Modeling of Earth System Data

Jacobs University Bremen, Course CA-EES-803, Fall 2022

— Project description —

2 Parametric Modeling of Earth System Processes

The topics addressed in the second course chapter *Parametric Modeling of Earth System Processes* include the linear algebra perspective on basic statistics and data analysis, the correspondence of linear correlation and alignment of data vectors, the orthogonality principle in least squares modeling, the least squares normal equations and singular value decomposition. In a series of Jupyter notebooks, the methods are presented and illustrated using synthetic data and geophysical time series.

In this project, key concepts presented in the course chapter *Parametric Modeling of Earth System Processes* are to be demonstrated using a combined polynomial and harmonic regression model applied to climate time series showing periodicities, namely, time series of the Quasi-Biennial Oscillation (QBO) provided by the Institute of Meteorology at the FU Berin as well as monthly mean values of globally averaged CO_2 concentrations provided by the NOAA Global Monitoring Laboratory.

2.1 Preparation

Consult the presentation file of the course chapter *Parametric Modeling of Earth System Processes* to review the theoretical basis of linear parametric modeling. Go through the computational exercises to recall the numerical methodology and the syntax of relevant Python functions. Of particular importance is the Numerical Software Lab (Jupyter notebook) associated with the section *Data modeling and numerical linear algebra*, where a harmonic regression model is applied to the QBO data provided by the Institute of Meteorology at the FU Berlin through their website *The Quasi-Biennial Oscillation (QBO) Data Series*:

- `https://www.geo.fu-berlin.de/met/ag/strat/produkte/qbo/qbo.dat`.
- `https://www.geo.fu-berlin.de/en/met/ag/strat/produkte/qbo/index.html`

Familiarize with the scientific context and the data format.

Visit the web archive on *Carbon Cycle Greenhouse Gases* hosted by the *Global Monitoring Laboratory (GML)* at the US National Oceanic and Atmospheric Administration (NOAA):

- `https://gml.noaa.gov/ccgg/`
- `https://gml.noaa.gov/`

Globally averaged marine surface monthly mean data of CO_2 concentrations are available from the web page *Trends in Atmospheric Carbon Dioxide* maintained by *Ed Dlugokencky and Pieter Tans, NOAA/GML*:

- `https://gml.noaa.gov/webdata/ccgg/trends/co2/co2_mm_gl.txt`
- `https://gml.noaa.gov/ccgg/trends/gl_data.html`
- `https://gml.noaa.gov/ccgg/trends/`

On the course teamwork space you find the Jupyter notebook `med-prj2-param-jnb.ipynb` with further instructions, code templates, and scripting examples.

2.2 Implementation and test of the model function

The function used to model the data $d = d(t)$ combines a polynomial fit with harmonic regression:

$$m(t) \;=\; \sum_{j=0}^{D} a_j t^j \;+\; \sum_{k=1}^{L} \{a_{2k+1} \cos(2\pi k f t) \;+\; a_{2k+2} \sin(2\pi k f t)\} \;\; .$$

Here D denotes the degree of the polynomial, and L is the number of harmonics considered in the function.

The model is implemented in the function `FitPolyharm()`. Check the code to understand its logic and syntax. In addition to the model parameter vector $\mathbf{a} = (a_0, a_1, \ldots)^\mathsf{T}$, the function returns the prediction $m(t)$ and the root-mean-square (RMS) misfit

$$\text{RMSmf} \;=\; \sqrt{\langle [d(t) - m(t)]^2 \rangle_t}$$

where $\langle \cdots \rangle_t$ denotes (time) averaging. Error weighting is disregarded for the purpose of the current project. The RMS misfit calculated this way suggests that the model captures the essential process(es) generating the data if its numerical value is comparable with the measurement accuracy (standard error of a single datum).

Test the function `FitPolyHarm()` using the synthetic data example from the Jupyter notebook *Data modeling and numerical linear algebra*, and also by means of synthetic data produced from a model function that combines a polynomial of degree $D = 1$ and a harmonic function involving $L = 2$ cosine and sine terms.

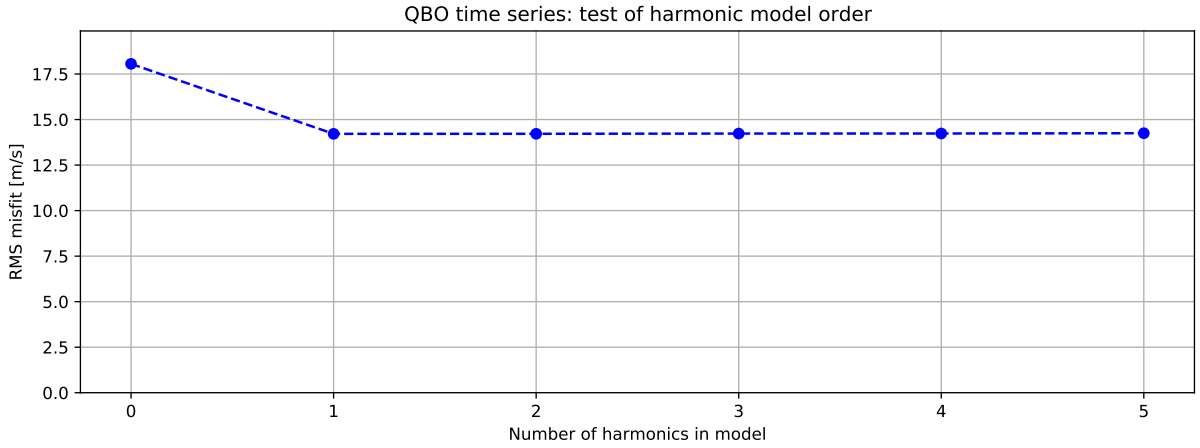

Figure 1: RMS misfit of the QBO time series versus model order L as fitted by `FitPolyHarm()` to a model function consisting of a constant and L harmonics. The curve starts to flatten at $L = 1$ which means that the mean square deviation is not reduced further by considering more than just the first harmonic.

2.3 QBO time series

Load the QBO time series as in the Jupyter notebook *Data modeling and numerical linear algebra*. Set the QBO period to 28.2 months and call `FitPolyHarm()` with $L = 1, 2, 3, 4, 5$ harmonics to obtain fits for different model orders (numbers of harmonics) L. Plot and discuss the RMS misfit versus L. See Figure 1.

Compute the least squares model parameters and compute the coefficient of linear correlation between the data and the best fit. Display the data together with the model. Add a second panel with a plot of the residuals. Discuss how well the model captures the dynamics of the process(es) generating the QBO time series. See Figure 2.

2.4 Globally averaged marine CO_2 concentrations

This part of the data modeling project is concerned with *globally averaged marine surface monthly mean data* of CO_2 concentrations from the web page *Trends in Atmospheric Carbon Dioxide* maintained by *Ed Dlugokencky and Pieter Tans, NOAA/GML*, see `https://gml.noaa.gov/ccgg/trends/`. Load the CO_2 concentrations from the file `co2_mm_gl.txt` and follow the instructions in the Jupyter notebook `med-prj2-param-jnb.ipynb`.

CO_2 concentrations are affected by the annual vegetation cycle, hence we set the fundamental period of the polynomial-harmonic model to one year. Plot the RMS misfit versus the number of harmonics L. At which value of the model order L do you observe the "knee" of the RMS misfit curve, i.e., where does the curve begin to flatten? Compare the RMS misfit with the measurement error specified in the data file. See Figure 4.

Compute the least squares model parameters and the coefficient of linear correlation between the data and the best fit. Identify the CO_2 concentration growth rate a_1 in the reference year,

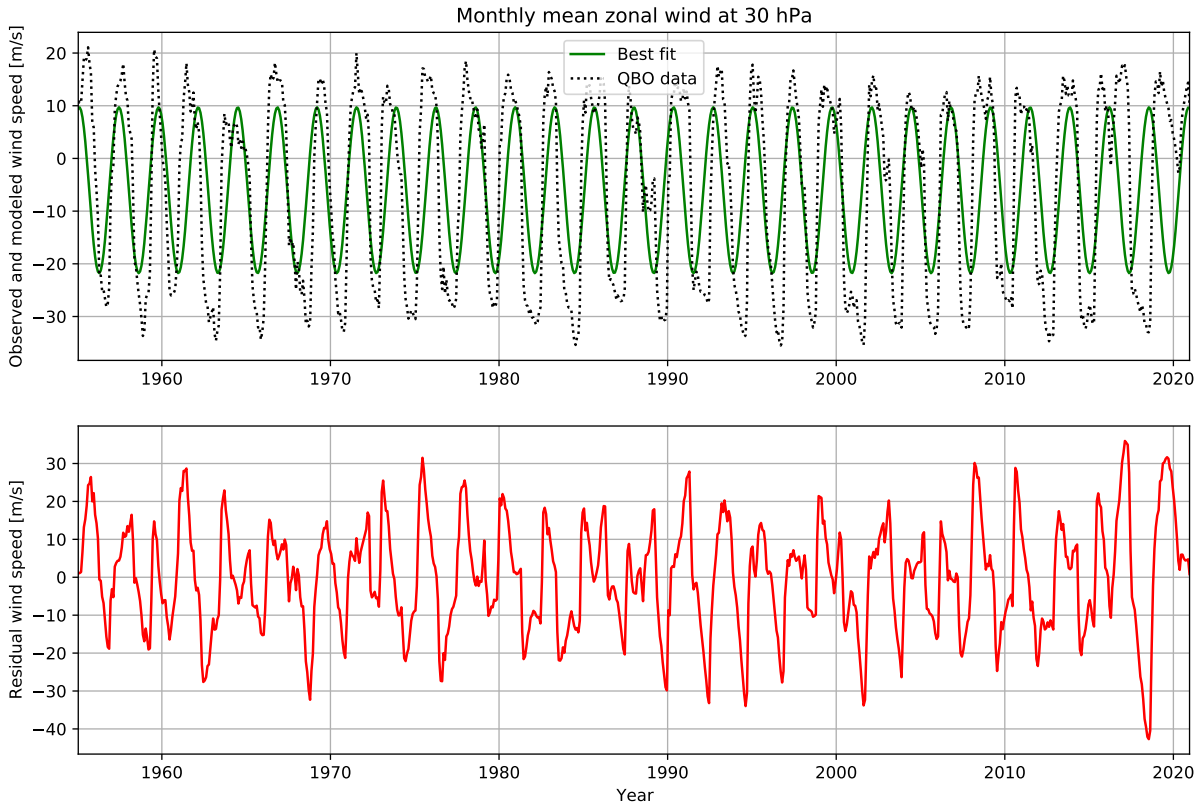

Figure 2: Upper panel: QBO time series provided by the Institute of Meteorology at the FU Berlin together with the best fit to a model function consisting of a constant and one harmonic. Lower panel: Residual of the best fit.

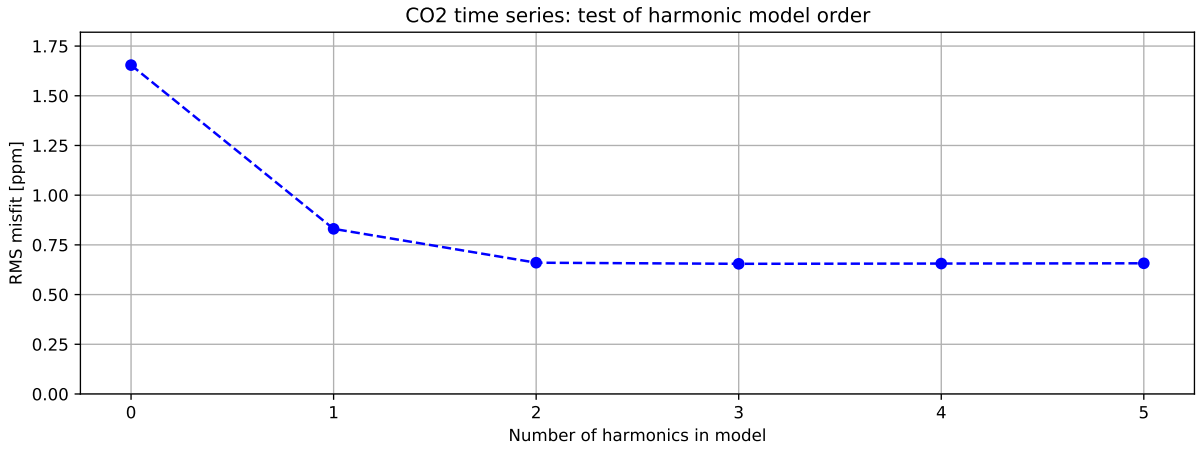

Figure 3: RMS misfit of the globally averaged CO_2 concentration time series provided by NOAA/GML versus model order L as fitted by `FitPolyHarm()` to a model function consisting of a quadratic and L harmonics.

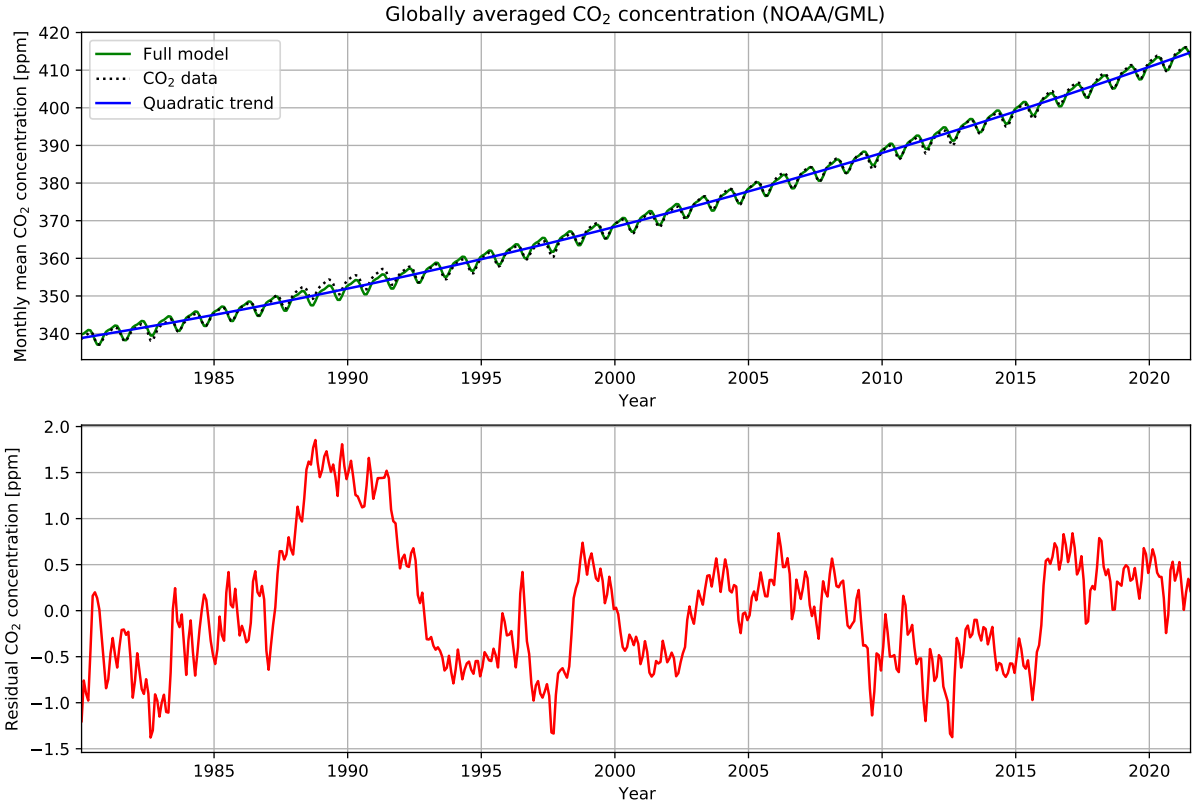

Figure 4: Upper panel: Globally averaged CO_2 concentrations provided by NOAA/GML together with the best fit to a model function consisting of a quadratic and two harmonics. Lower panel: Residual of the best fit.

and the acceleration parameter a_2. Display the data together with the model. Plot also the deseasonalized model, i.e., the polynomial variation only. In a second panel display the residuals. Perform the Shapiro-Wilk normality test and print the probability value. See Figure 4.

Discuss your findings. Which aspects of the CO_2 concentration measurements are captured by the combined polynomial-harmonic model? Which processes are missing?

2.5 Projections of CO_2 concentrations

Using the polynomial (quadratic trend) contribution to the CO_2 concentration model, the dynamics is to be projected until the mid of the 21st century. The function `QTrendCoeffBS()` applies the bootstrap approach (random resampling with replacement) to generate ensembles of quadratic trend coefficients a_0, a_1, a_2.

2.5.1 Tests of the projection procedure

To demonstrate the projection procedure, generate ensembles of models from bootstrap replications of (training) data from the ten-year period 2000–2010. The projections are to be compared

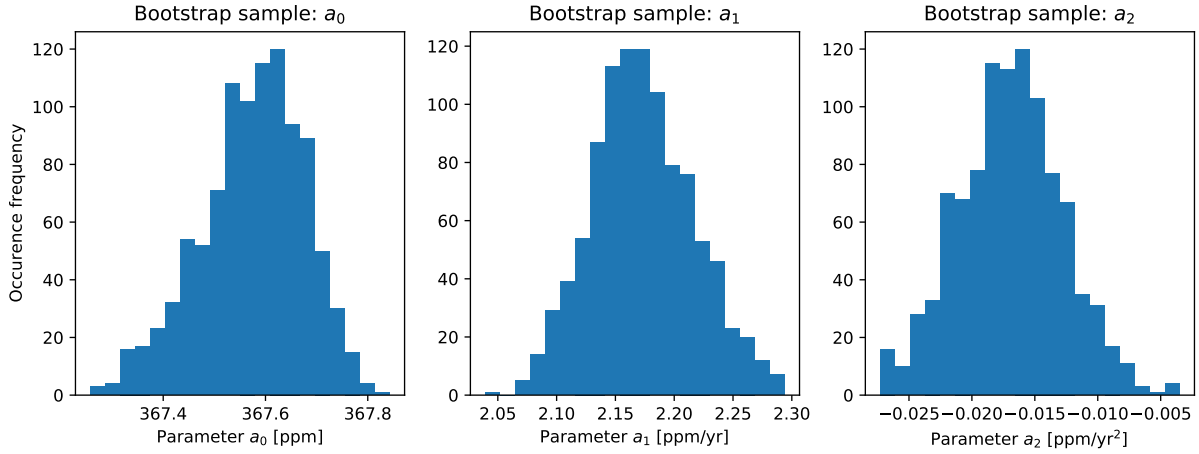

Figure 5: Histograms of quadratic trend parameter bootstrap ensembles obtained from fitting NOAA/GML CO_2 concentration to a parametric model function consisting of a quadratic and two harmonics for the time period 2000–2010.

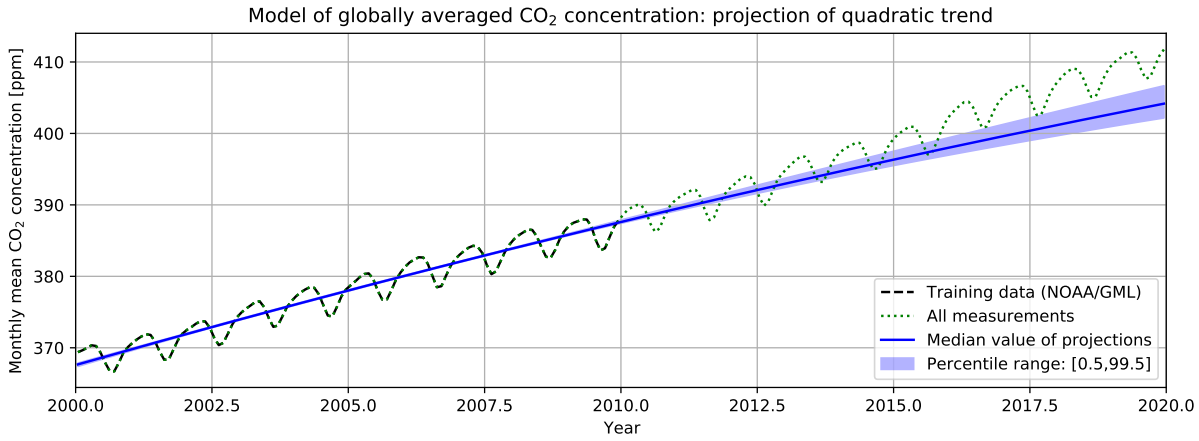

Figure 6: Projection of the quadratic trend obtained from fitting NOAA/GML CO_2 concentration to a parametric model function consisting of a quadratic and two harmonics for the time period 2000–2010. The percentiles were obtained from a bootstrap ensemble of quadratic trend parameters.

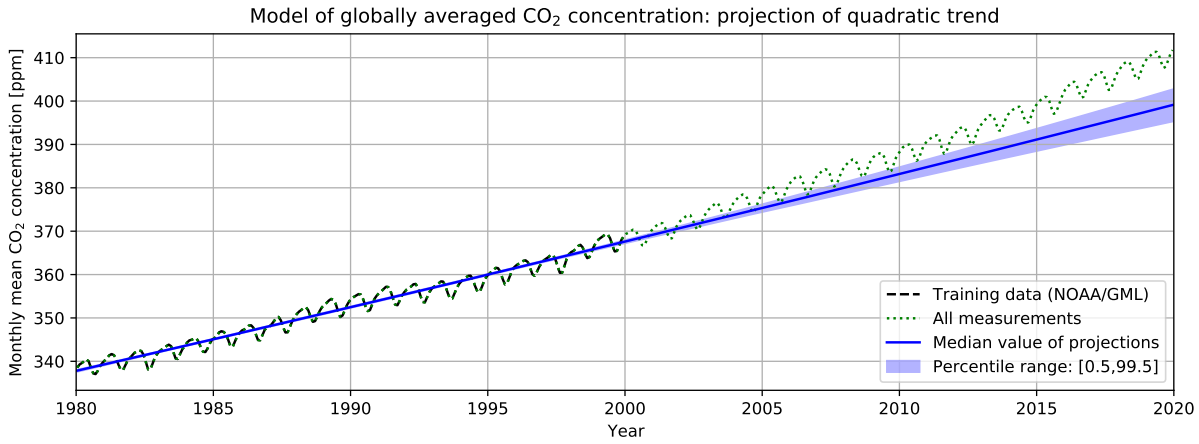

Figure 7: Projection of the quadratic trend obtained from fitting NOAA/GML CO_2 concentration to a parametric model function consisting of a quadratic and two harmonics for the time period 1980–2000. The percentiles were obtained from a bootstrap ensemble of quadratic trend parameters.

with (test) data from the subsequent 10-year period 2010–2020. Display the distributions of the polynomial parameters as in Figure 5.

The bootstrap ensembles of quadratic trend coefficients give rise to an ensemble of quadratic trend projections for the time interval 2000–2020, encompassing both the training period 2000-2010 and the test period 2010–2020. From the quadratic trend projections extract the percentiles at 0.5% and 99.5%, and also the median (50% percentile). Display the percentiles as functions of time together with the data.

For these particular 10-year and 20-year periods, the projections agree reasonably well with the measurements as shown in Figure 6. To see that projections can be much worse, repeat the modeling exercise with other choices of the reference year (1980,1985,1990,1995). As a further test of the projection procedure you may model the quadratic trend from the 20-year period 1980–2000 and then compare with all data from the 40-year period 1980–2020, see Figure 7.

2.5.2 Projection until 2060

Using data from the 40-year time period 1980-2020, proceed as before to obtain projections for another 40 years until 2060. Start with constructing bootstrap ensembles of quadratic trend coefficients. Display the boostrap distributions of the quadratic model parameters. From the bootstrap samples of quadratic trend coefficients construct the quadratic trend projections. Extract the percentiles and plot them as functions of time together with the training data.

Discuss the results in the light of the numerical experiments carried out before. Assuming that the processes not captured by the model are marine system variations on sub-decadal timescales, would you put more or less trust in this long-term projection?

2.6 Project report and digital supplements

Scientific reports typically show the following structural elements.

- *Introduction*: background, motivation, objectives.
- *Methodology*: briefly describe key theoretical concepts and computational tools.
- *Results*: comments on implementation and procedure, description of graphics.
- *Discussion*: discuss the results in the light of the objectives.
- *Conclusions*: suggestions for further work, outlook.
- *References*

Digital supplements (Jupyter notebook, graphics, data files) are to be uploaded to your personal folder on the course teamwork space. Organize the teamwork wiki and add comments.