El Nino

A Network Tour of Data Science

A. Bonelli M. Cotting M. Allani

January 2017

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Goals

- Predicting future temperatures in specific zones of the Pacific ocean.
- Trying to find to best model to make predictions.

Outline

Outline

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Data Exploration

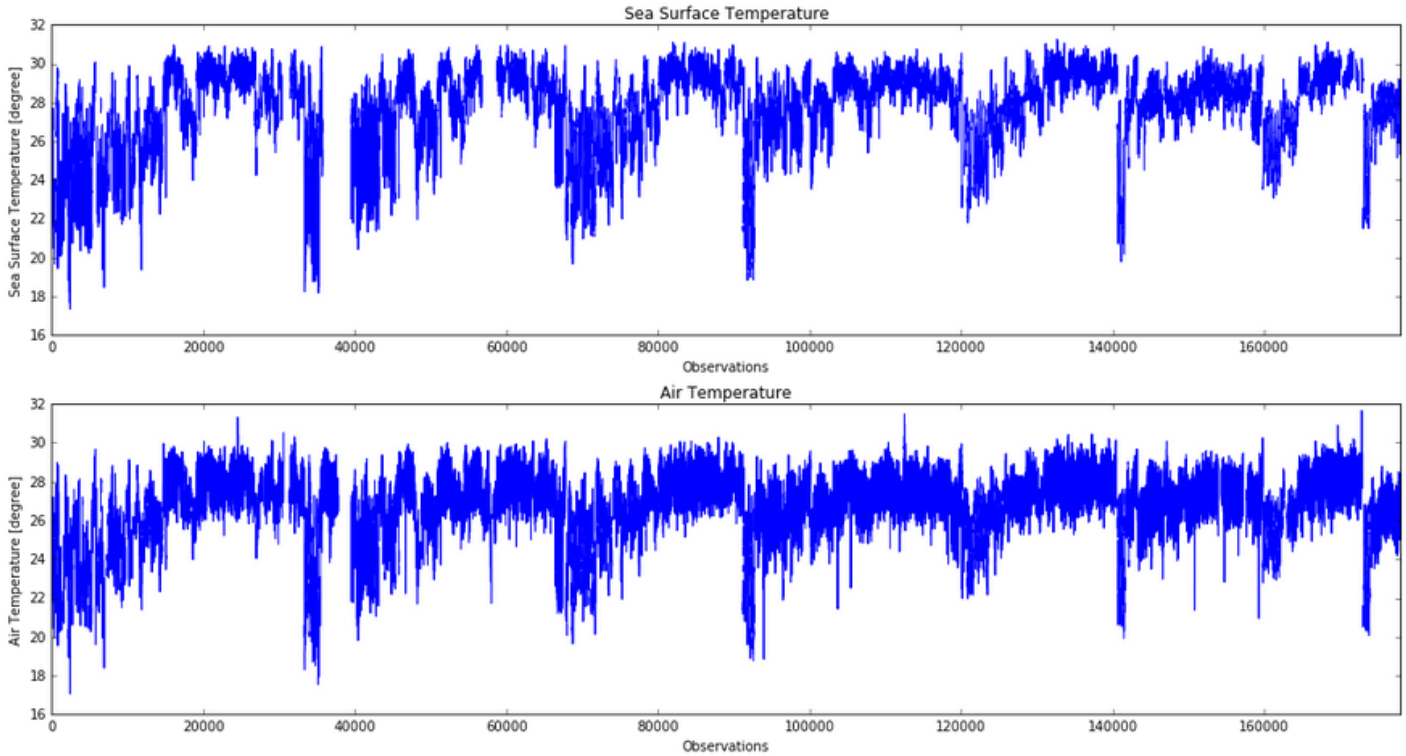


Figure: Exploration of the Air and Sea Surface Temperature

Data Exploration

Base Map

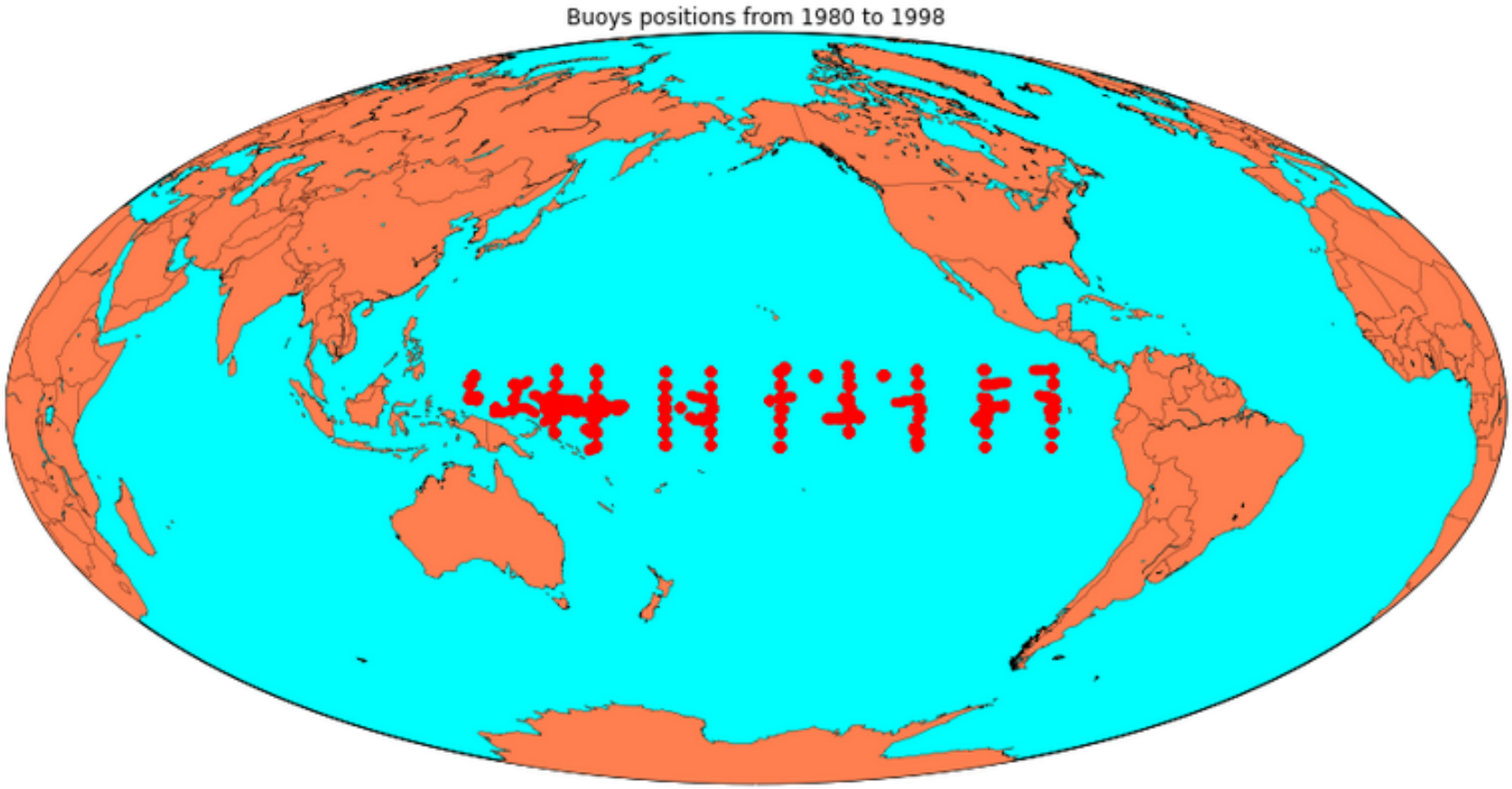


Figure: Positions of Buyos

Data Exploration

Base Map

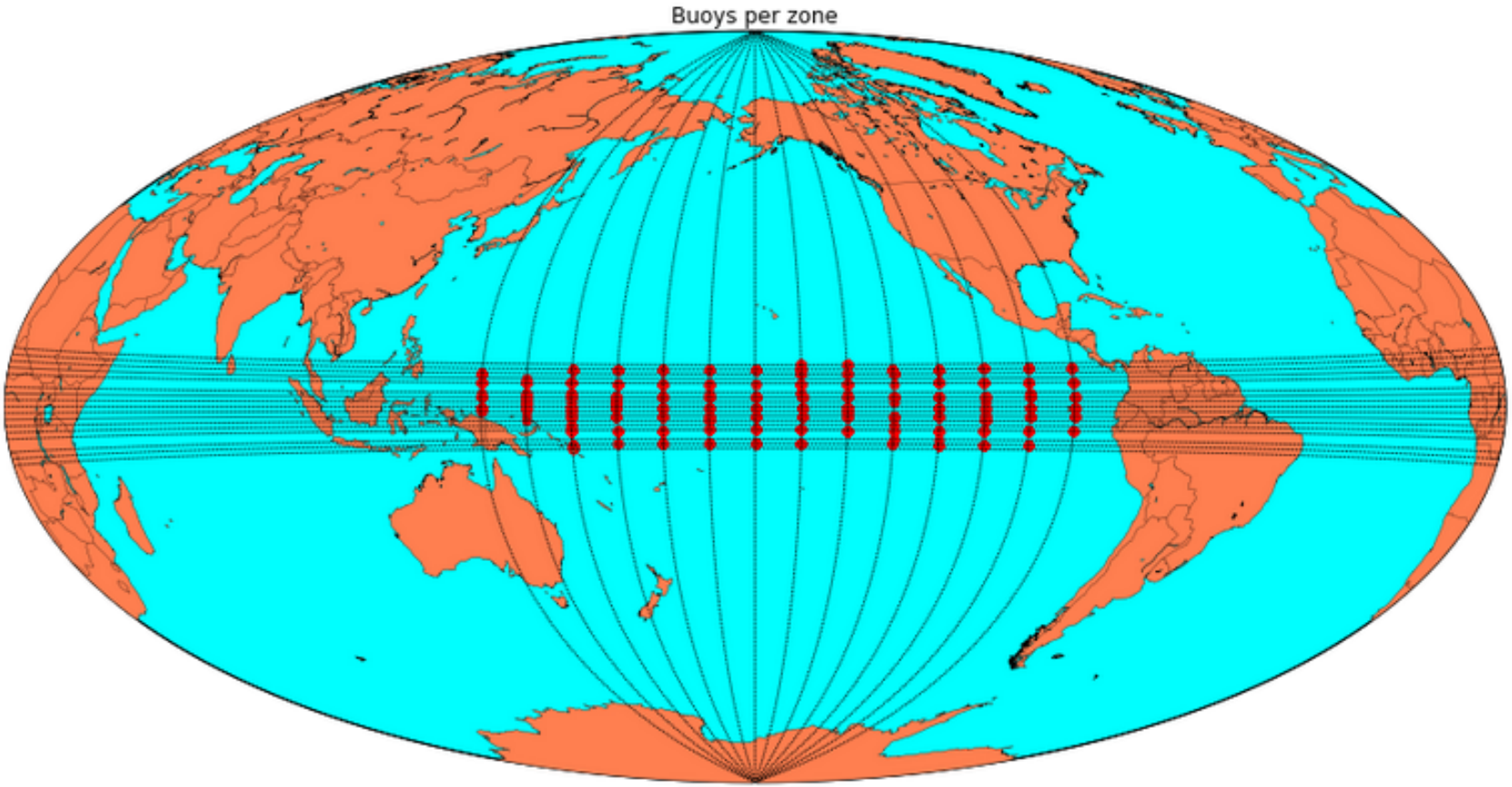


Figure: Positions of Buyos rearranged

Data Exploration

Data evolution

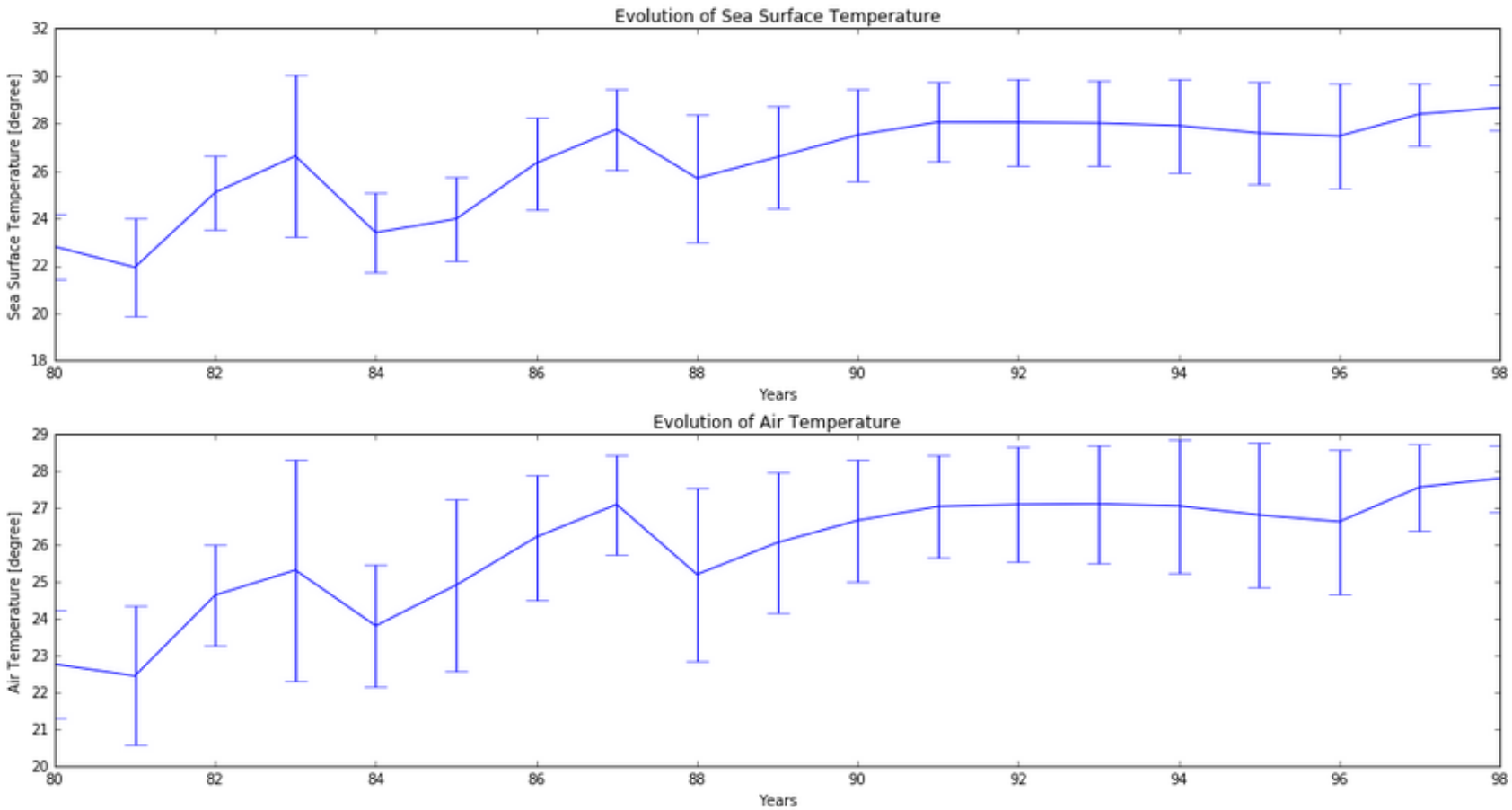


Figure: Evolution of the Air and Sea Surface Temperature

Data Exploration

Data evolution

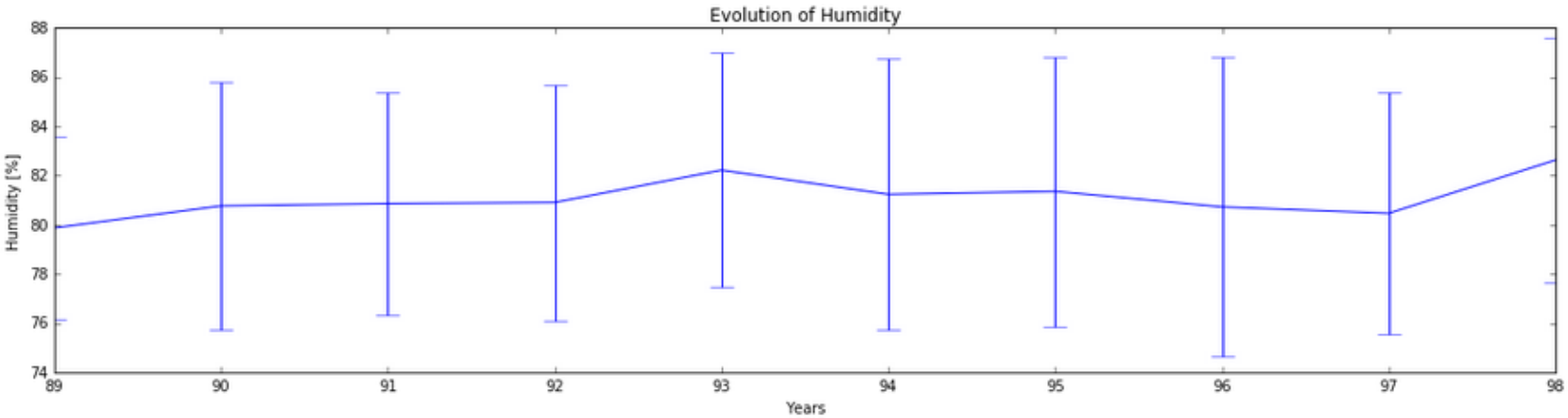


Figure: Evolution of Humidity

Data Exploration

Data evolution

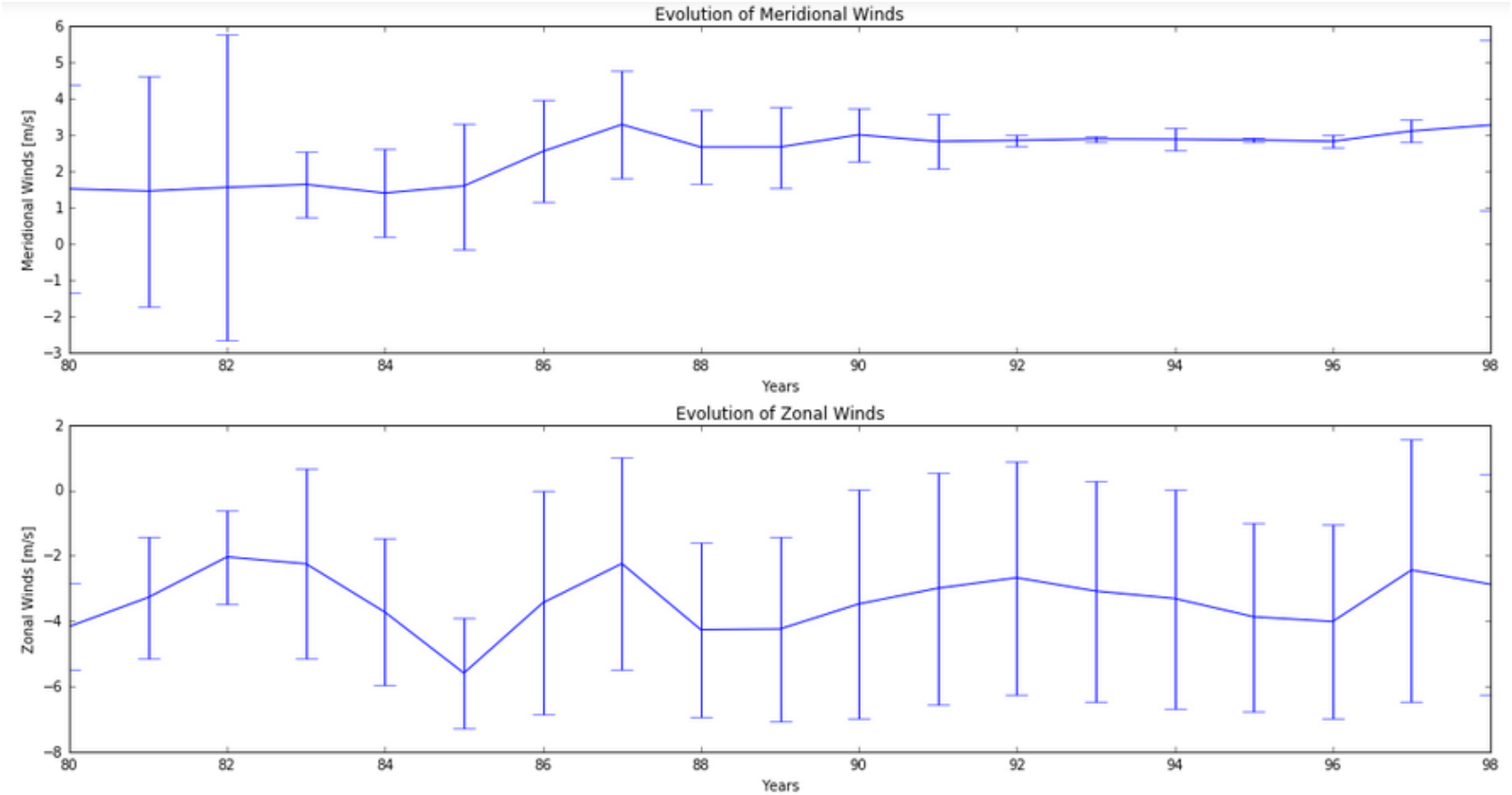


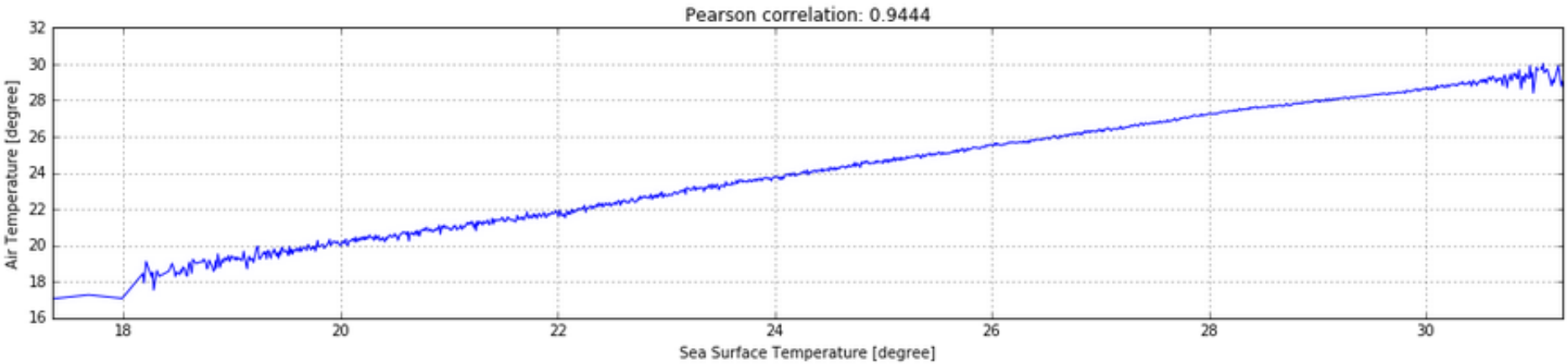
Figure: Evolution of Winds

Outline

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Data Exploration

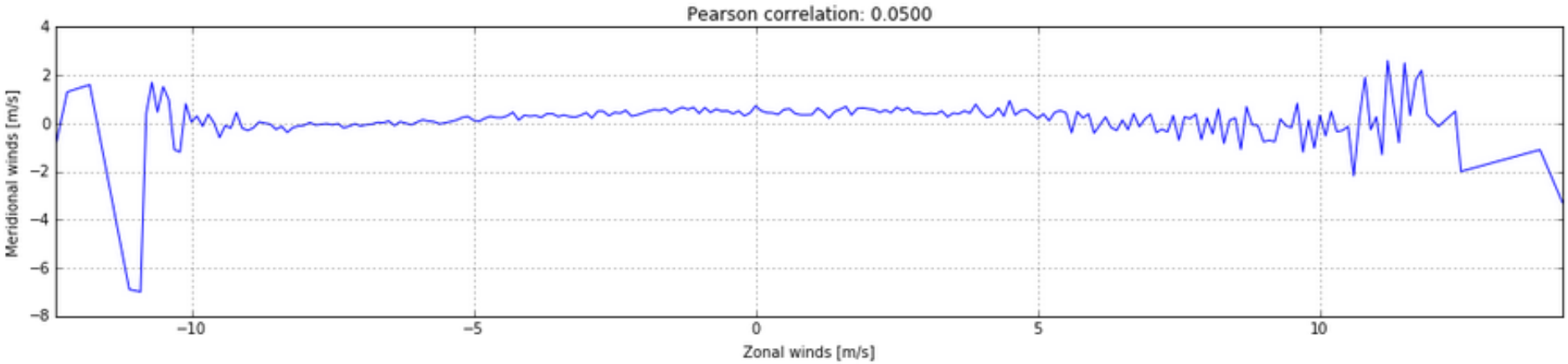
Correlation between Air and Sea Surface Temperature



ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

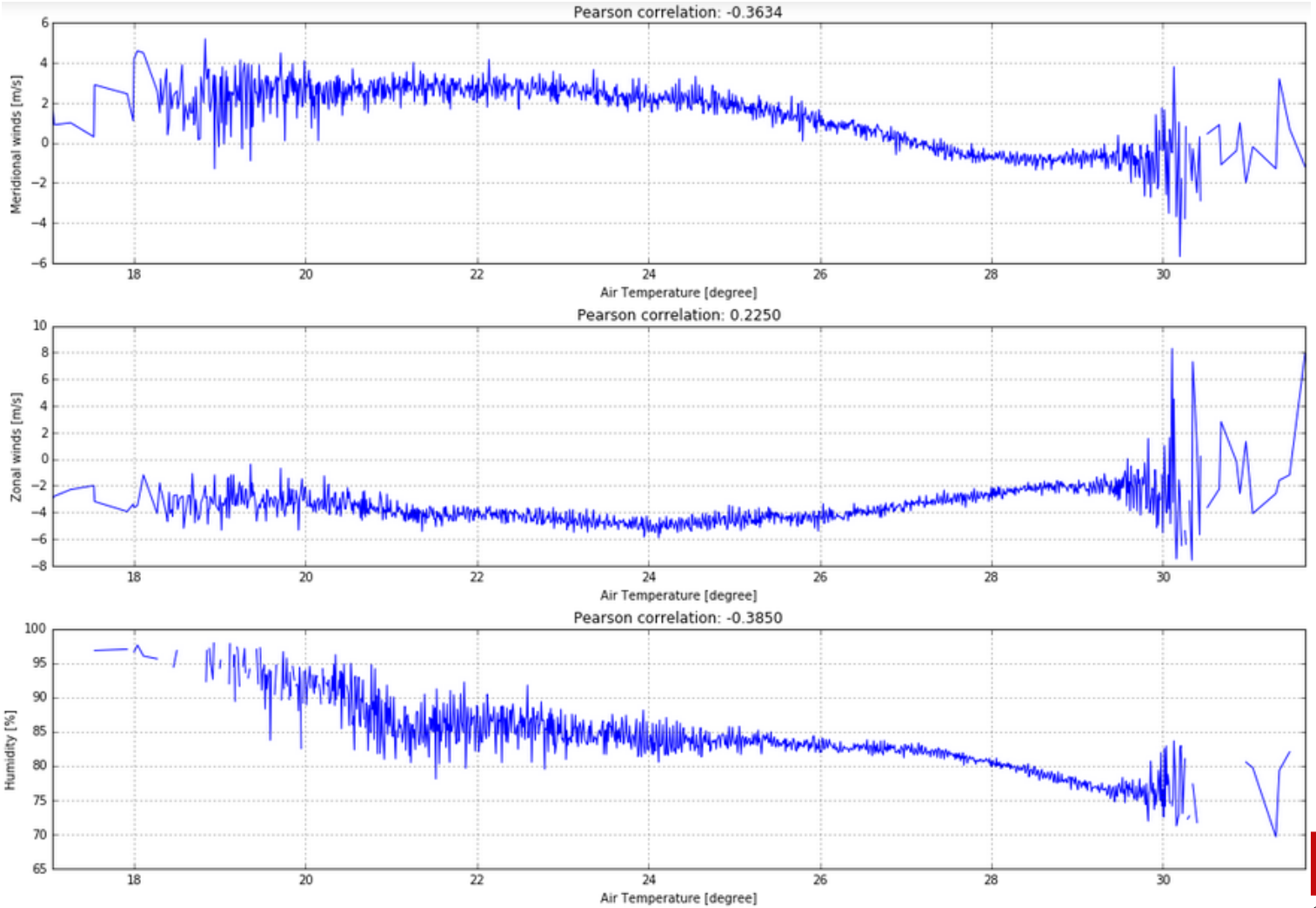
Data Exploration

Correlation between Winds



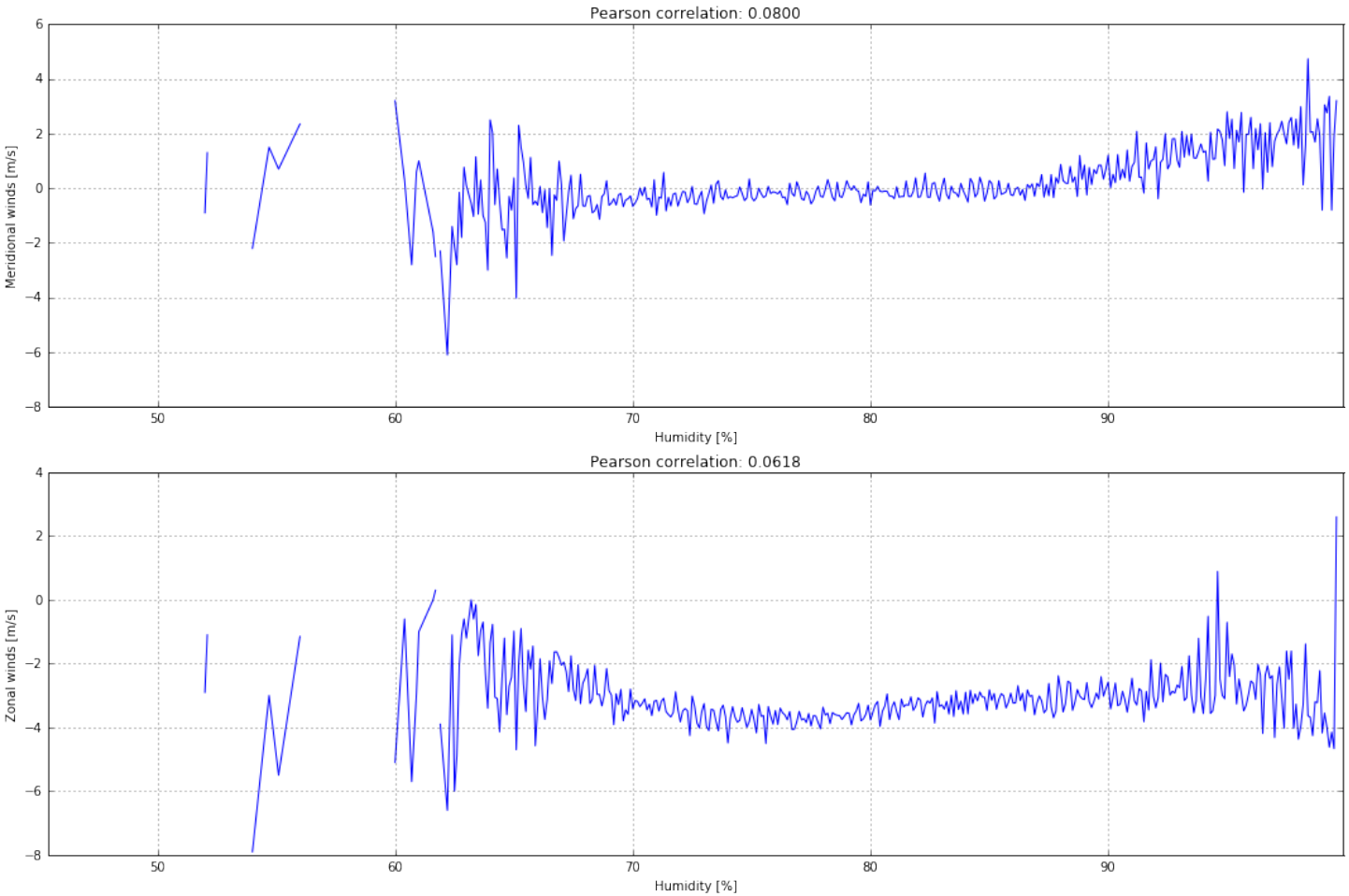
Data Exploration

Correlation between Air Temperature, Winds and Humidity

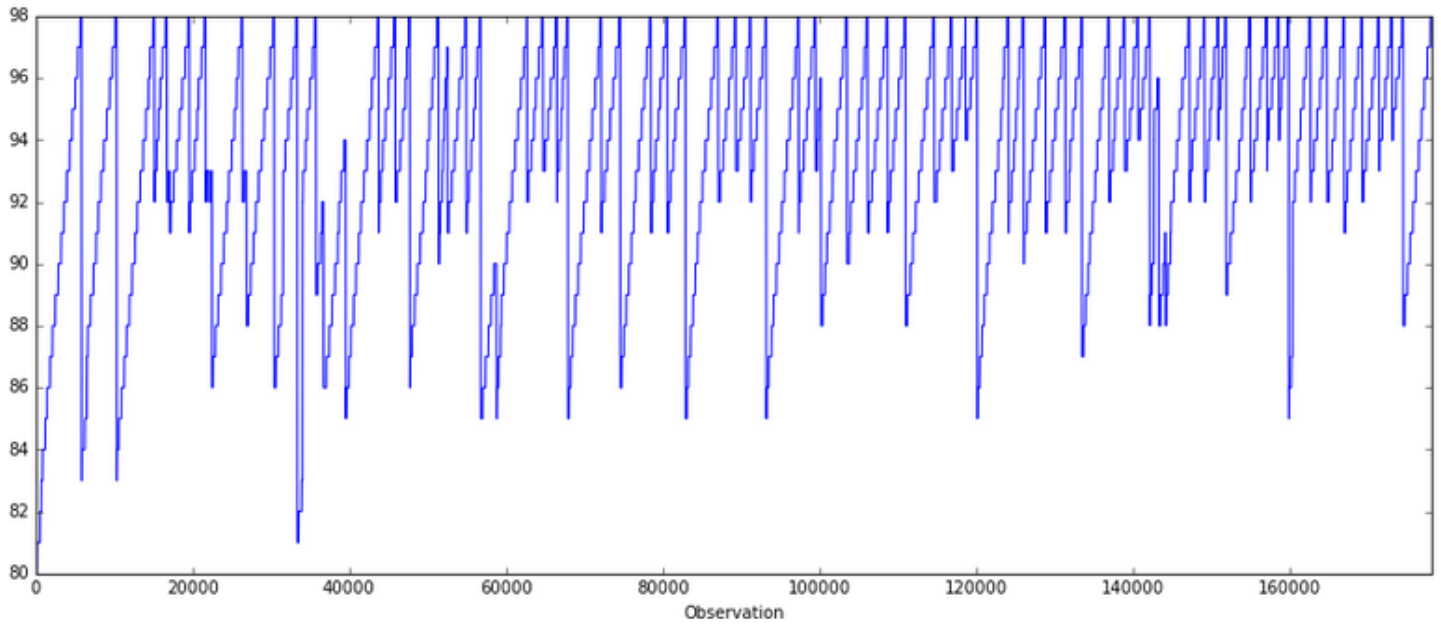


Data Exploration

Correlation between Winds and Humidity



Data Exploration



- Observations are not ordered with time.
- We focus our analysis in one zone in order to be able to make time series predictions.

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Data Exploration

Yearly evolution of meteorological variables for the month of July in a specific zone

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Outline

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

- Monthly air temperature variation from 1980 to 1998
- Selected zone: $(0, -110)$
- Some missing measurements

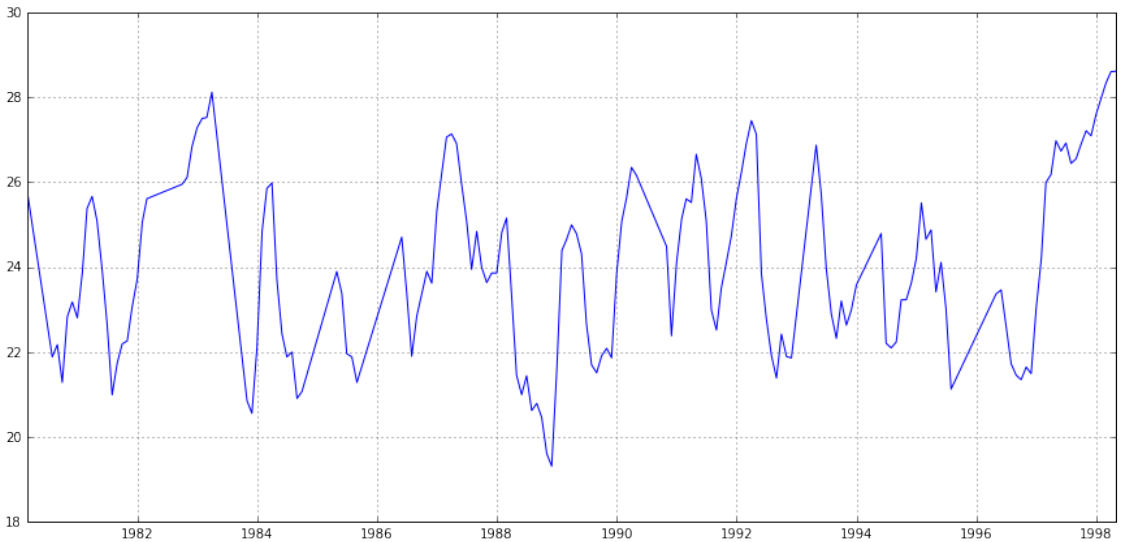


Figure: Air temperature evolution from 1980 to 1998

ARMA

Stationary Check

- Need stationary signal to fit ARMA model
  - Time independent mean
  - Time independent variance
- Perform Dickey-Fuller Test to check stationarity

Results	
Test Statistic	-4.857
p-value	0.000042
Nbre Observations	164
Critical Value (5%)	-2.879
Critical Value (1%)	-3.470
Critical Value (10%)	-2.576

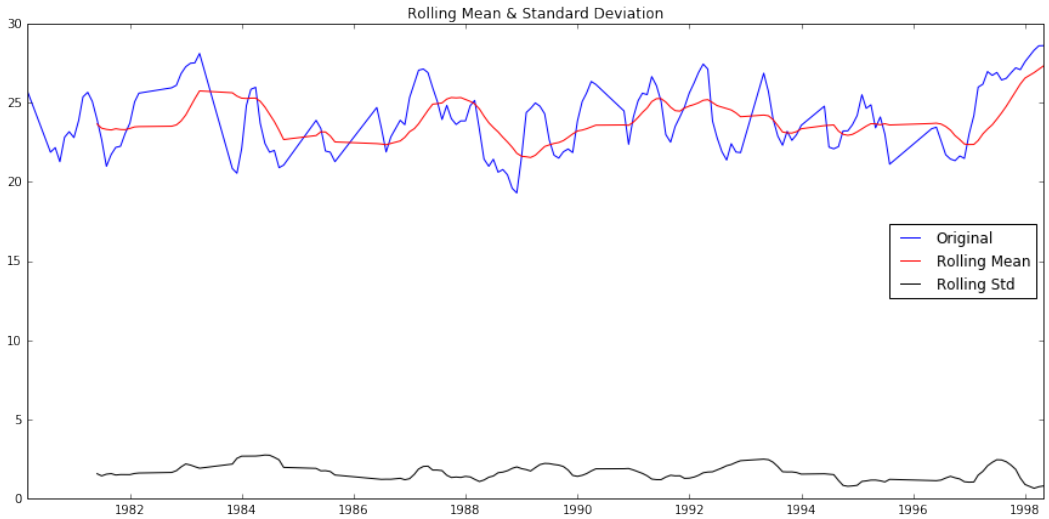


Figure: Rolling mean and standard deviation

- Crital value (1%) $<$ Test statistic
  $\rightarrow$ signal is at 99% stationary

- Compute Autocorrelation (ACF) and Partial Autocorrelation (PACF) function
  $\rightarrow$ determine degree of Auto-Regressive (AR) and Moving Average (MA) model

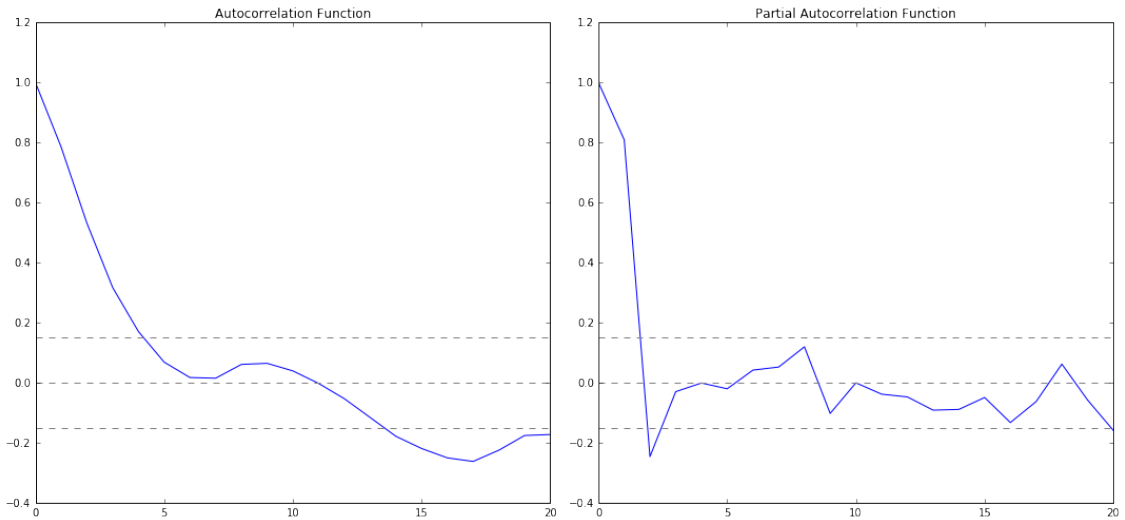


Figure: ACF and PACF

- Min. degrees: AR $\rightarrow$ 3 and MA $\rightarrow$ 4

- Best model ARMA = (5,5)
- MSE = 1.567
- Missing values influence a lot the prediction

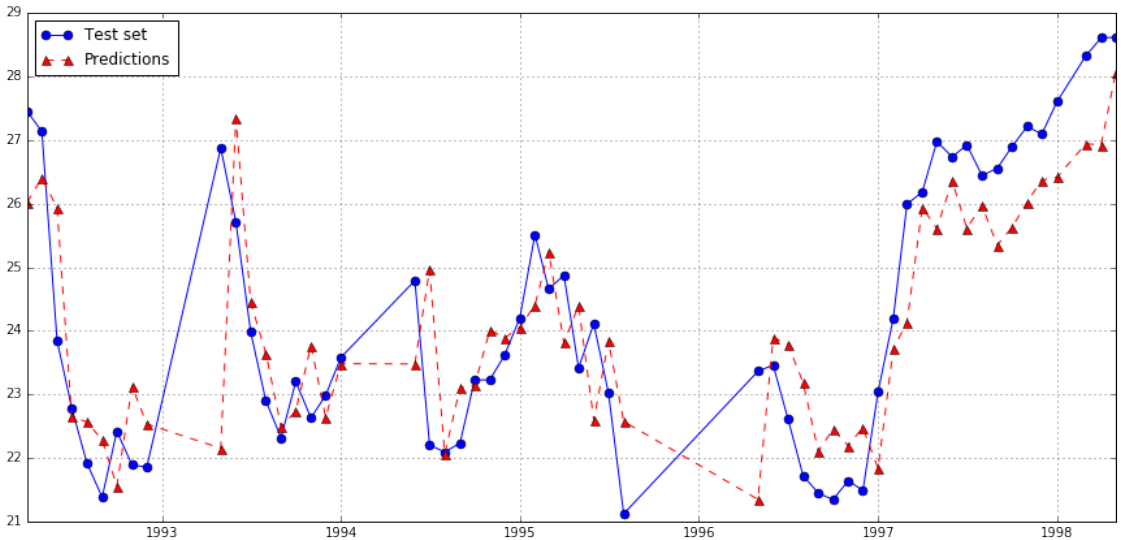


Figure: Predictions

Outline

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

- Several regression methods have been investigated:
  1. Linear
  2. Ridge
  3. LASSO
  4. ElasticNet
  5. Polynomial up to degree 5
- Used metrics:
  1. $R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$

     with $SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$

     $SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$
  2. MSE $= \frac{1}{n} \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2$

- Best results with polynomial regression of degree 2
- Fitting time comparison:

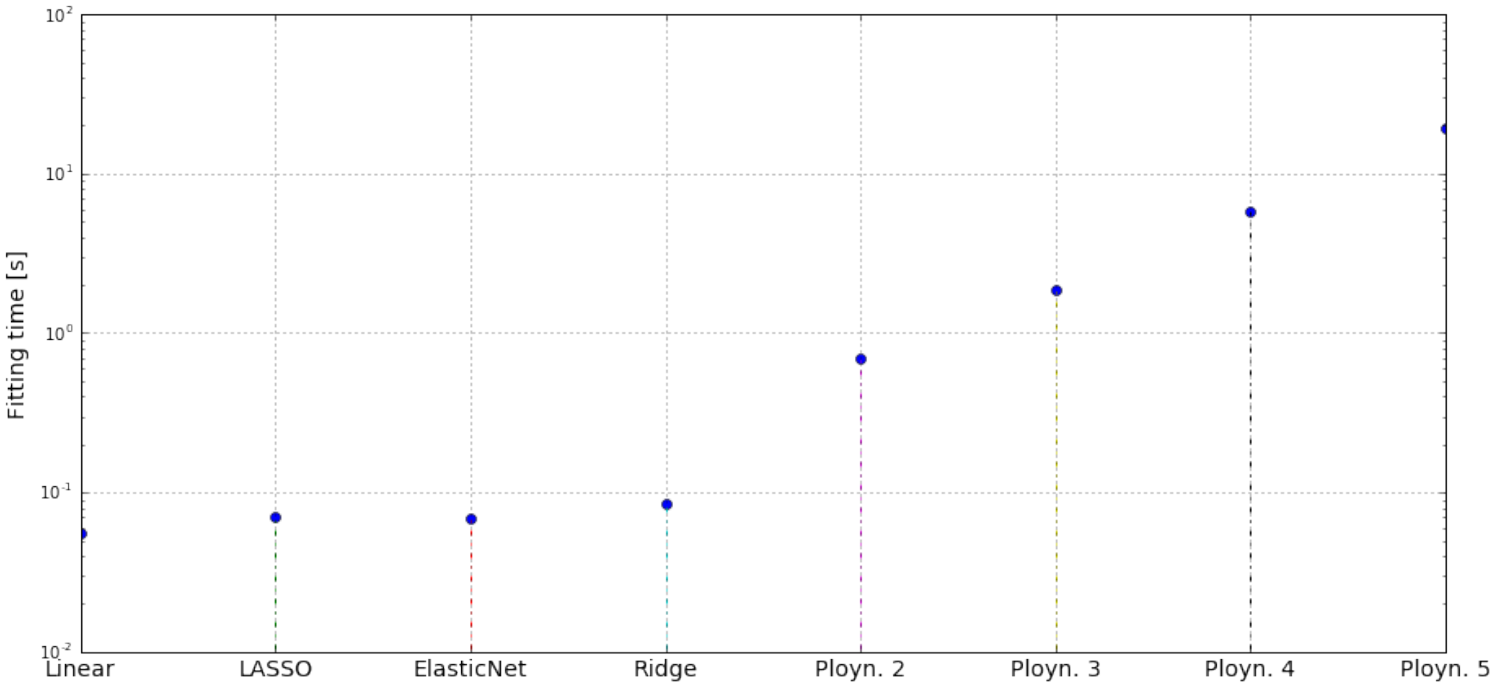
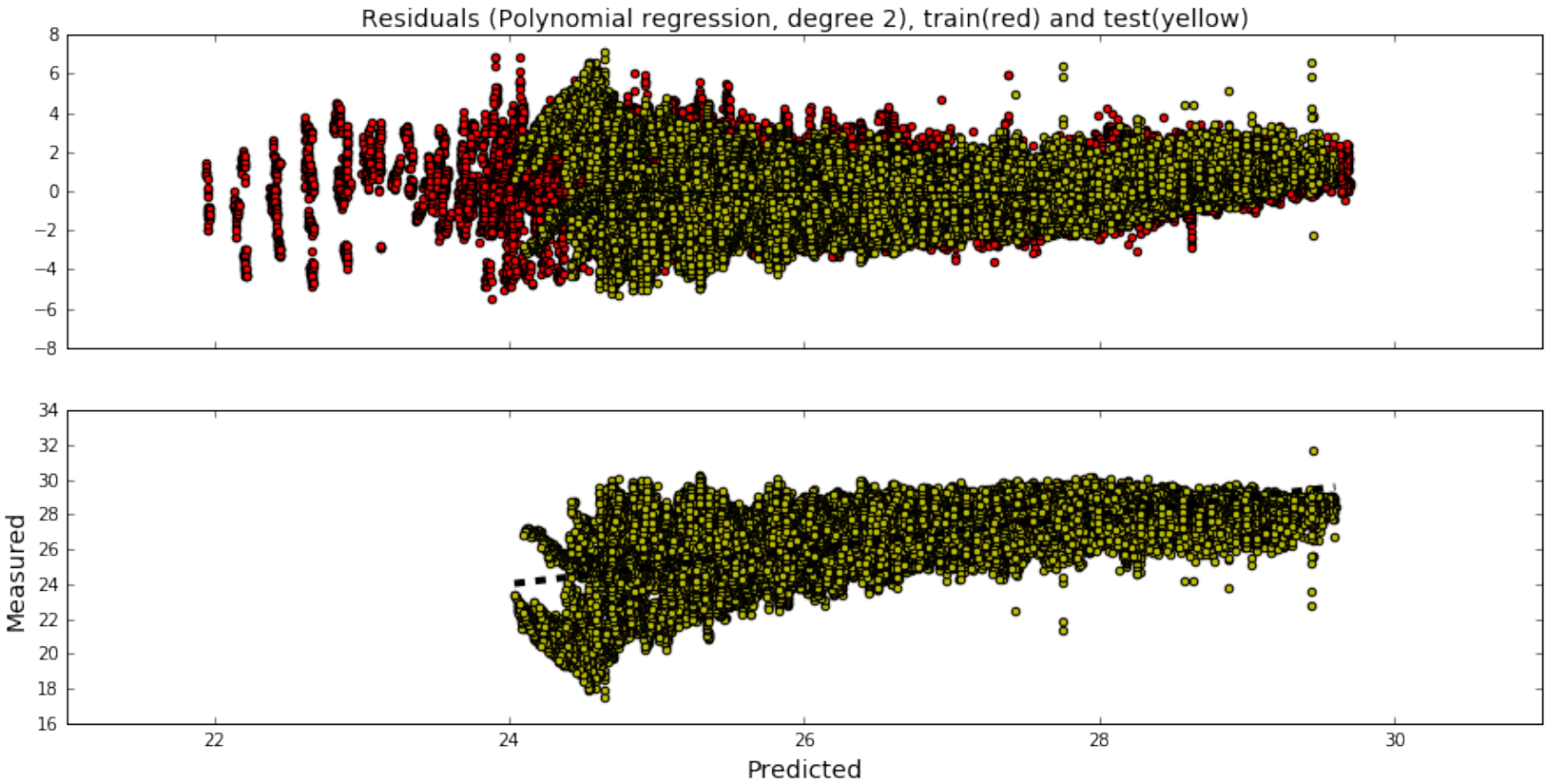


Figure: Fitting time

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Regression

Polynomial Regression



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

- Train on fold 1 $\rightarrow$ Test on fold 2

  Train on fold 1+2 $\rightarrow$ Test on fold 3

  $\ldots$

  Train on fold 1+2+3+4+5+6+7+8+9 $\rightarrow$ Test on fold 10
- Minimization problem for Lasso regression:

$$\min_w \left(\frac{1}{2n_{samples}} ||y - Xw||_2^2 + \alpha ||w||_1 \right)$$

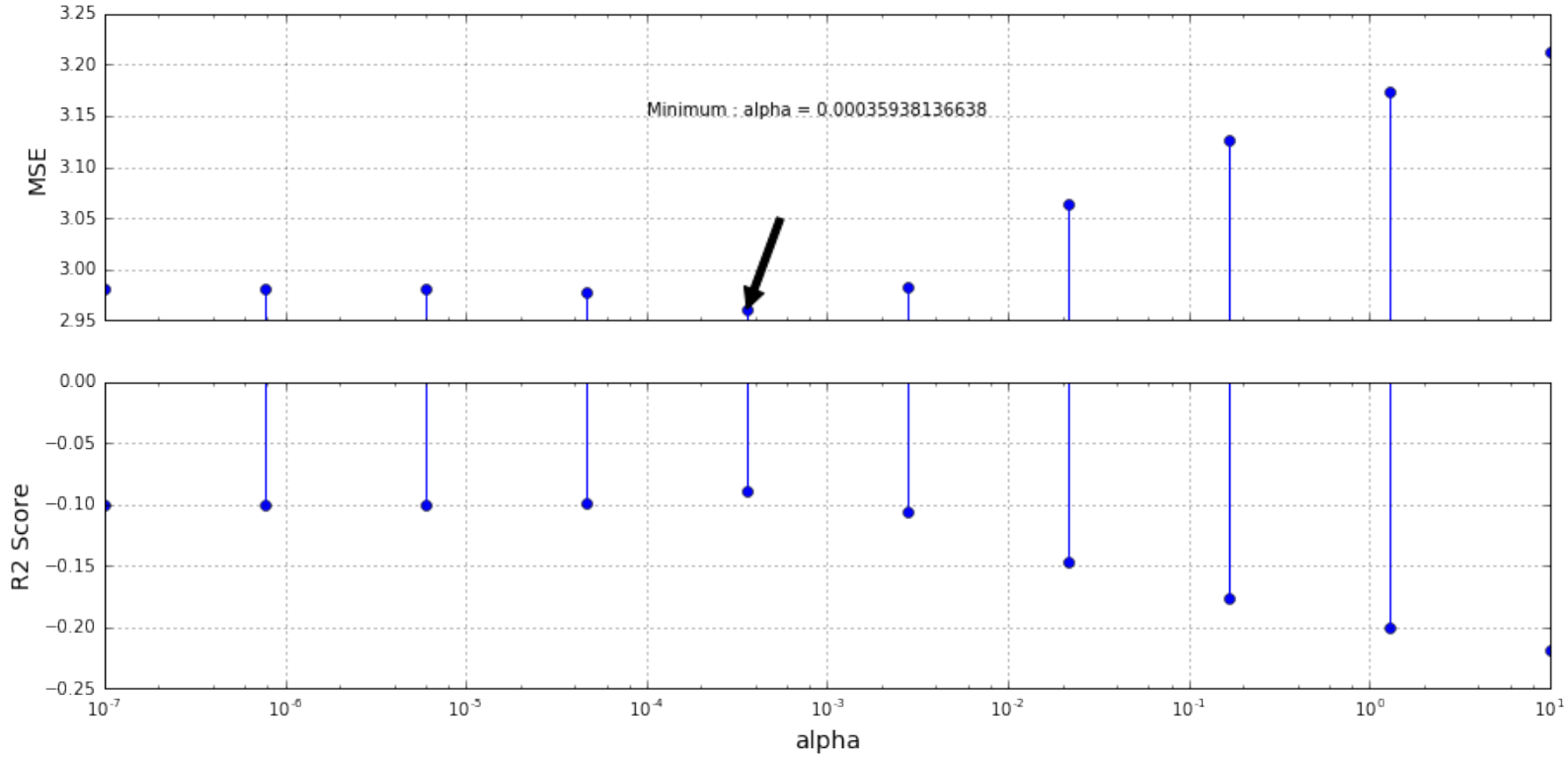


Figure: Optimal α

- Lasso's MSE $>$ polynomial of degree 2's MSE

- ARMA's MSE $<$ polynomial's MSE
  $\rightarrow 1.567 < 6.285$
- Best results: ARMA model

Thank you for your attention !!

"Great things are done by a series of small things brought together."

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE