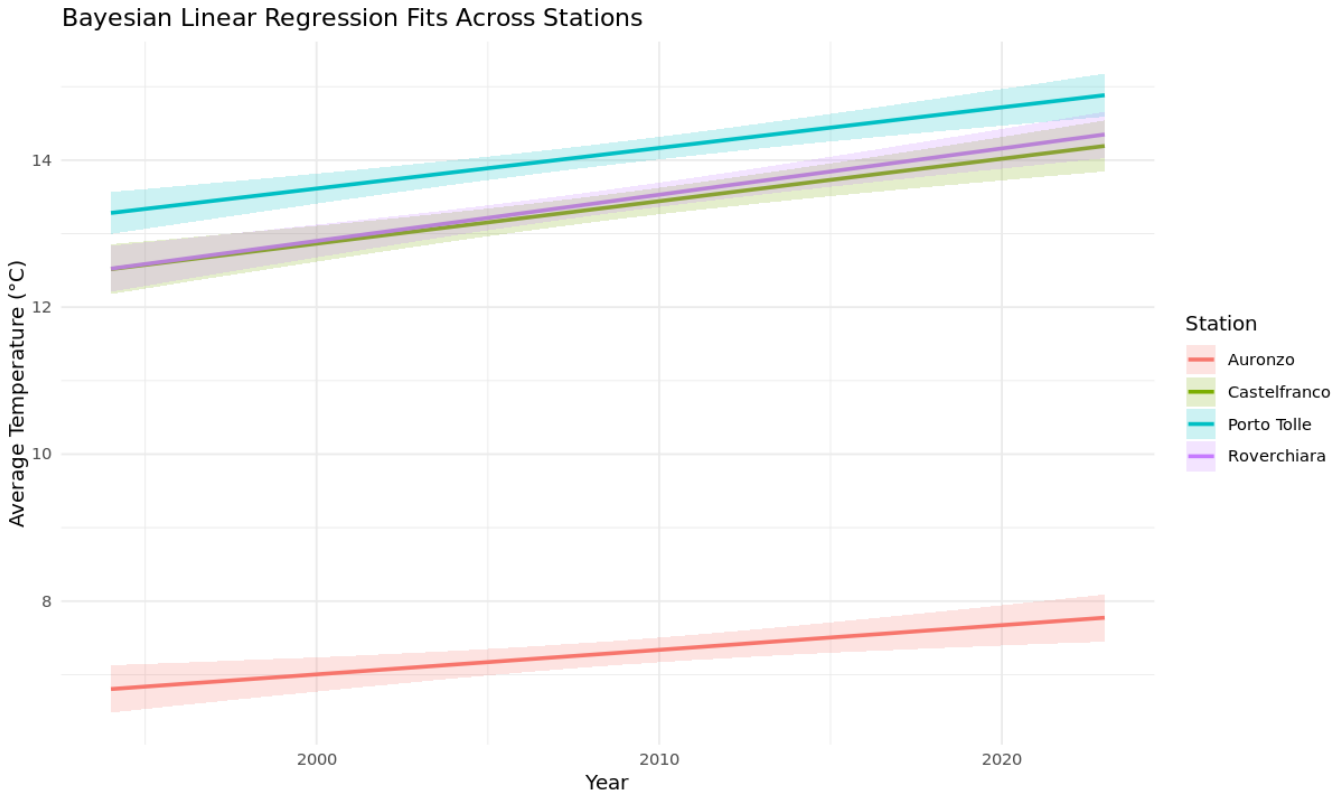


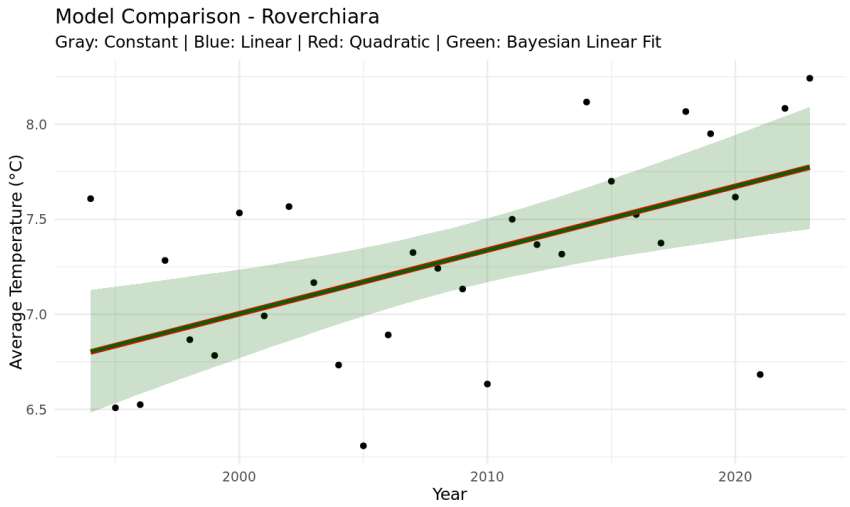
Parameters prediction	Auronzo Di Cadore	Castelfranco Veneto	Porto Tolle	Roverchiara
Intercept	-59.92	-102.6	-96.96	-112.92
Slope	0.03	0.06	0.06	0.06
Sigma	0.46	0.49	0.41	0.44

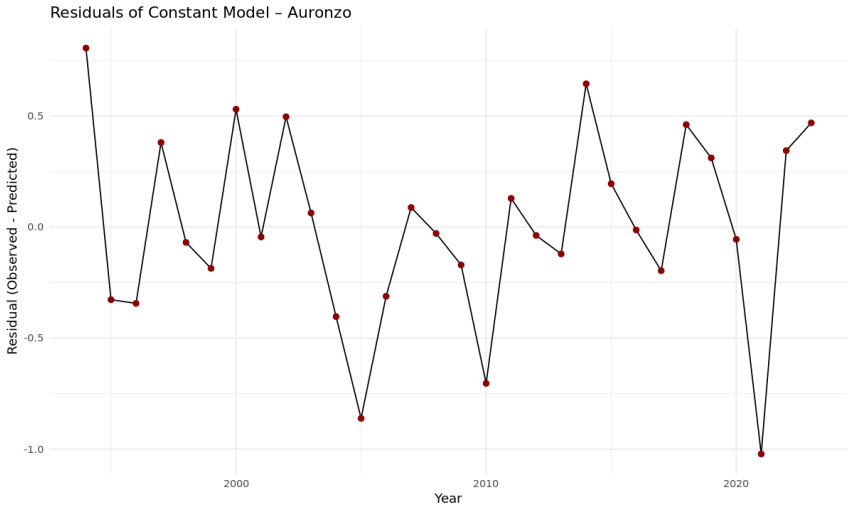


Part 1: Bayesian and Frequentist approach comparison

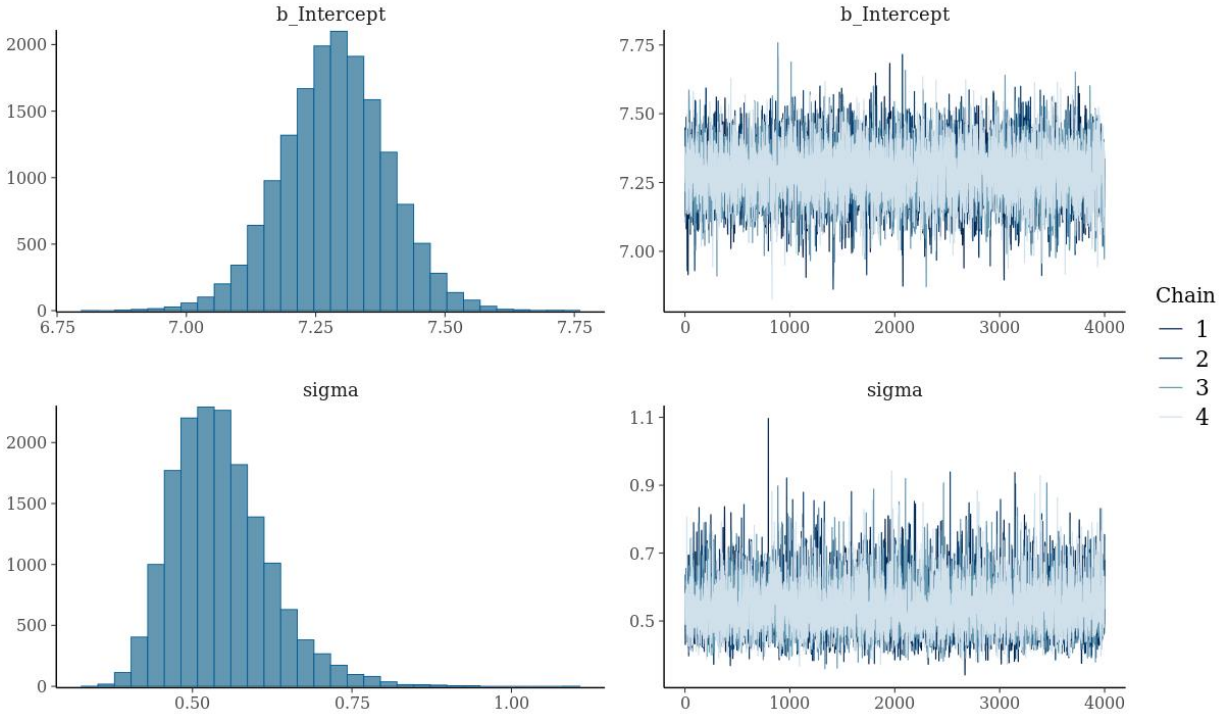
The results are consistent between bayesian and frequentist approach





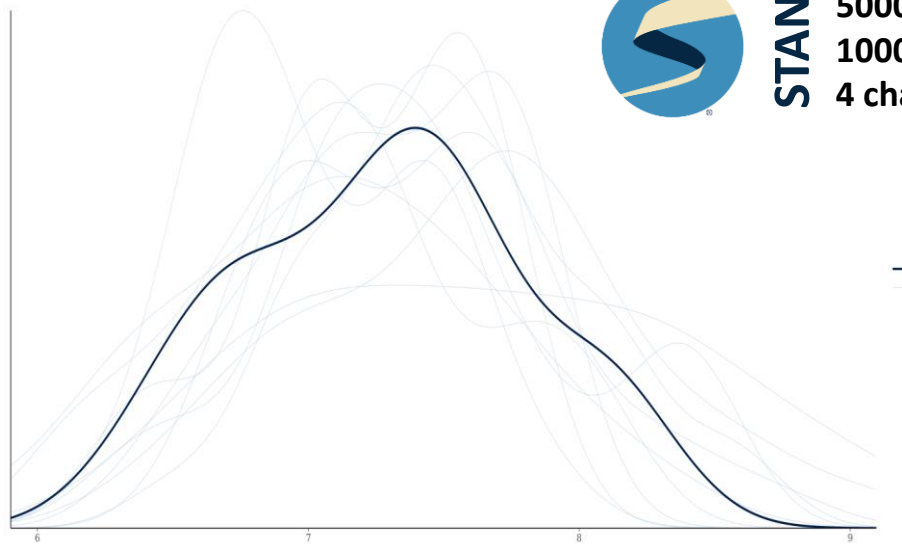


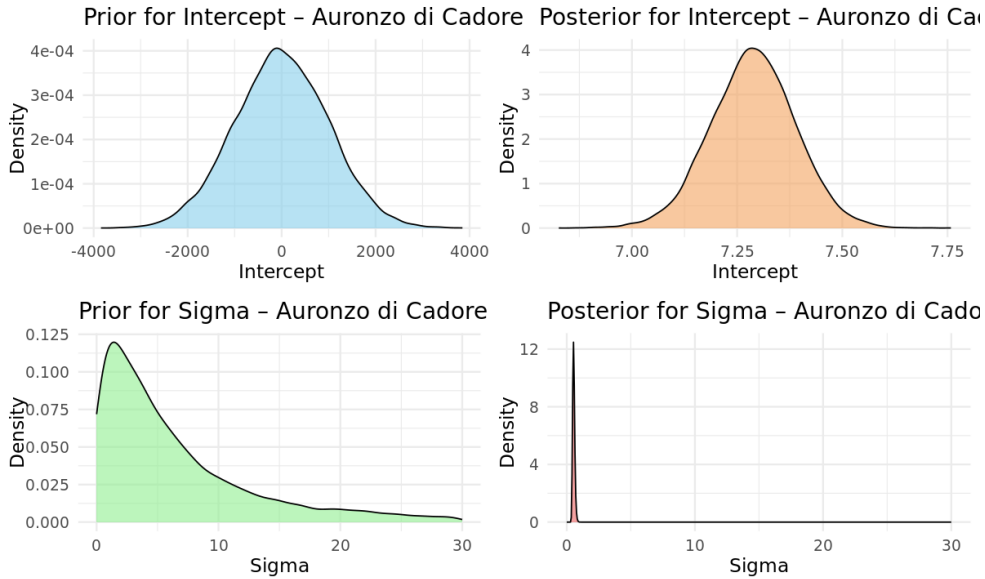
Part 1: Constant regression with STAN



STAN

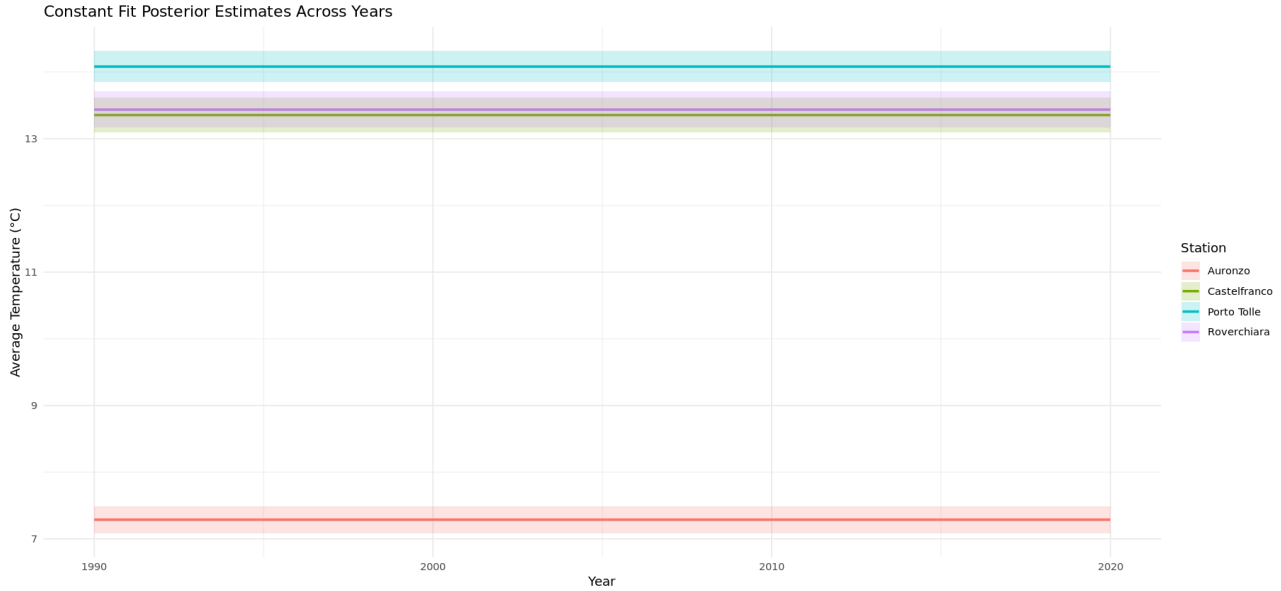
5000 iterations
1000 warmup
4 chains

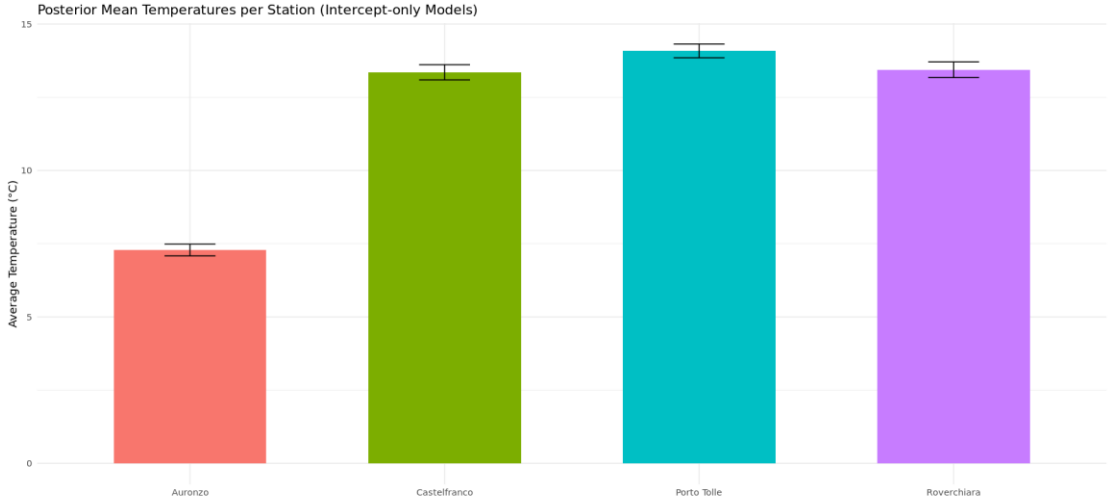


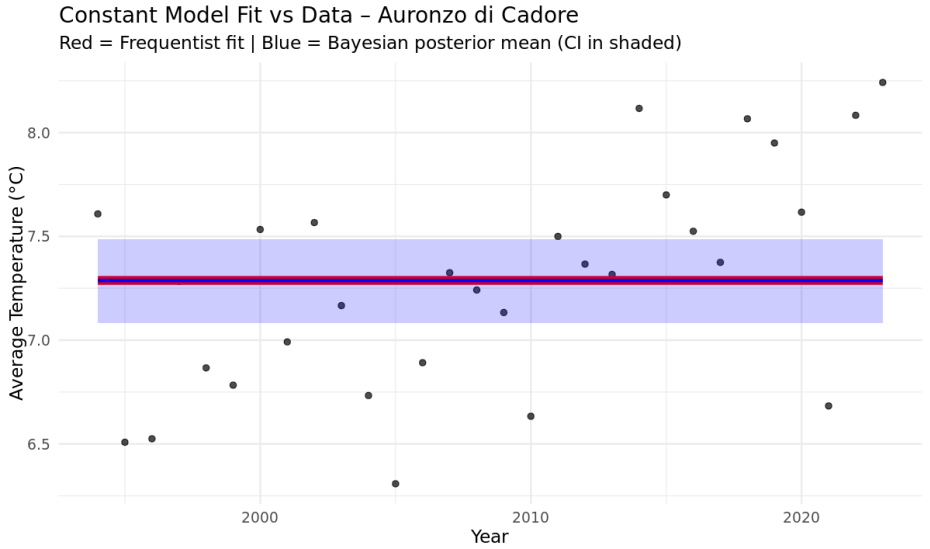


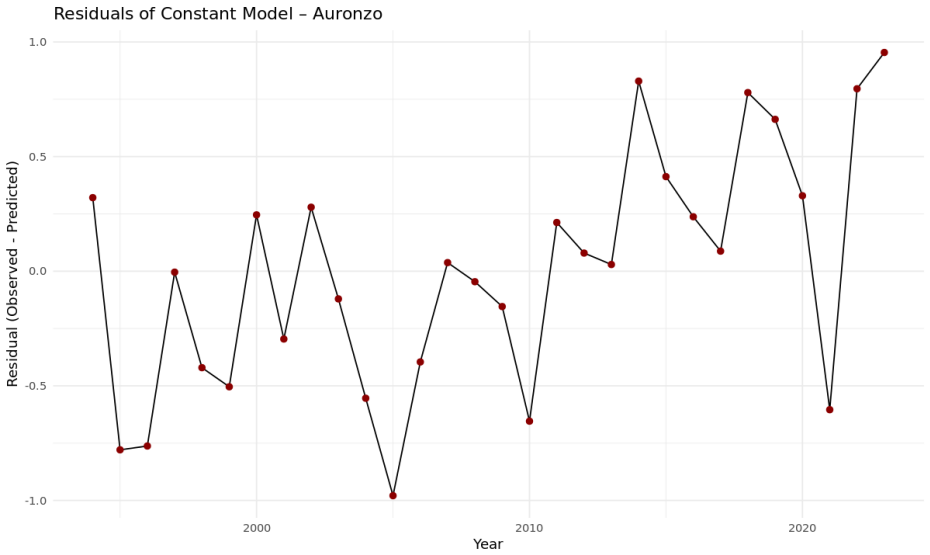
Parameters prediction	Auronzo Di Cadore	Castelfranco Veneto	Porto Tolle	Roverchiara
Intercept	7.29	13.35	14.08	13.44
sigma	0.54	0.72	0.65	0.72

Part 1: Constant regression







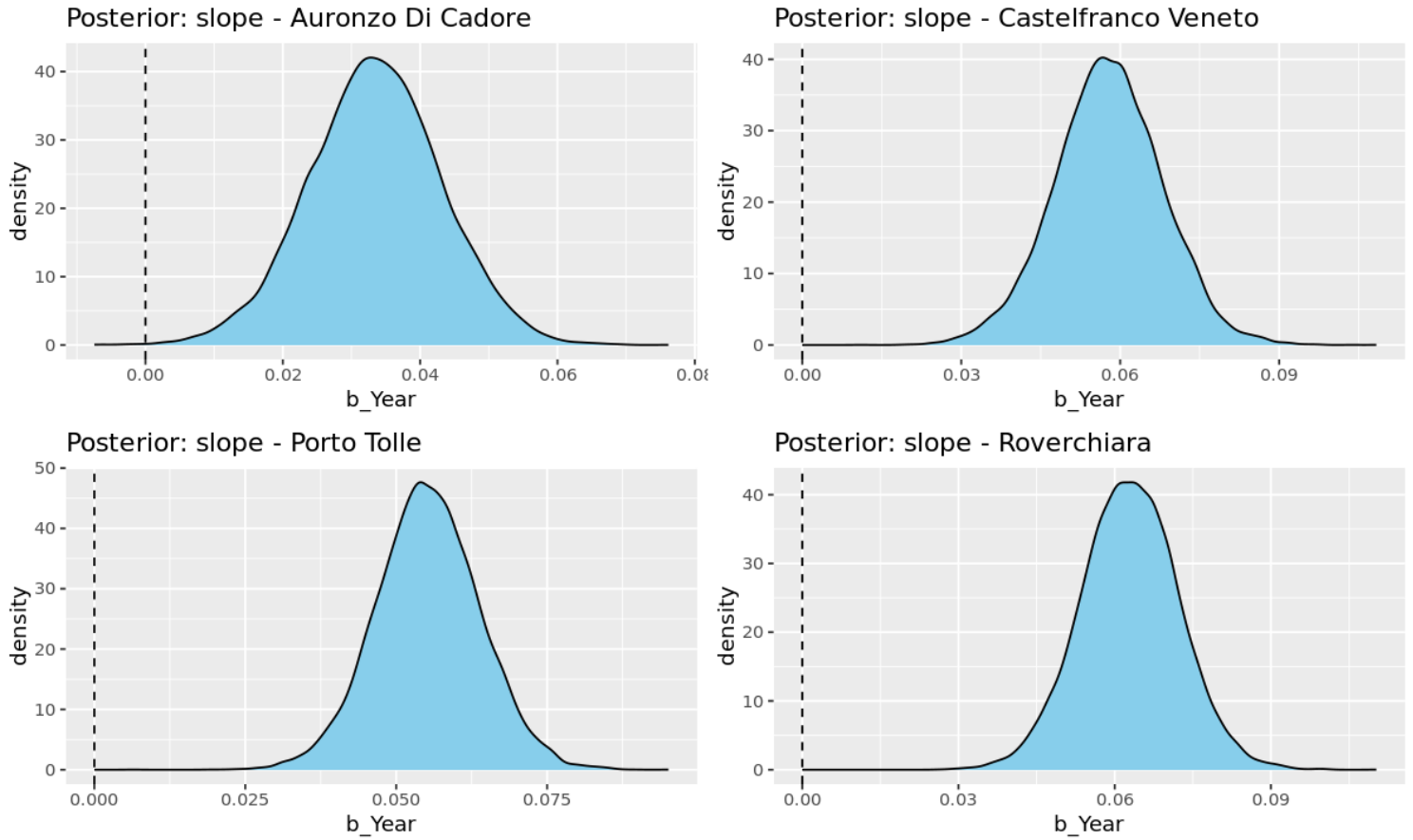


Part 1: Hypothesis Testing

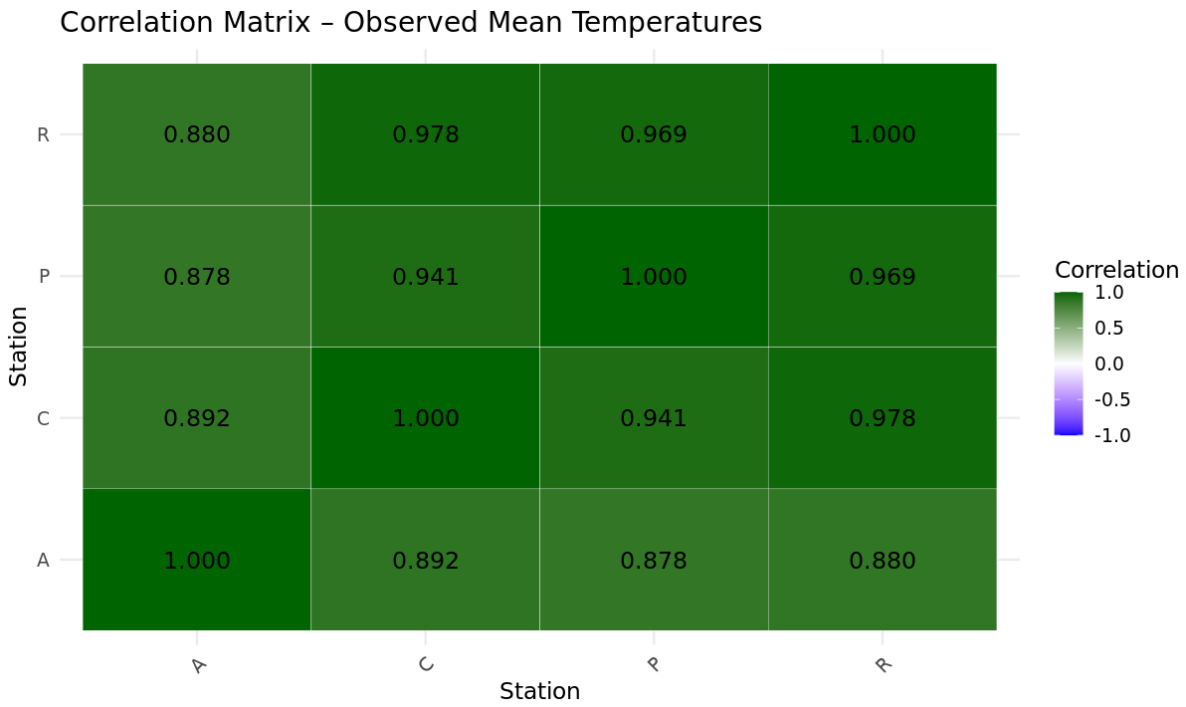
$H_0: b = 0$ (constant trend)

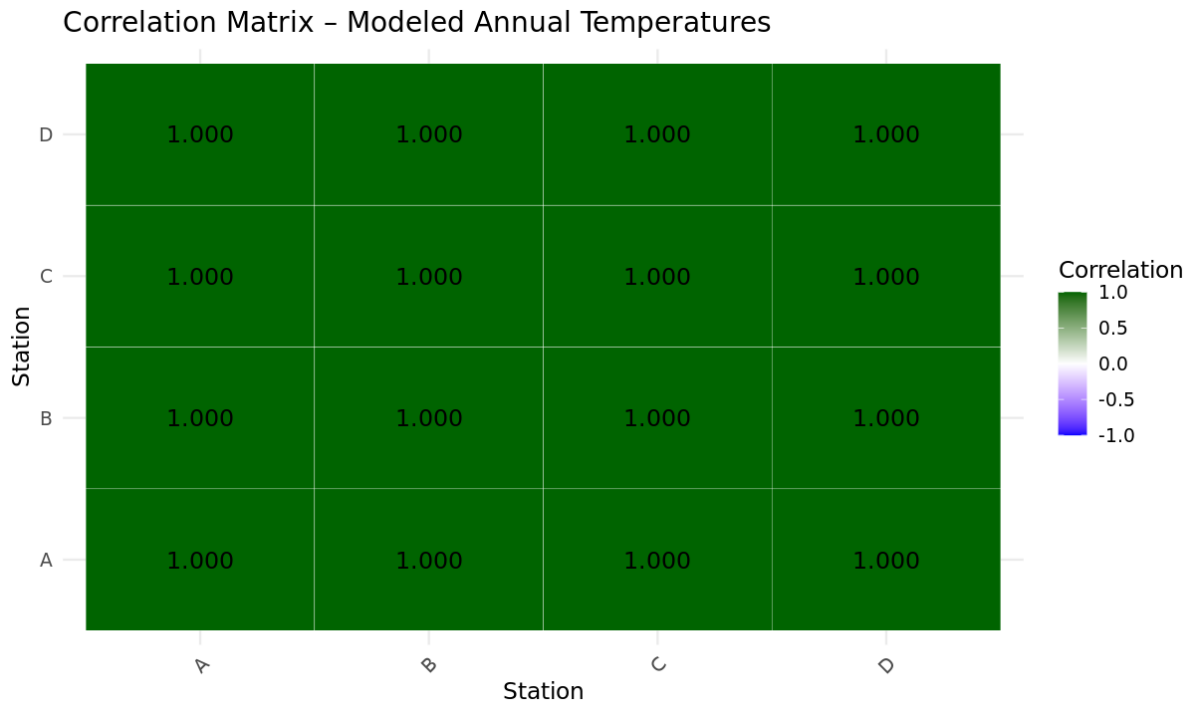
$H_1: b > 0$ (increasing trend)

	Auronzo di Cadore	Castelfranco Veneto	Porto Tolle	Roverchiara
$P(b > 0)$	0.9992	~ 1	~ 1	~ 1



Part 1: Correlation between stations





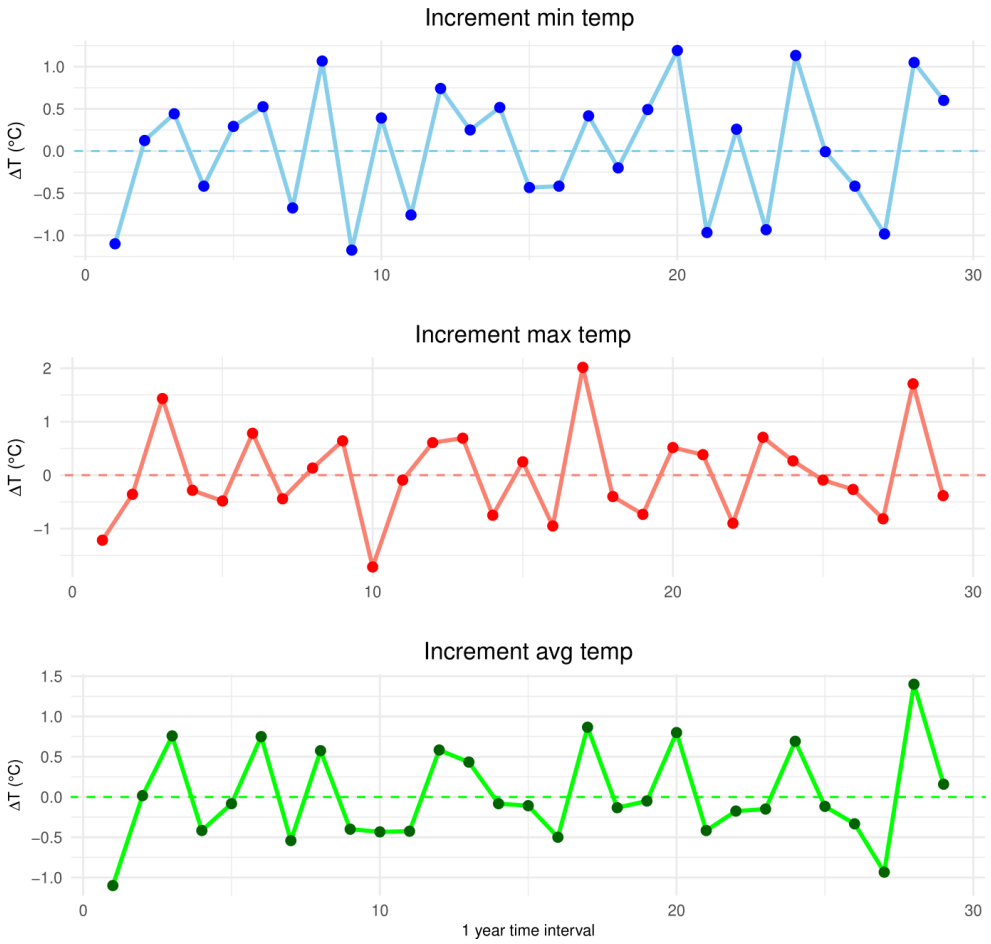
Part 2: Annual temperature changes

To further investigate the increasing trend in temperatures and compare it with official data, we define:

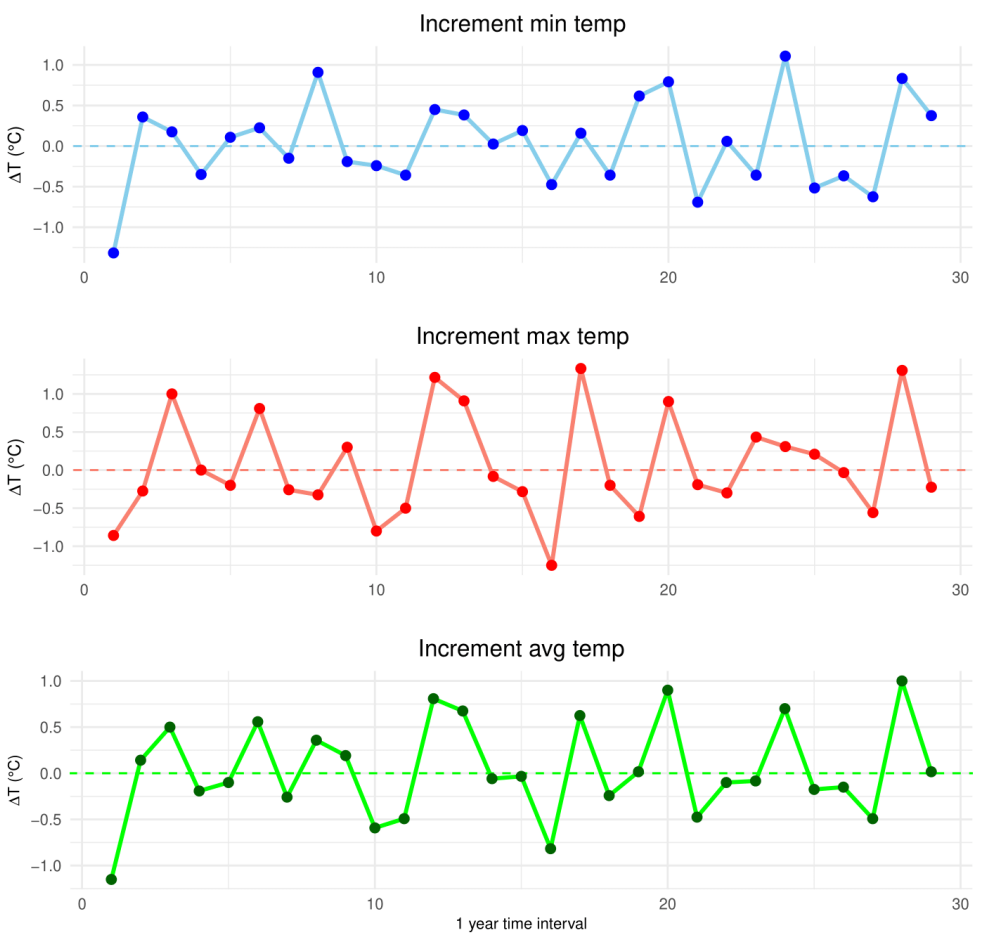
$$\Delta T_j = T_{j+1} - T_j$$

where T_j is one of the (mean) temperatures evaluated at year *j*

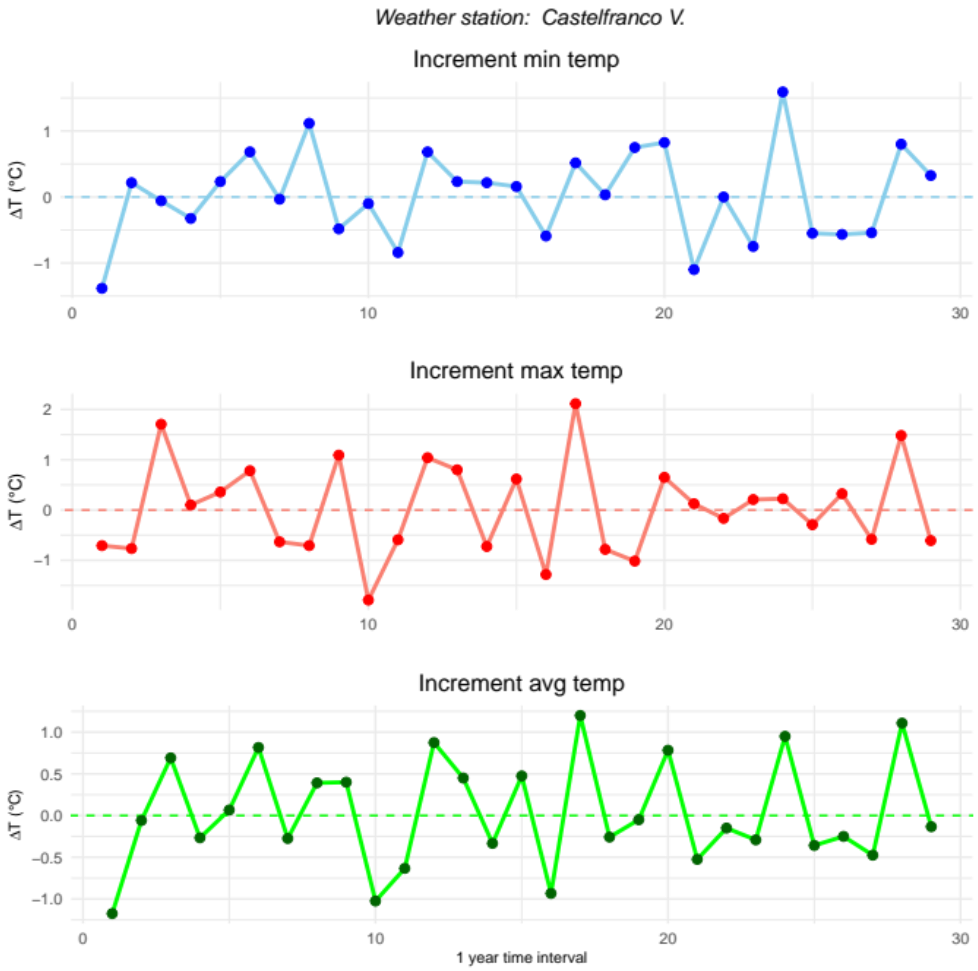
Weather station: Auronzo



Weather station: Porto Tolle



Part 2: Annual temperature changes



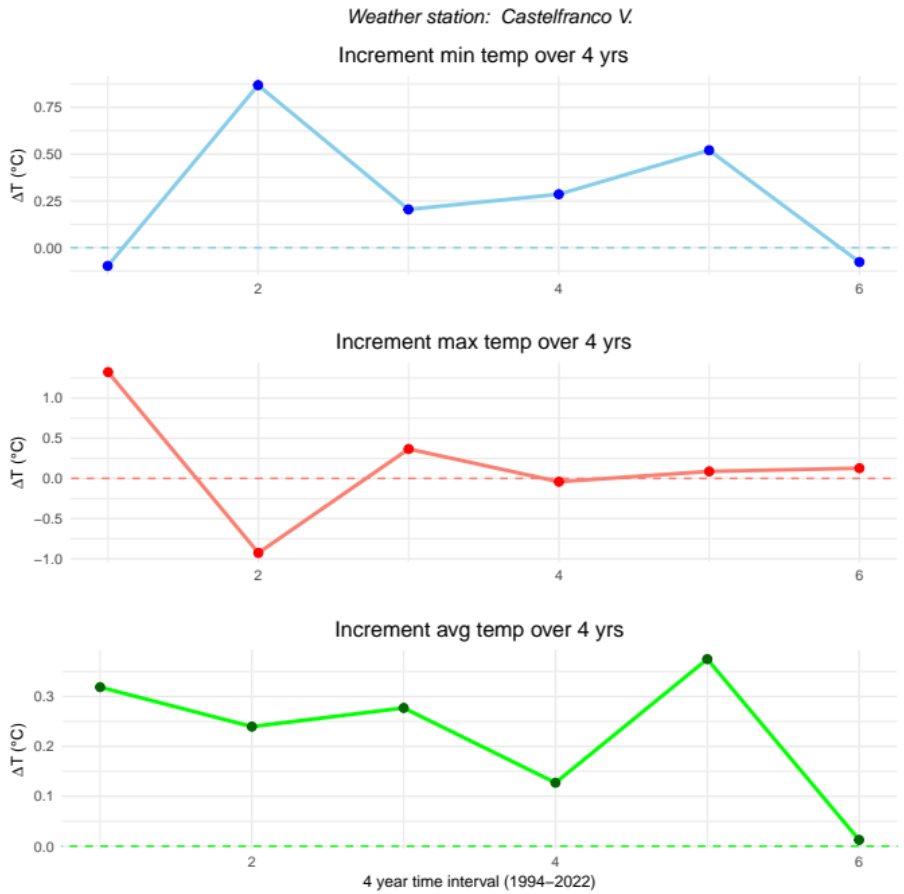
Let's compute the average ΔT for all stations to get an initial sense of the overall temperature changes

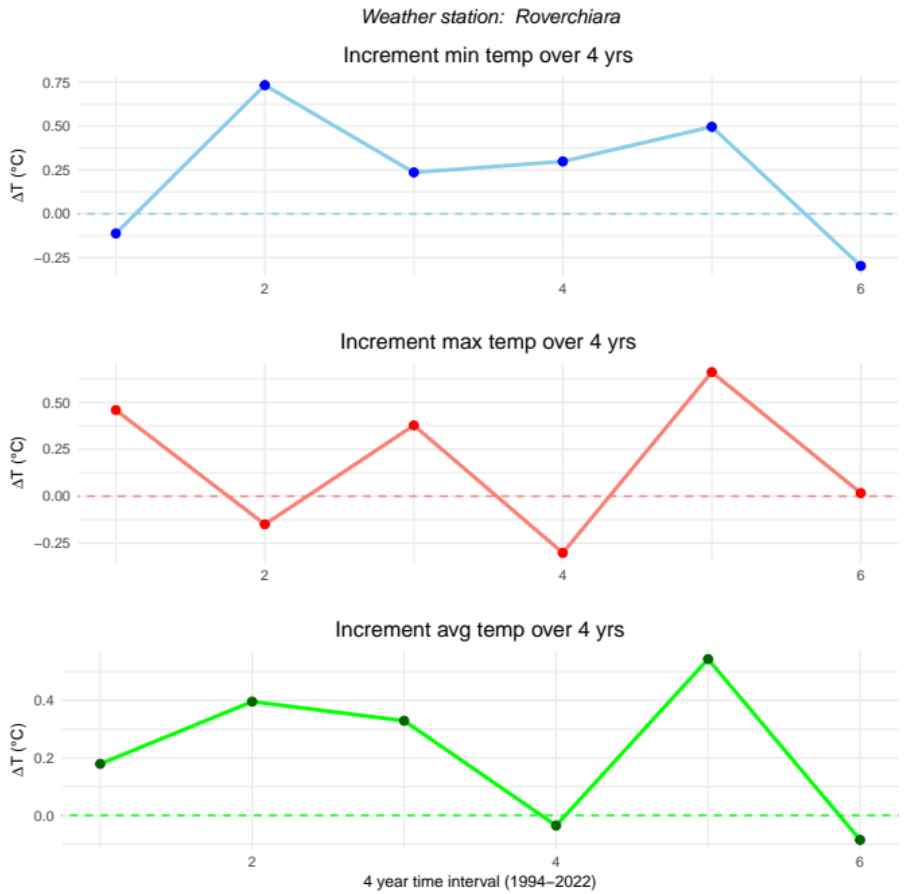
Weather station	$\Delta T_{\min}(^{\circ}C)$	$\Delta T_{\max}(^{\circ}C)$	$\Delta T_{\text{avg}}(^{\circ}C)$
Auronzo	+ 0.0089	+ 0.0348	+ 0.0218
Castelfranco V.	+ 0.0336	+ 0.036	+ 0.0351
Porto Tolle	+ 0.0612	+ 0.0264	+ 0.0374
Roverchiara	+ 0.0503	+ 0.0379	+ 0.0437

The average of ΔT is positive for all temperatures types and stations.

This is still a very basic and descriptive insight, not an evidence, just motivating further analysis.

Part 2: 4 Year time interval

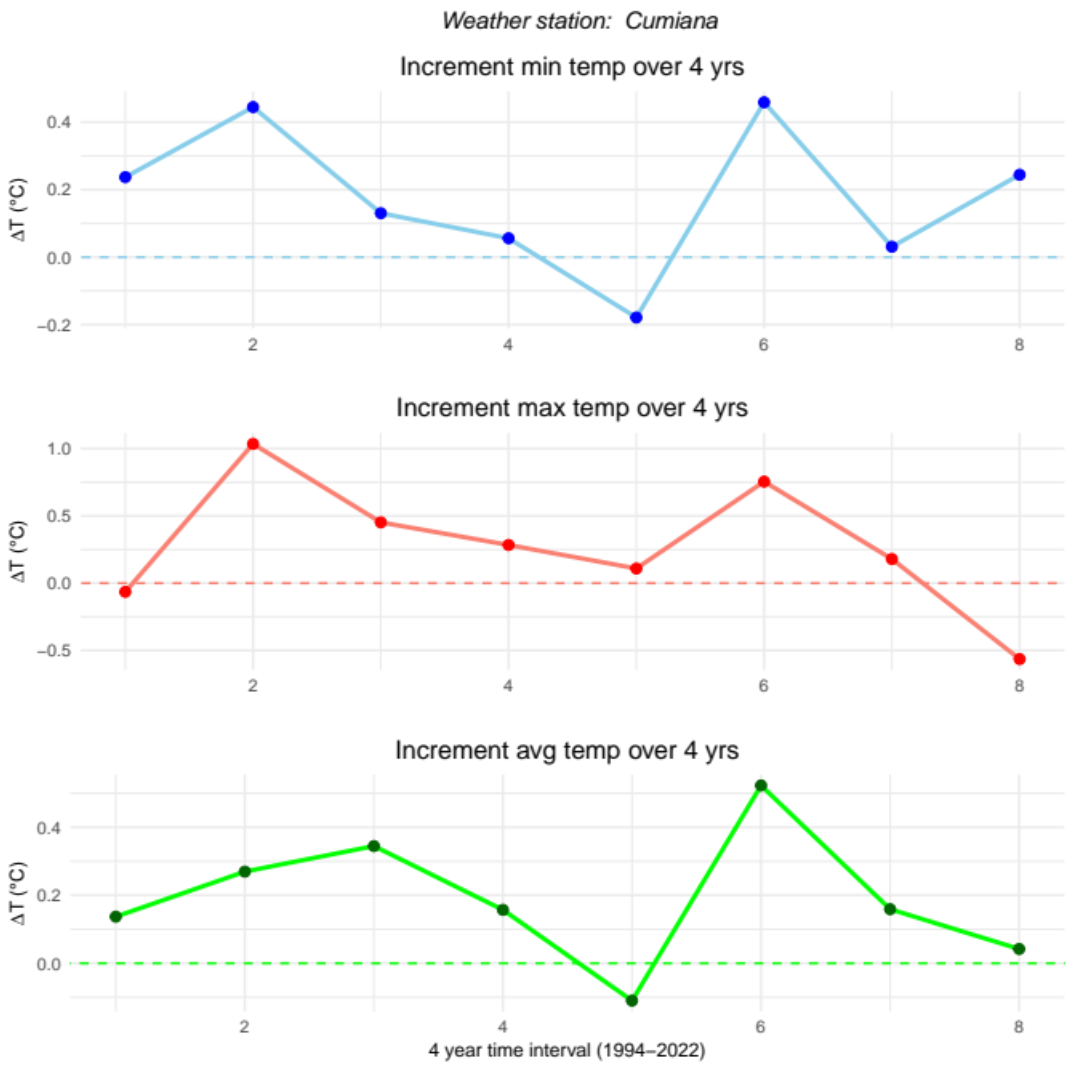




To reduce yearly noise and smooth out the signal, we'll group the time series into 4-year intervals (from 1994 to 2022)

Now the positive trend is a bit more evident, especially when it comes to the daily average temperature

Part 2: More data



ARPA Veneto provides climate data covering only the period from 1994 to the present. Since this results in only 6 points when grouped in 4-year intervals, we searched for older climatic data

Fortunately, ARPA Piemonte offers a bit more coverage (starting from 1987), so we performed our statistical analysis of 4 Piemontese stations too:

- **CUMIANA (*TO*) – 327 m a.s.l.**
- **ACQUI TERME (*AL*) – 217 m a.s.l..**
- **COLLE BARANT (*TO*) – 2294 m a.s.l.**
- **GARESSIO (*CN*) – 980 m a.s.l.**

This additional data will come in handy later, when we will compare regional trends to nation-wide predictions made by SNPA

PIEMONTE Arpa Agenzia Regionale per la Protezione Ambientale

ARPA Piemonte Open data

Part 2: Bayesian framework and MCMC analysis

Let us now perform a linear regression with STAN over ΔT

- **Generative model: Following the scientific literature, we assume that the *temperature increase is constant over time* (the same assumption adopted by SNPA). Even if the true trend were accelerating, it would appear approximately linear within such a limited temporal span**

- **Parameters: We model our data as**

$$\Delta T_j \sim \mathrm{Norm}(b, \sigma^2)$$

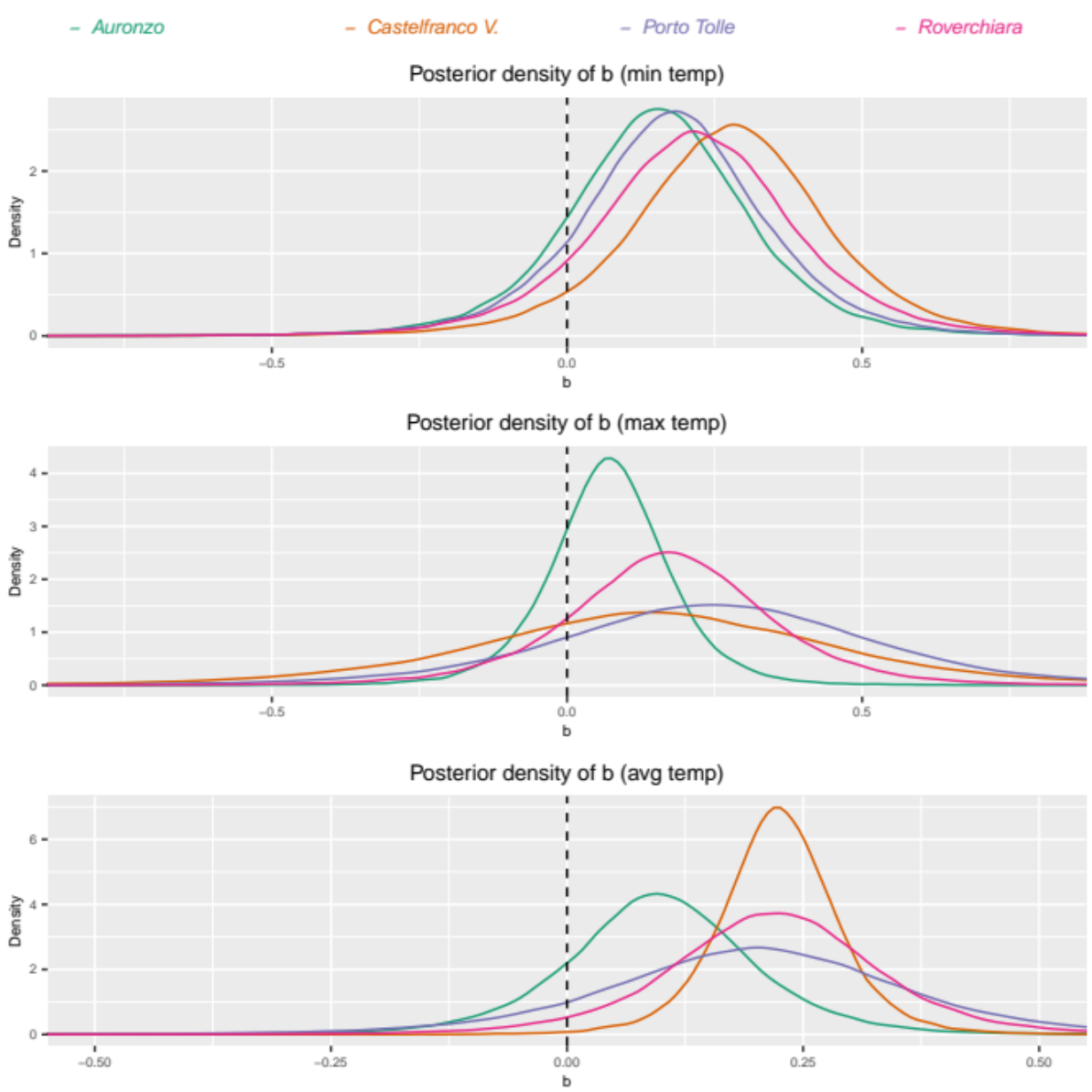
Gaussian inference with unknown variance

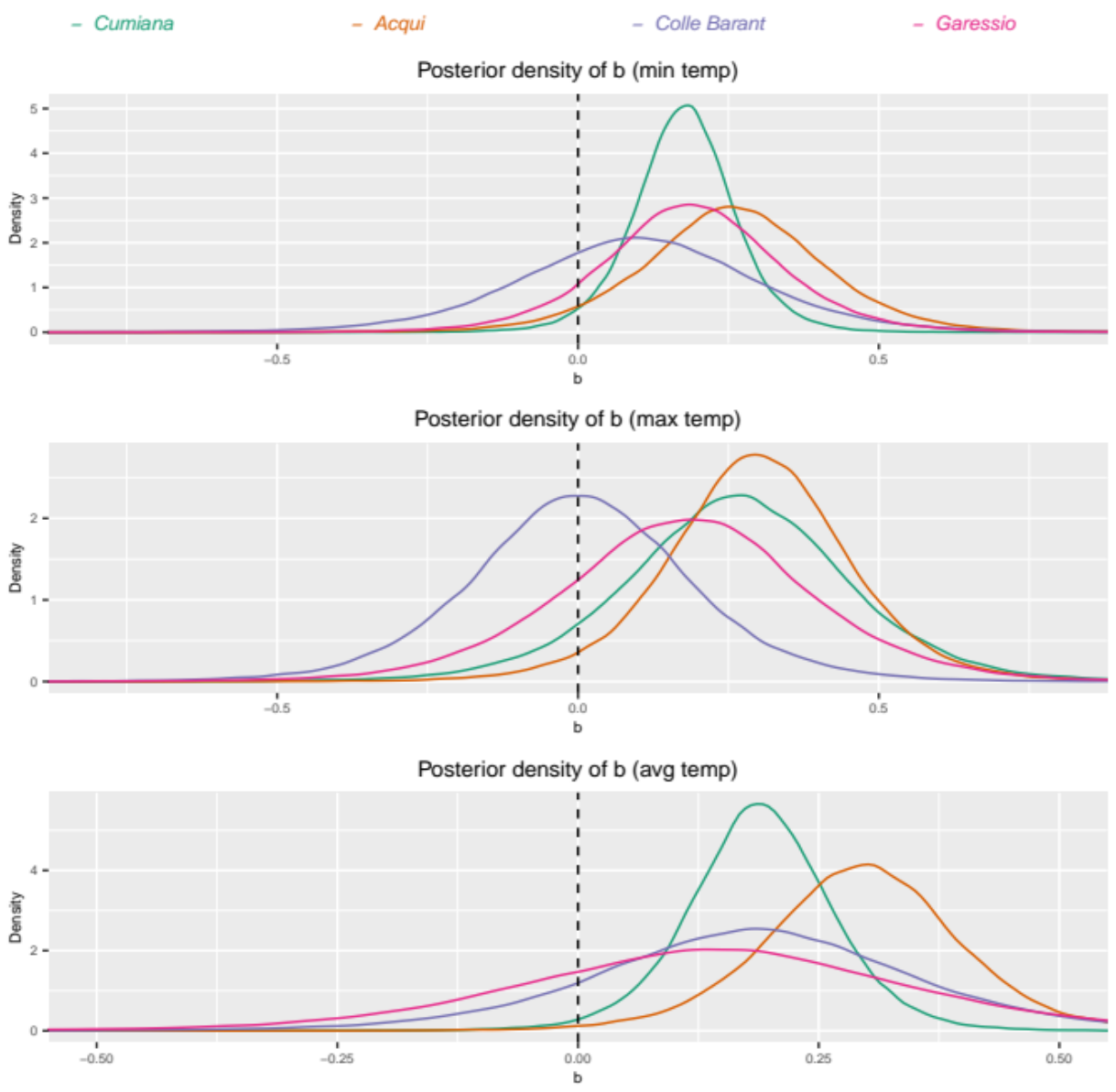
where b ($\mathcal{C}$ / 4 yrs) is the 4-year trend in temperature (min, max or daily) and σ represents its variability

- **Priors: We use a normal prior for b centered around 0 (we want to be unbiased) and an uniform prior in $[-1, 1]$. For the scale parameter, a Jeffreys prior is used:**

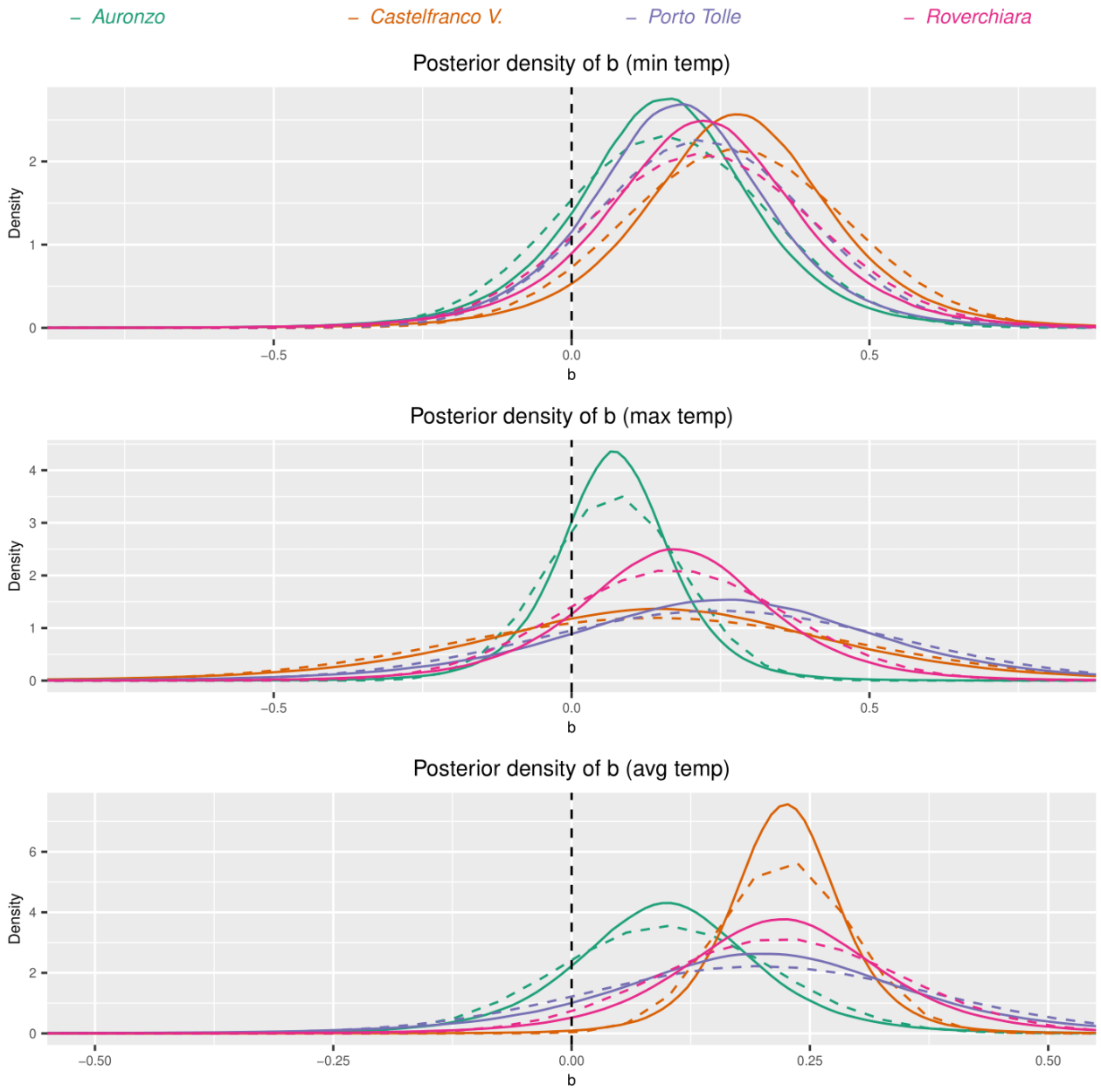
$$P(b) \sim \mathrm{Norm}(0,1), \qquad \mathrm{Unif}(-1,1), \qquad P(\sigma) \sim \frac{1}{\sigma}$$

Part 2: Posterior on b





Part 2: Posterior on b – normal or not normal?



A gaussian inference problem with unknown variance doesn't produce normal posteriors, especially when the samples are few.

Instead, a *t-student*-like distribution comes up (although not precisely a t-student…)

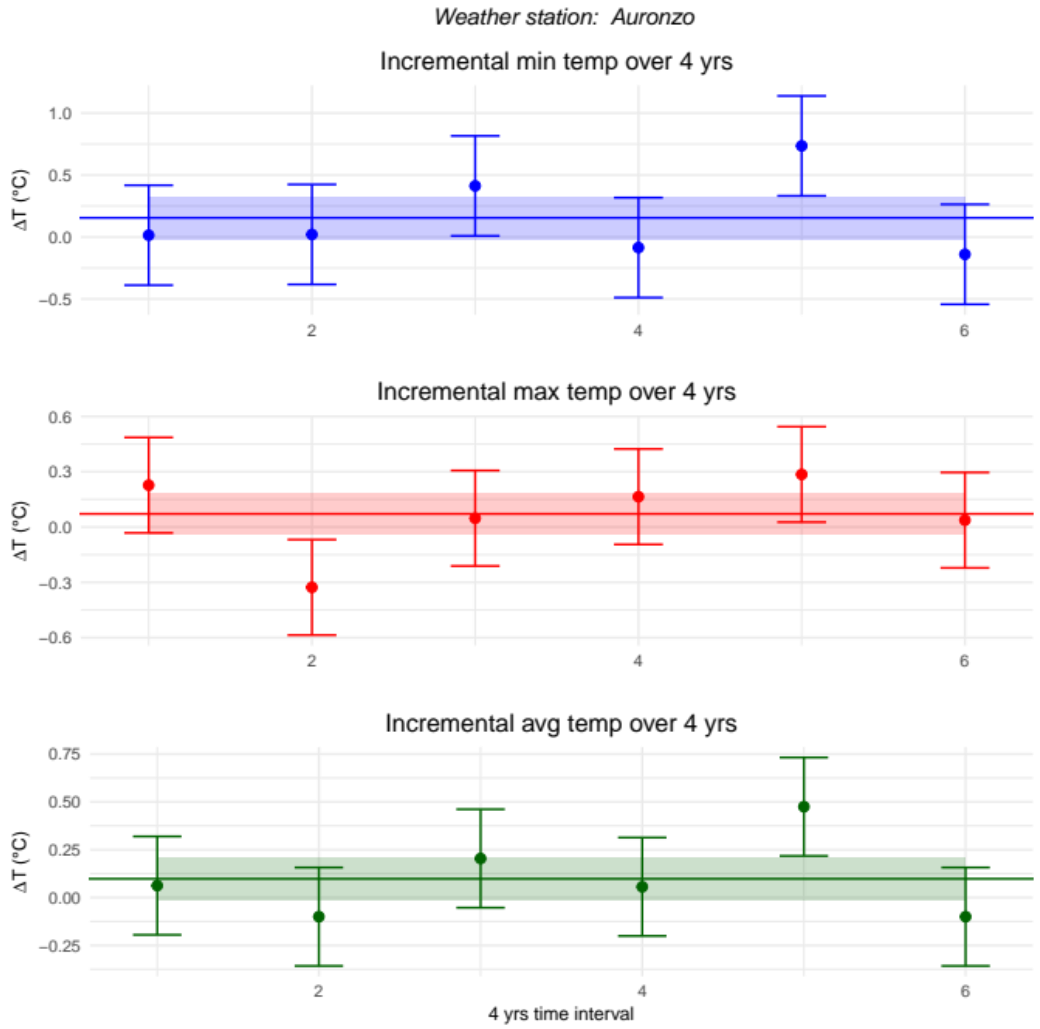
$$P(\mu, \sigma\{x_i\}) \propto P(\mu, \sigma)P(\{x_i\}\mu, \sigma)$$

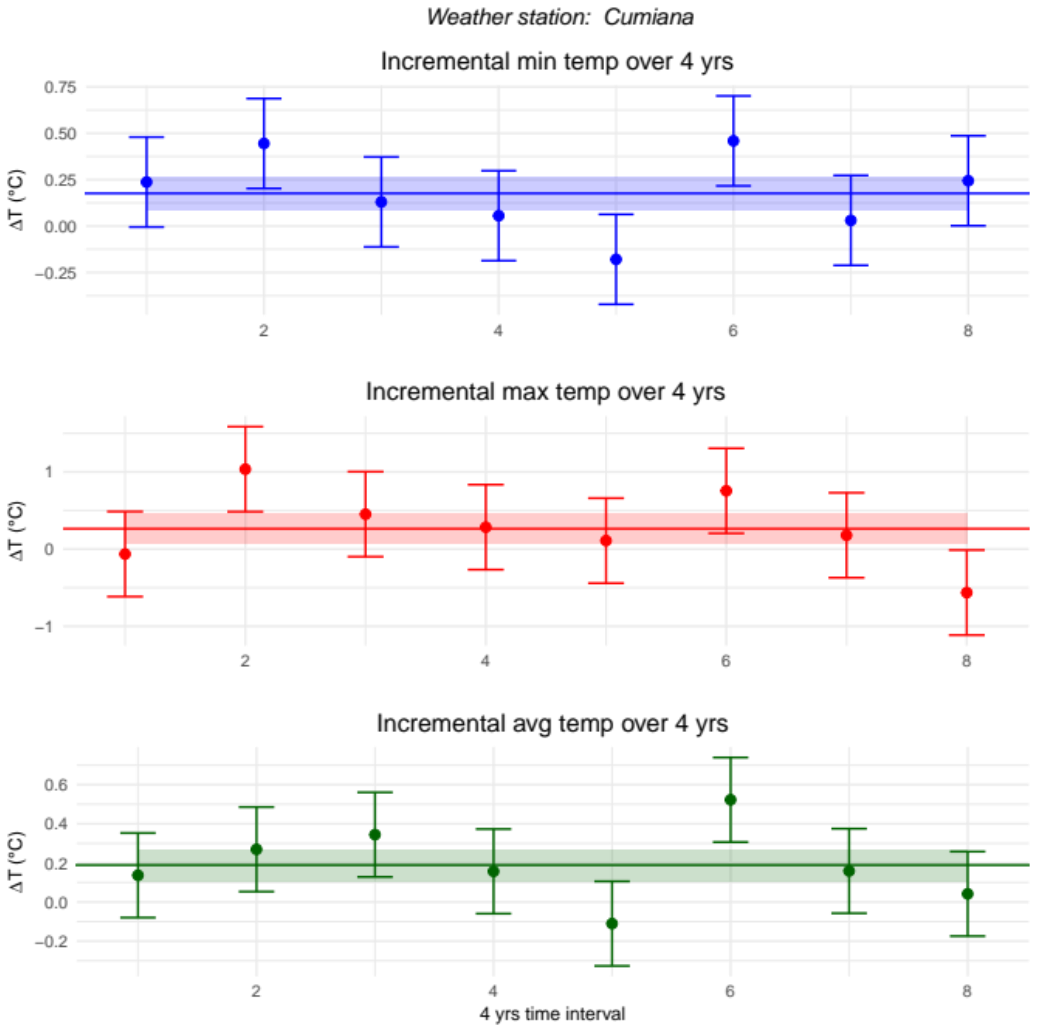
$$P(\mu, \sigma\{x_i\}) \propto \frac{1}{\sigma^n} exp\left(-\frac{1}{2}\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{\sigma^2}\right)$$

$$P(\mu) \propto \int_0^{\infty}\left(\frac{1}{\sigma^n} exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right)\right)d\sigma$$

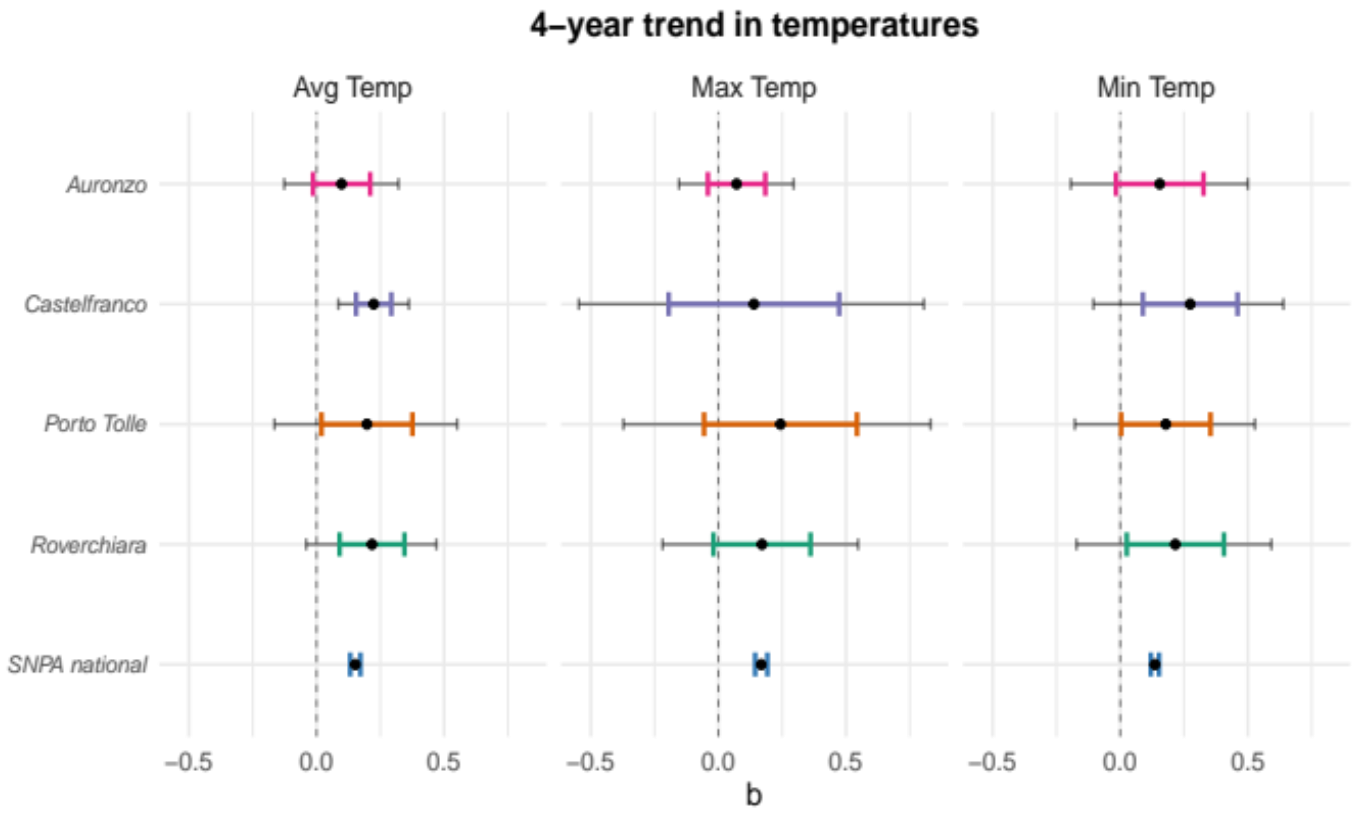
Part 2: Fitted time series

The continuous colored line represents the posterior mean of b, while the surrounding shaded area indicates its posterior standard deviation. The error bars are evaluated as posterior averages over the parameter σ





Part 2: Comparing trends with SNPA



SNPA periodically publishes reports in coordination with regional ARPA agencies.

In our analysis, we refer specifically to the 2021 report*. According to this report:

"La stima del rateo di variazione della temperatura media dal 1981 al 2019 è di +0,38 ± 0,05°C/10 anni; il rateo di variazione della temperatura massima (+0,42 ±0,06°C/10 anni) è maggiore di quello della temperatura minima (+0,34 ± 0,04°C/10 anni)"

Sistema Nazionale per la Protezione dell'Ambiente

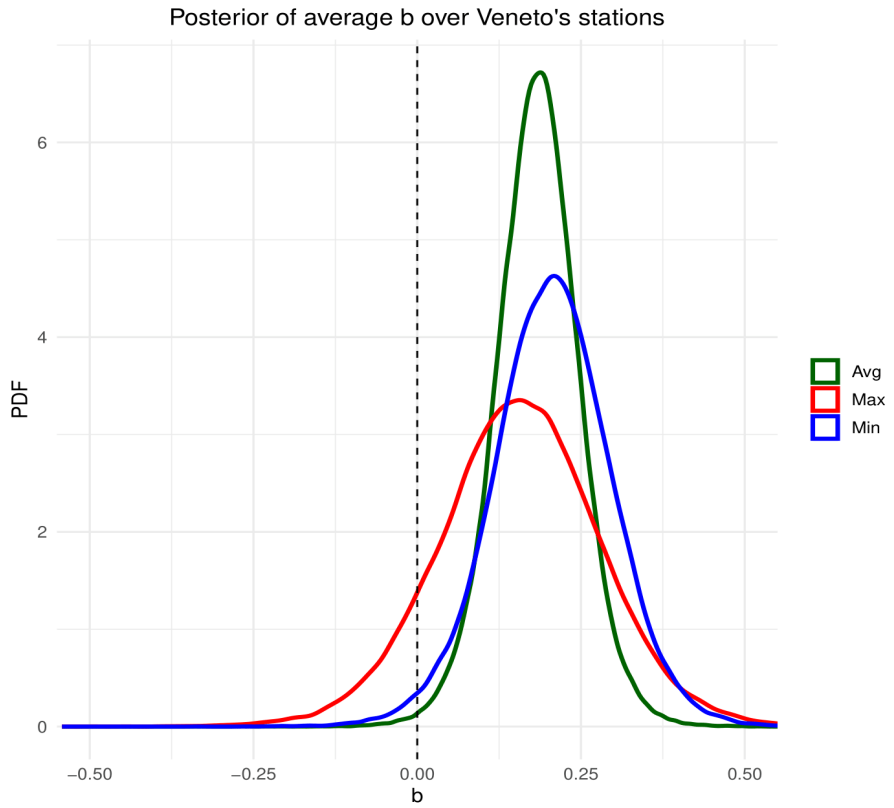
(*) https://www.snpambiente.it/wp-content/uploads/2021/06/Rapporto-SNPA-21_2021.pdf

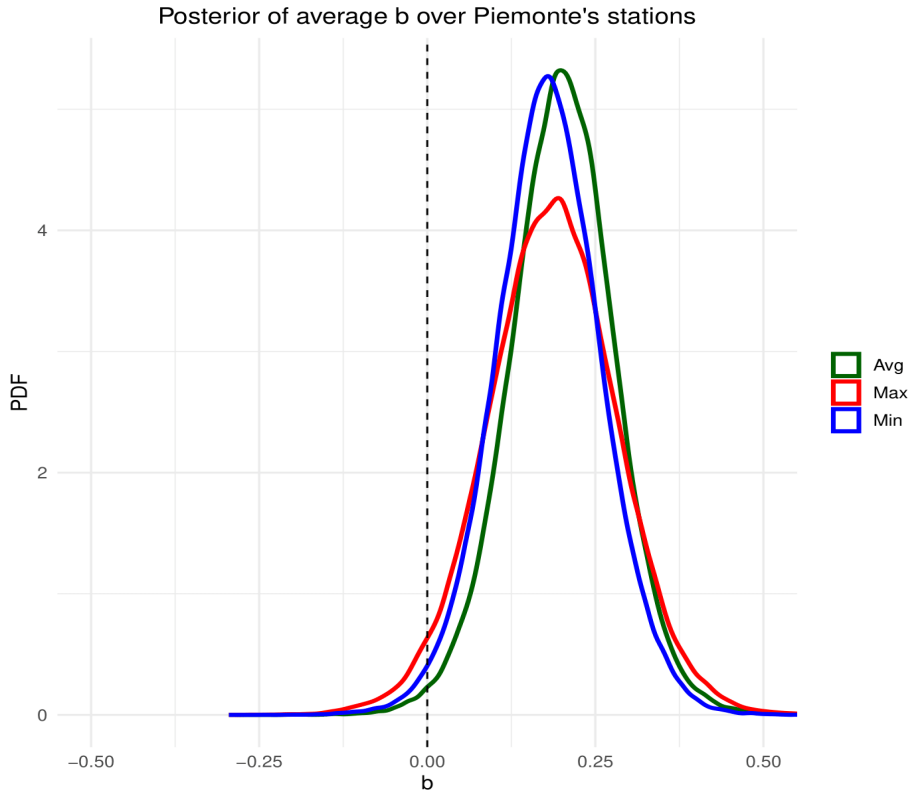
Part 2: Averaging over stations

We'd like to extract a single value for each region by averaging over all stations, i.e. get the regional trend:

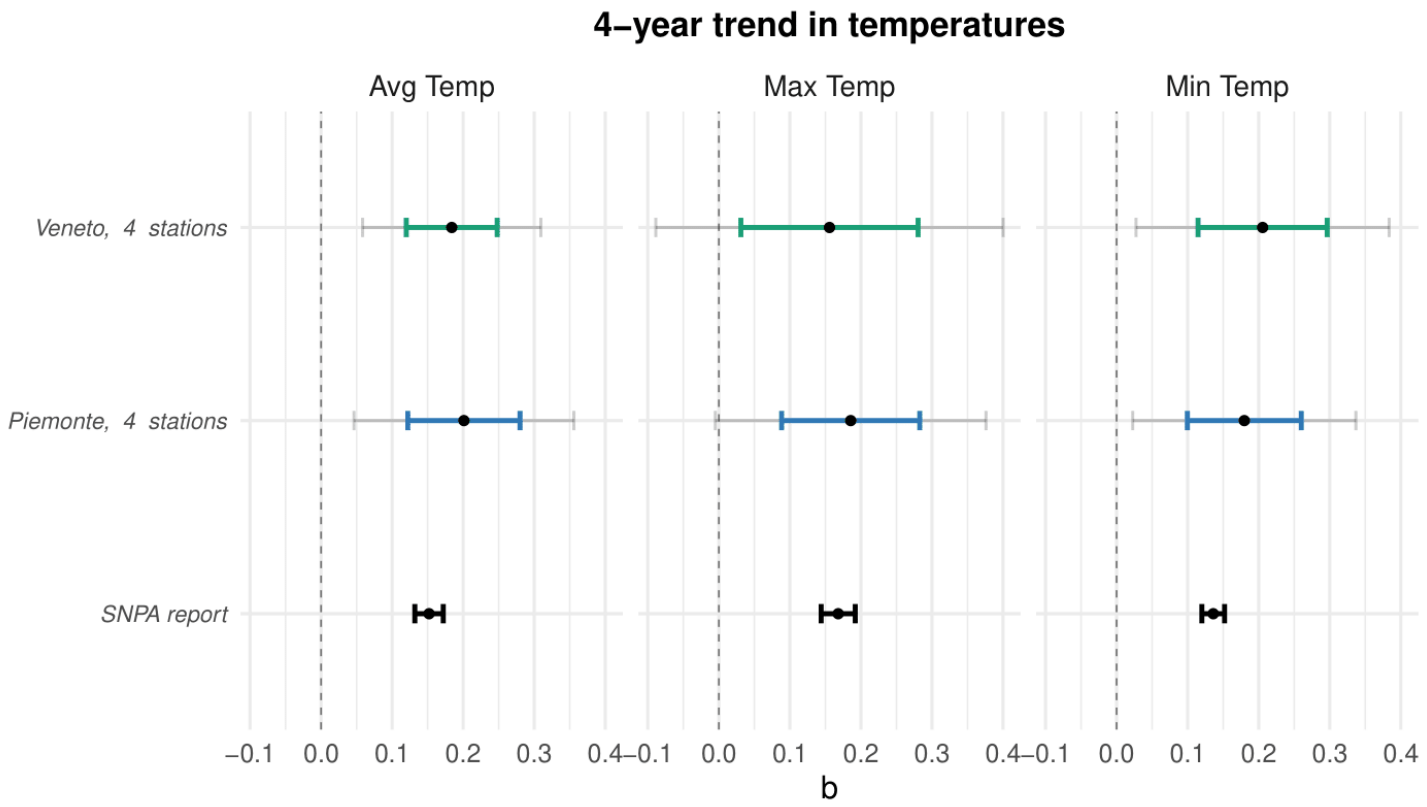
$$b_{reg} = \frac{1}{4}(b_1 + b_2 + b_3 + b_4)$$

Unfortunately, b_i is not normally distributed and the above average can't be computed in a close form except for its expected value and variance. Still, we can estimate it numerically with MCMC





Part 2: Final results

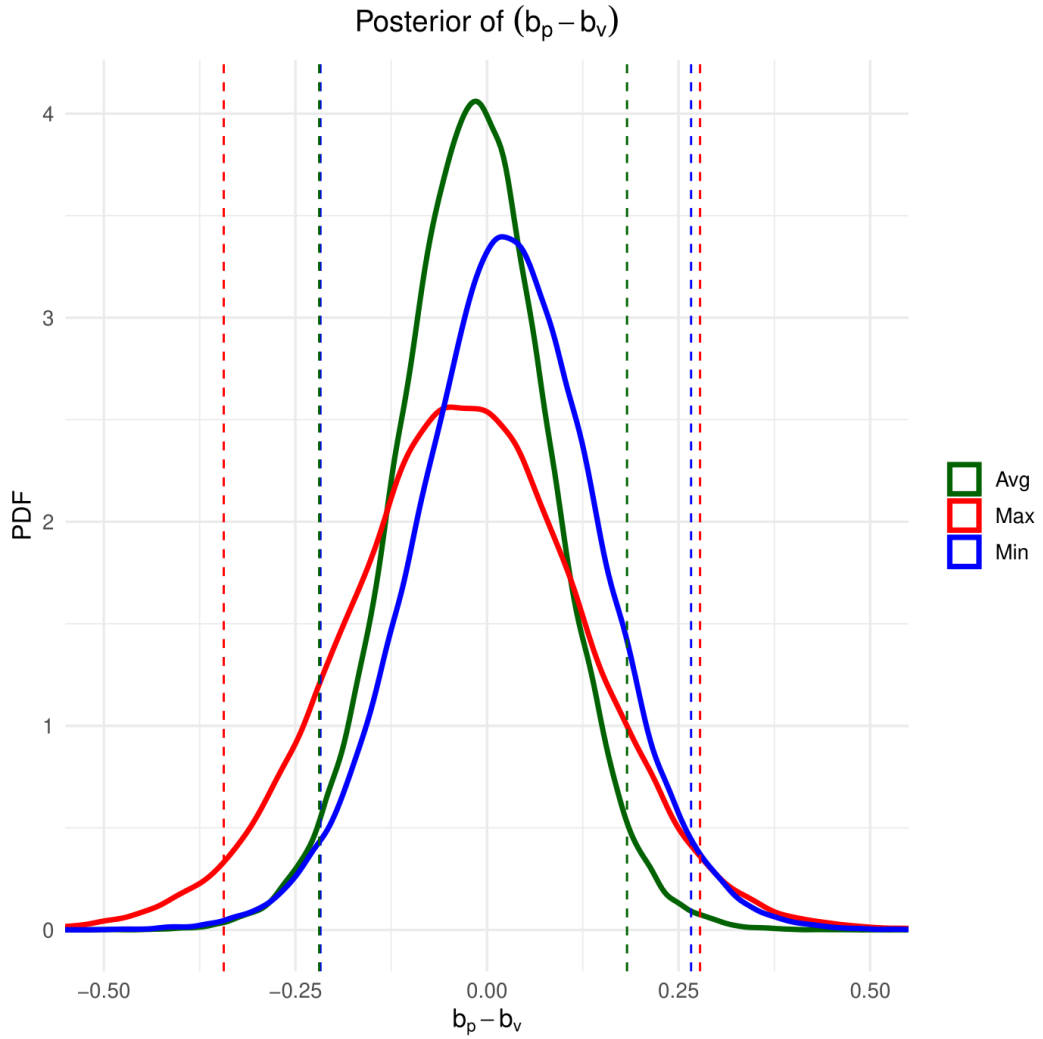


Weather station	$\Delta T_{\min}(^{\circ}C\ /\ 4\ \text{yrs})$	$\Delta T_{\max}(^{\circ}C\ /\ 4\ \text{yrs})$	$\Delta T_{\text{avg}}(^{\circ}C\ /\ 4\ \text{yrs})$
Avg Veneto	0.21 ± 0.09	0.16 ± 0.12	0.18 ± 0.06
Avg Piemonte	0.18 ± 0.08	0.19 ± 0.09	0.20 ± 0.08
SNPA Report	0.136 ± 0.016	0.17 ± 0.02	0.15 ± 0.02

Comparing everything with the national-scale SNPA data …

- **At a 5% level of significance, at least for the min and daily avg temp, we can reject the null hypothesis H_0: $b_{reg} = 0$**
- **The results are quite satisfying and appear to be consistent across regions. However, one should test this hypothesis**
- **Results from SNPA always lie in the 95% C.I., so we can't reject the null hypothesis: $H_0: b_{reg} = b_{SNPA}$**

Part 2: Comparisons between regions



We have samples coming from the posterior distribution of b_p, b_v (MCMC w/ STAN). It's thus easy to build the posterior distribution of the random variable $b_p - b_v$ and test whether 0 lies in the acceptation region with a significance level of 5%

As a result, we can't reject the null hypothesis $H_0: b_p - b_v = 0$ and we don't have evidence that support differences in the two regions

$ARIMA(p, d, q)$

AutoRegressive Integrated Moving Average

Way to describe non-stationary stochastic proceses X_t

- $AR(p)$ Autoregressive: the next value of X at time t is given by a linear combination of previous p variables
- $I(d)$ Integrated: differentiation order d in order to make the time series trend stationary
- $MA(q)$ Moving average: the next error term is given by a linear combination of the previus q white noises

$$X_t = \varepsilon_t + \underbrace{\sum_{i=1}^{p} \varphi_i X_{t-i}}_{AR(p)} + \overbrace{\sum_{i=1}^{q} \theta_i \varepsilon_{t-i}}^{MA(q)}$$

For time series with a strong seasonal component SARIMA (seasonal ARIMA) is used

$$SARIMA(p, d, q)(P, D, Q)_m$$

$AR(P), MA(Q)$ are the same as ARIMA but specified at seasonal lag. $I(D)$ accounts for the number of differentiation to remove the seasonality. m is the number of observations per year

Hyndman, R.J., & Athanasopoulos, G. (2018) Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia.

Part 3: Differentiation order d and D

Non-seasonal first-order differentiation $X'_t = X_t - X_{t-1}$

```
diff(x, lag = 1, differences = d)
```

Seasonal first-order differentiation $X'_t = X_t - X_{t-m}$

```
diff(x, lag = m, differences = D)
```

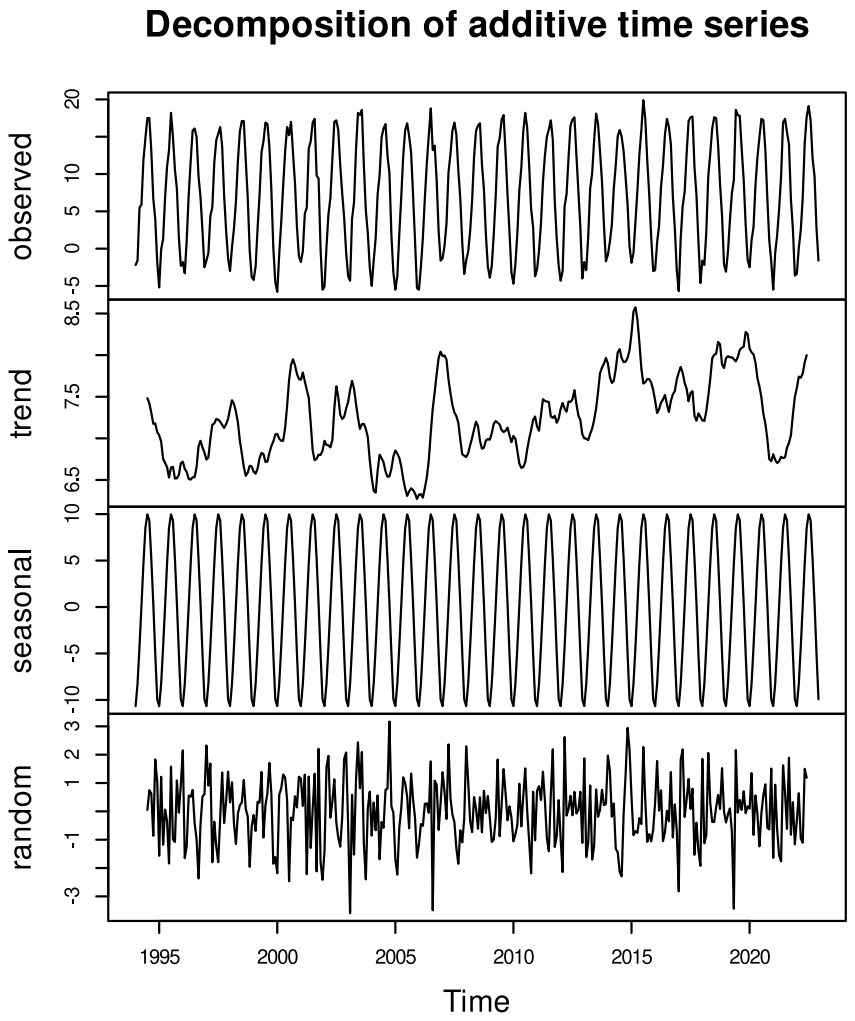
We can determine the number of non-seasonal d and seasonal D differences required for time series to became stationary: KPSS (Kwiatkowski–Phillips–Schmidt–Shin) tests:

- H_0: the time series is stationary around a deterministic trend
- H_1: the time series has a unit root (not stationary)

```
d <- ndiffs(x, alpha = 0.05)
D <- nsdiffs(x, alpha = 0.05)

Differentiation order d = 0
Seasonal differentiation order D = 1
```

This analysis suggests an $ARIMA(p, 0, q)(P, 1, Q)_{12}$ model



Auronzo di Cadore 1994-2022 average temperatures data

Part 3: AR and MA parameters

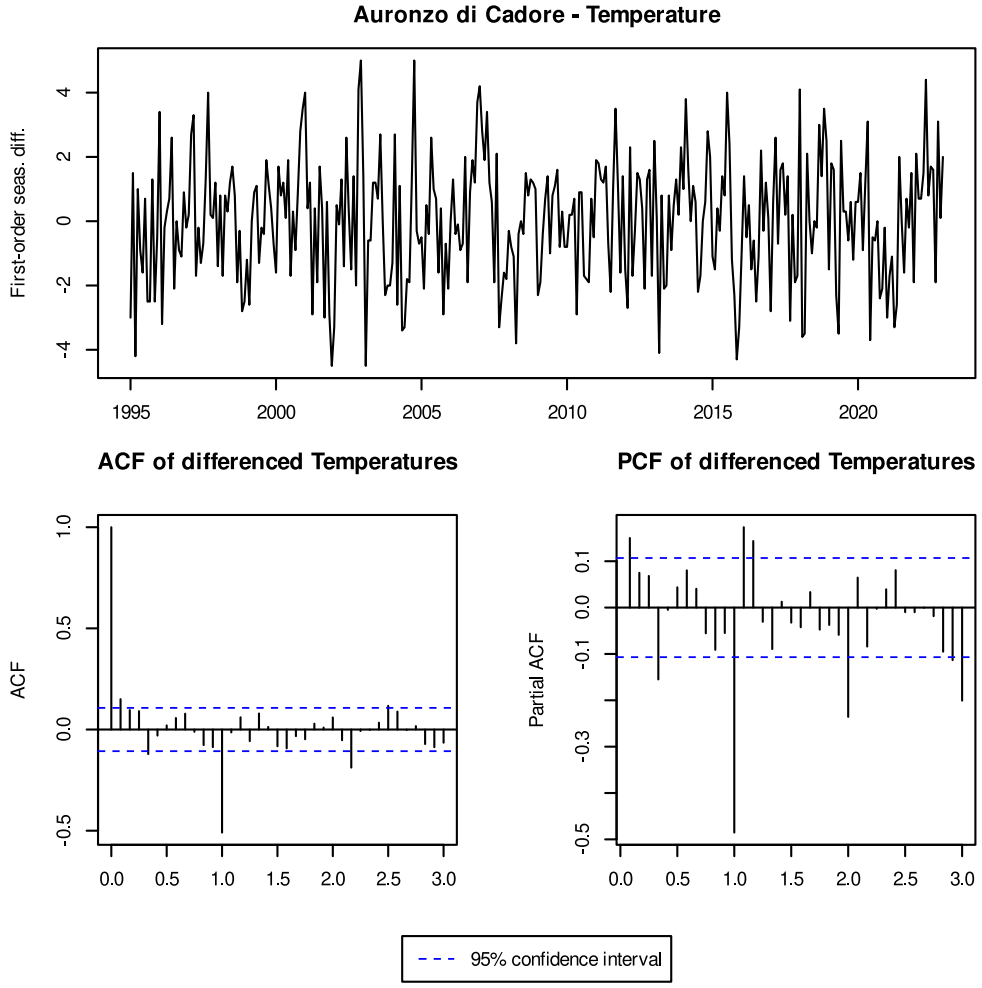
The autocorrelation function (ACF) shows the relationship between X_t and X_{t-k} for different lags k. The partial autocorrelation function (PACF) measure the relationship between X_t and X_{t-k} after removing the effects of lags $1, \dots, k-1$

↓ ACF and PACF can provide qualitative information about the AR and MA orders

- $AR(p)$: if the ACF is exponentially decaying and there is a significant spike at lag p in the PACF but none beyond lag p
- $MA(q)$: if the PACF is exponentially decaying and there is a significant spike at lag q in the ACF but none beyond lag q

Seasonality complicates distinguishing between AR and MA components, though their values should be small given the fast decay of both ACF and PACF.

The right figure shows prominent spikes at a one-year lag, suggesting non-zero values for seasonal orders P and/or Q



Part 3: AR and MA parameters

We perform an $ARIMA(p,1,q)(P,1,Q)_{12}$ test, evaluating various parameter combinations for $(p,q,P,Q) \in \{0,1,2\}$ and recording the best results

We use the Akaike Information Criterion (AIC) to evaluate model fit. AIC balances model likelihood $\mathcal{L}$ with the number of estimated parameters N, penalizing complexity:

$$AIC = 2N - 2\ln\mathcal{L}$$

This analysis suggests an $ARIMA(1,0,1)(0,1,1)_{12}$ model with AIC of 1169.568

```
order.val <- c(p, 0, q)
seasonal.val <- list(order = c(P, 0, Q), period = 12)
fit <- Arima(ts, order.val, seasonal.val)
aic <- AIC(fit)
```

```
Series: ts ARIMA(1,0,1)(0,1,1)[12] Coefficients:
          ar1      ma1      sma1
       0.5882 -0.4103 -0.9453
s.e.   0.1646  0.1821  0.0473
sigma^2 = 1.732: log likelihood = -580.78
AIC=1169.57  AICc=1169.69  BIC=1184.84
```

While the `forecast` package in R offers `auto.arima`, it has limitations: it doesn't allow applying the estimated model to new data and can sometimes yield suboptimal results. For instance, `auto.arima` suggests $ARIMA(0,0,3)(2,1,0)_{12}$ model, but its AIC of 1232.97 is higher than our best-performing model

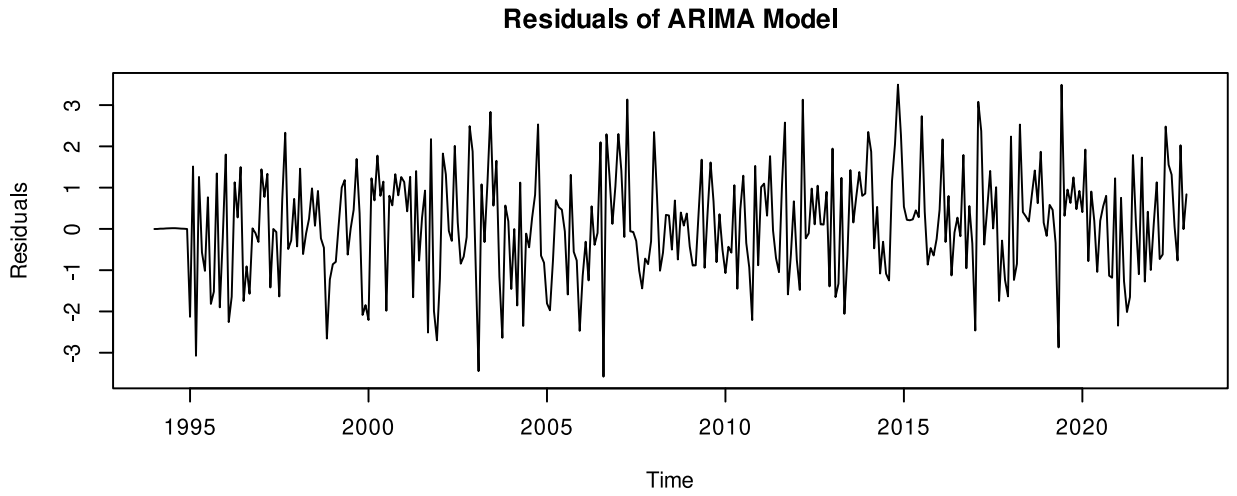
Part 3: AR and MA parameters

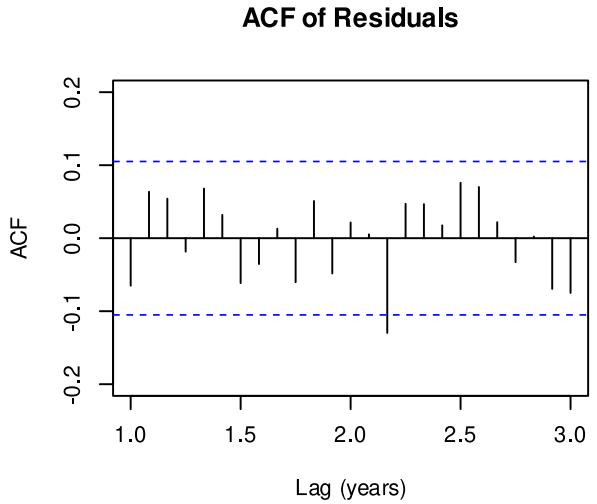
As a sanity check, we confirmed that the residuals behave as Gaussian white noise. The Ljung-Box test quantitatively verifies this by checking if residual autocorrelations are zero.

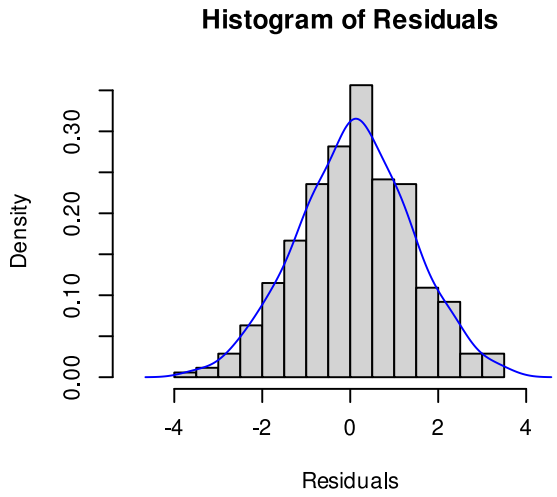
```
        Ljung-Box test

data:  Residuals from ARIMA(1,0,1)(0,1,1)[12]
Q* = 19.825, df = 21, p-value = 0.5323
```

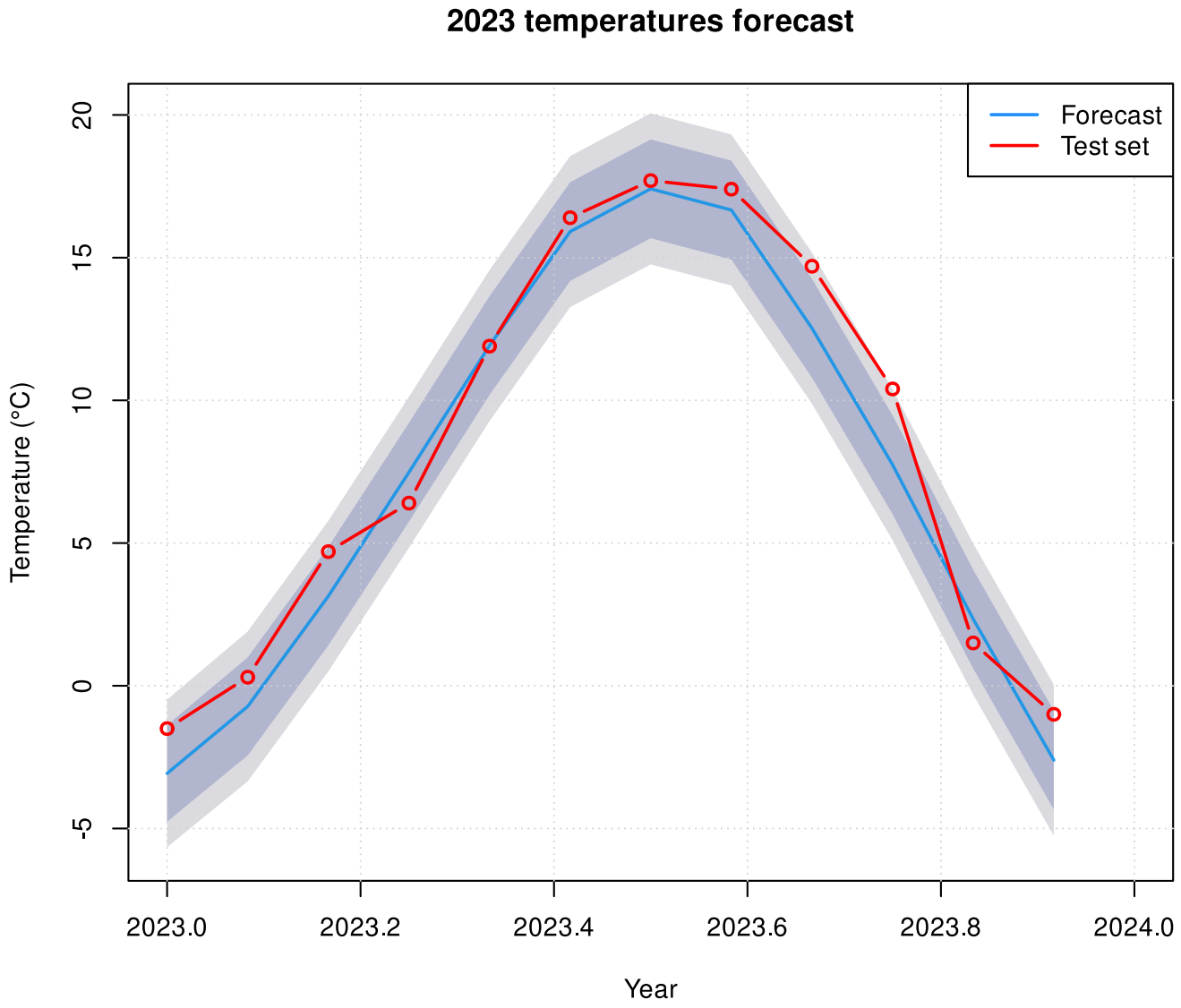
We fail to reject the null hypothesis. This indicates that the ARIMA model has adequately captured the autocorrelations in the data





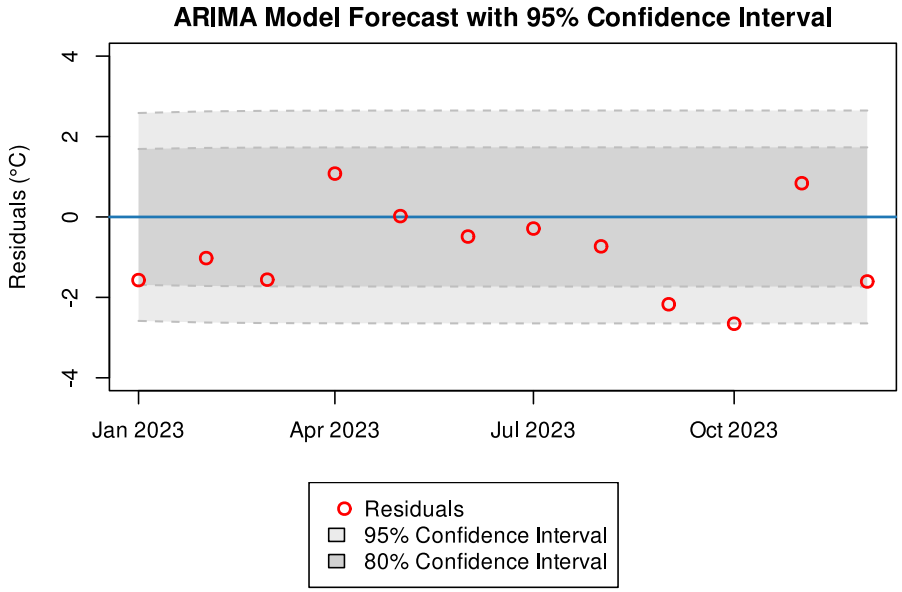


Part 3: Forecasting



ARIMA model effectively forecasts 2023 temperatures, with actuals data within the 95% confidence interval

The Root Mean Squared Error is 1.38°C.
The t-test on the forecast residuals (p=0.027) indicates a possible bias. Specifically, our forecast under-estimates the temperature by an average of -0.85°C



Auronzo di Cadore 2023 average temperatures data

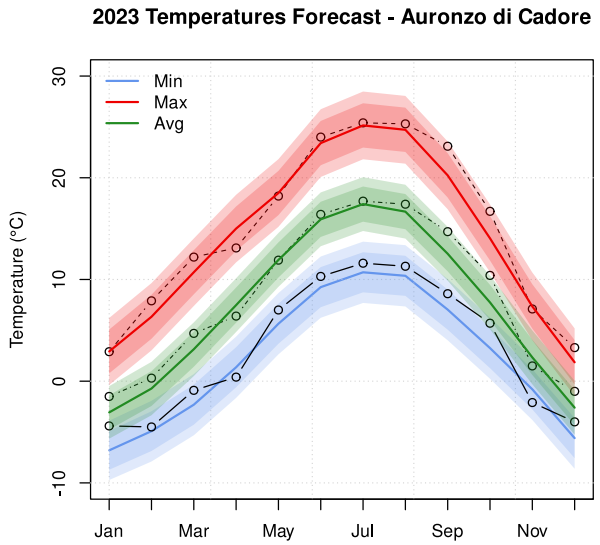
Part 3: Forecasting

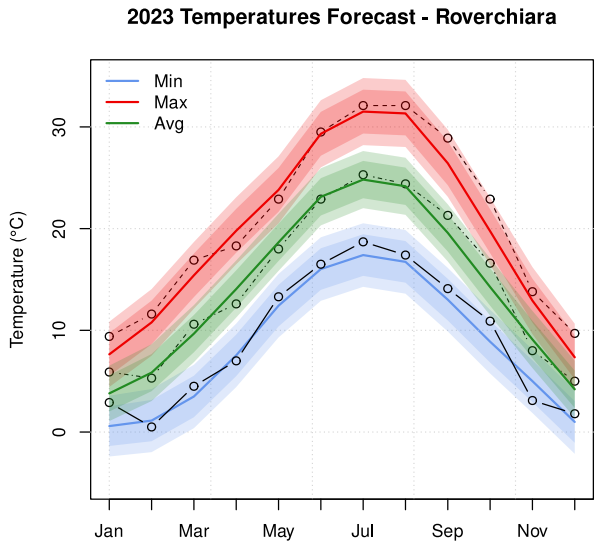
AURONZO DI CADORE

$ARIMA(2,0,0)(0,1,1)_{12}$
AIC: 1247.97
RMSE: $1.48\ °C$

$ARIMA(1,0,1)(1,1,1)_{12}$
AIC: 1333.73
RMSE: $1.49\ °C$

$ARIMA(1,0,1)(1,1,1)_{12}$
AIC: 1169.57
RMSE: $1.38\ °C$





ROVERCHIARA

$ARIMA(2,0,1)(2,1,1)_{12}$
AIC: 1270.70
RMSE: $1.28\ °C$

$ARIMA(2,0,2)(1,1,1)_{12}$
AIC: 1313.01
RMSE: $1.64\ °C$

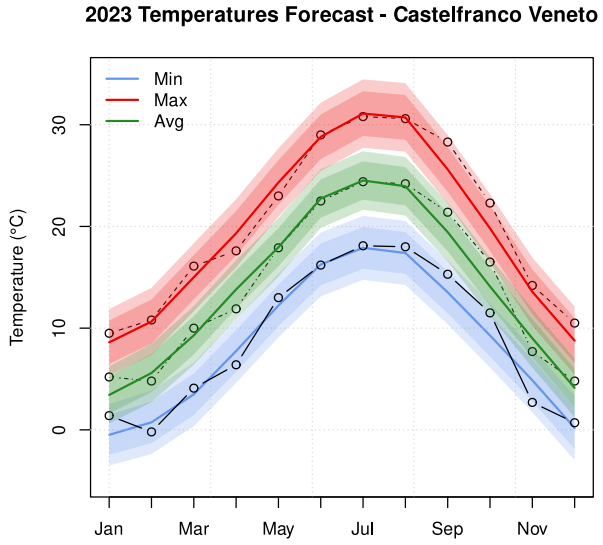
$ARIMA(2,0,2)(2,1,1)_{12}$
AIC: 1208.64
RMSE: $1.25\ °C$

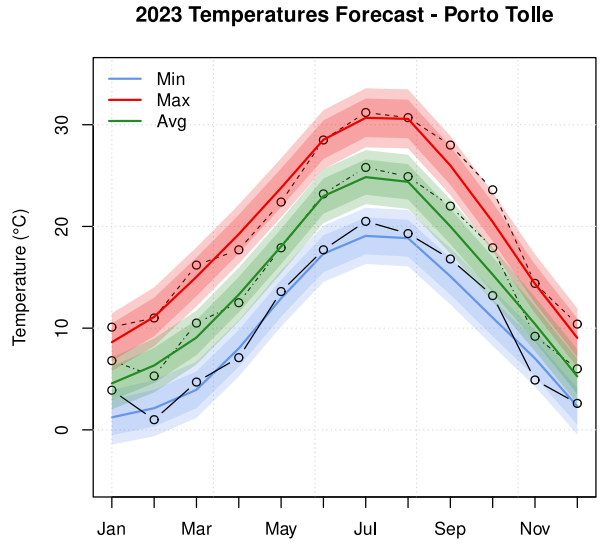
CASTELFRANCO VENETO

$ARIMA(2,0,1)(1,1,1)_{12}$
AIC: 1271.02
RMSE: $1.31\ °C$

$ARIMA(1,0,1)(0,1,1)_{12}$
AIC: 1331.08
RMSE: $1.40\ °C$

$ARIMA(2,0,1)(2,1,1)_{12}$
AIC: 1222.68
RMSE: $1.28\ °C$





PORTO TOLLE

$ARIMA(1,0,0)(0,1,1)_{12}$
AIC: 1184.56
RMSE: $1.44\ °C$

$ARIMA(2,0,1)(1,1,1)_{12}$
AIC: 1237.38
RMSE: $1.41\ °C$

$ARIMA(1,0,0)(0,1,1)_{12}$
AIC: 1151.41
RMSE: $1.38\ °C$

Backup 2: Posterior on a

```
#########  Auronzo , FIT RESULT #########
4 yrs increase trend in min temp.: (0.0273 ± 0.104) °C / 4 yrs
4 yrs increase trend in max temp.: (0.0279 ± 0.0708) °C / 4 yrs
4 yrs increase trend in avg temp.: (0.0221 ± 0.0717) °C / 4 yrs

#########  Castelfranco , FIT RESULT #########
4 yrs increase trend in min temp.: (-0.00935 ± 0.112) °C / 4 yrs
4 yrs increase trend in max temp.: (-0.0503 ± 0.175) °C / 4 yrs
4 yrs increase trend in avg temp.: (-0.0344 ± 0.0411) °C / 4 yrs

#########  Porto Tolle , FIT RESULT #########
4 yrs increase trend in min temp.: (0.0136 ± 0.106) °C / 4 yrs
4 yrs increase trend in max temp.: (0.0304 ± 0.163) °C / 4 yrs
4 yrs increase trend in avg temp.: (0.0151 ± 0.108) °C / 4 yrs

#########  Roverchiara , FIT RESULT #########
4 yrs increase trend in min temp.: (-0.0285 ± 0.113) °C / 4 yrs
4 yrs increase trend in max temp.: (-0.00446 ± 0.113) °C / 4 yrs
4 yrs increase trend in avg temp.: (-0.028 ± 0.0793) °C / 4 yrs
```

What if we had assumed, for example:

$$\Delta T_j \sim Norm(aj + b, \sigma^2)$$

Always compatible with 0

Backup 2: Compatibility between regions

The regional trends b_p and b_v both have a posterior which is not strictly speaking a gaussian distribution. To test whether the two results are compatible, one should use the t-student distribution.

Using independent flat priors, we estimate the variance as:

$$\hat{\sigma}^2 = \frac{\sum_i^4 \left(b_{v,i} - \bar{b}_v\right)^2 + \sum_i^4 (b_{p,i} - \bar{b}_p)\text{^}2}{n_1 + n_2 - 2}$$

And the (1-α)x100% credible interval for b_p - b_v becomes:

$$\bar{b}_p - \bar{b}_v \pm t_{\alpha\,2}\hat{\sigma}\sqrt{1/n\_1 \; + 1/n_2}$$

Making the computations, one finds that 0 is well inside this interval, so compatibility is statistically claimed with a level of significance of 5%