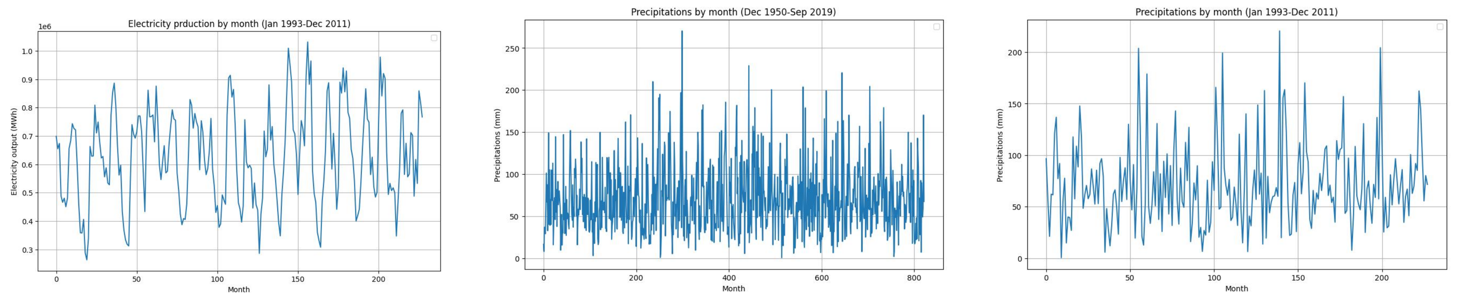
Introduction and exploratory data analysis

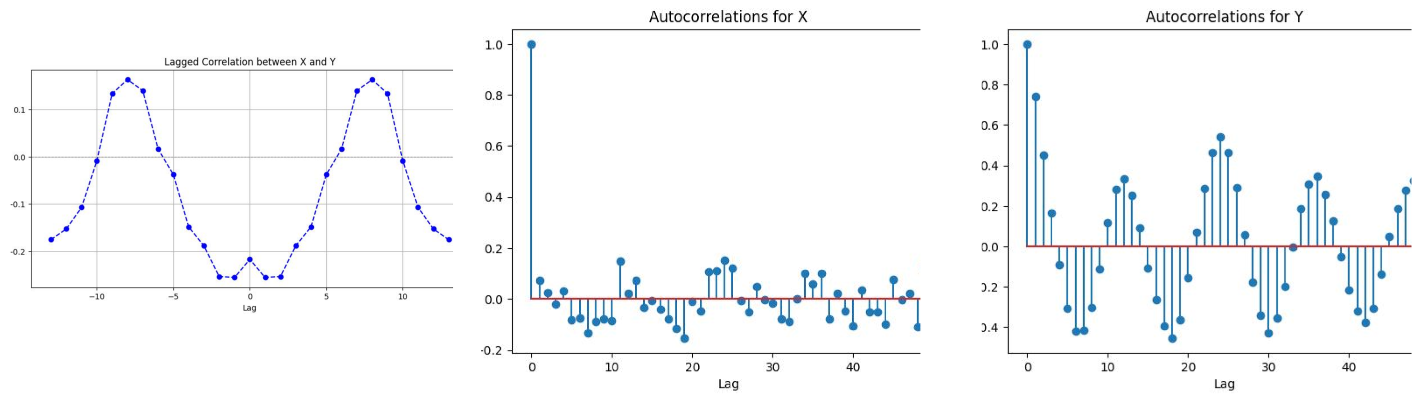
Our task consists of building a model to forecast electricity production in Trondheim using precipitation data as an explanatory variable.
I will call Xt the precipitation data (exogenous variable) and Yt the energy production (dependent variable).

Both quantities are plotted with on a monthly basis but the data for Xt spans a longer time window. (January 1993 to December 2011, 228 values VS December 1950 to September 2019, 822 values)
Because of the exogenous assumption I decided to model precipitations independently, therefore employing the whole data available.



Plotting the lagged correlations between the precipitations and the energy production we see there seems to be a noticeable positive correlation around the 8 months lag mark. On the shorter lags the quantities are negatively correlated.

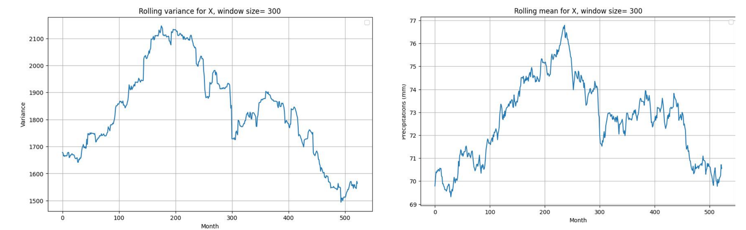
Looking at the auto-correlation functions for both X and Y we see that they exhibit seasonal behavior on annual scale.
Moreover, the energy production data presents higher regularity, likely due to its relationship to energy demand.



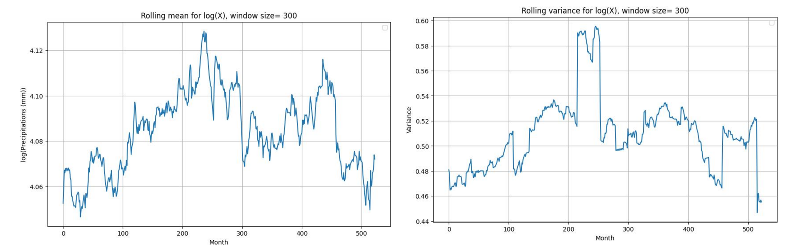
Stationarity and Transformations

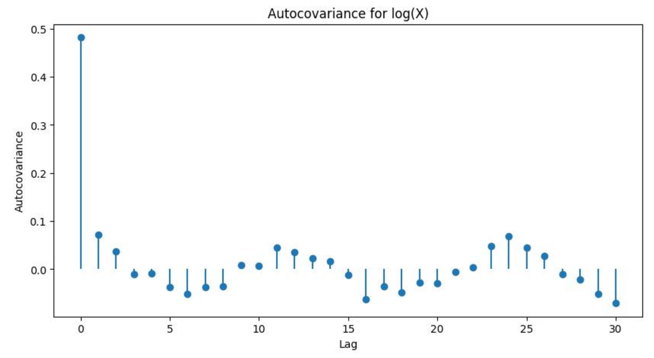
We want to explore the stationarity properties of our data in order to make a more informed modeling choice.
This looks to be particularly relevant for X since its dynamics exhibits less regularity.

I plotted rolling means and rolling variances with a time window of 300 observations in order to assess potential changes in the mean function and the covariance function over time.



I tried to apply an bijective transformation of the data in an attempt to stabilize the mean and the variance, in particular the function: f(x)=log(x).

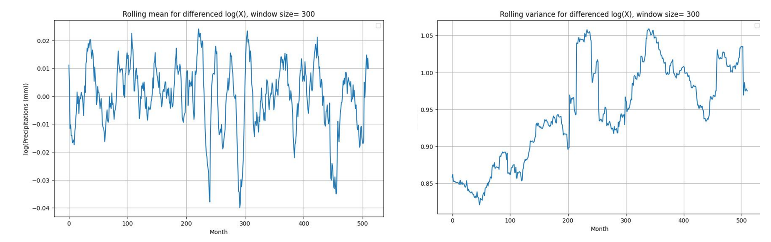




By plotting the auto-covariance for the logged version of X we see a major improvement in the regularity of its auto-regressive component.

Seasonal differencing and SARIMA - GARCH

Before any modeling choice we need to address the seasonality of our data. By applying seasonal differencing we can get the following behavior:



Both the mean and the variance exhibit a more stable pattern, however we notice a slight upward trend in the latter.
It becomes more apparent that the precipitation data is characterized by some sort of heteroskedasticity.

This fact could be addressed by a SARIMA model for the mean (that handles the seasonal differencing) supported by a GARCH model to model for the growing variance.

Analytically, our steps modeling steps can be written in this way:

Step 1: Log Transformation of Xt

$$X_t \mapsto \log(X_t)$$

Step 2: Fitting of a SARIMA((p,q,d)x(PQD)s) model to handle the seasonal and the auto-regressive components

$$\Phi(B^s)\phi(B)\nabla_s^d\nabla^d log(X_t) = \Theta(B^s)\theta(B)\epsilon_t$$

$$\phi(B) = 1 - \sum_{i=1}^{p}\phi_i B^i \qquad \Phi(B) = 1 - \sum_{i=1}^{P}\Phi_i B^i \qquad \nabla = 1 - B$$

$$\theta(B) = 1 + \sum_{i=1}^{q}\theta_i B^i \qquad \Theta(B) = 1 + \sum_{i=1}^{q}\Theta_i B^{is} \qquad \nabla_s = 1 - B^s$$

Step 3: Fitting a GARCH(p,q) model on the residuals of the SARIMA to handle heteroskedasticity.

$$\epsilon_t = \sigma_t z_t \qquad z_t \sim N(0,1)$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p}\alpha_i\epsilon_{t-i}^2 + \sum_{j=1}^{q}\beta_j\sigma_{t-j}^2$$

Model Implementation and Calibration

For the implementation and fitting of the SARIMA model I used a State Space representation of the model.
Using the fact that the polynomial equation in terms of the shift operator B can be expanded (I implemented a code for handling polynomial coefficients during multiplication) I reconstructed the State equation and the Observation Equation.

Given polynomials:

$$P_F(B) = \Phi(B^s)\phi(B)\nabla_s^d\nabla^d = 1 + \sum_{i=1}^{M} f_i B^i$$

$$P_H(B) = \Theta(B^s)\theta(B) = 1 + \sum_{i=1}^{N} h_i B^i$$

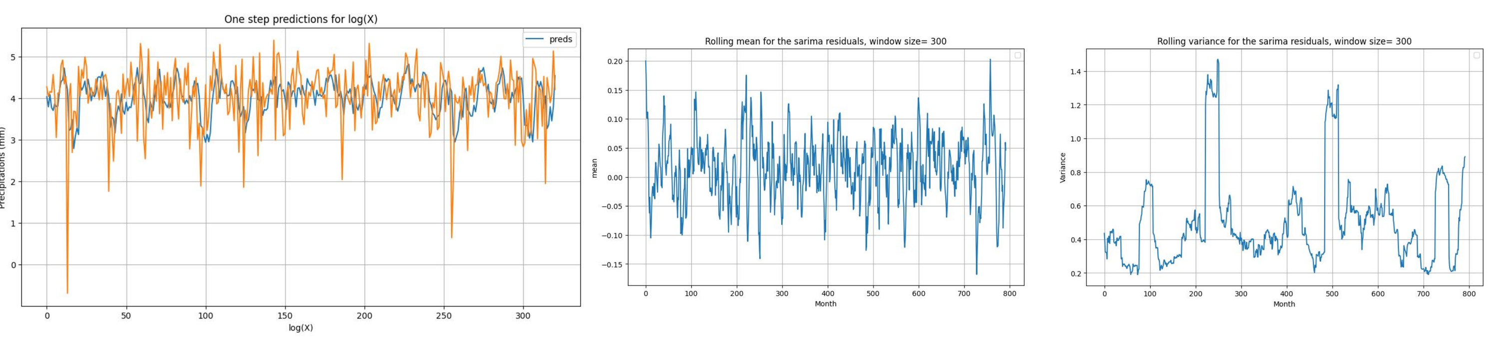
After defining n=max(N+1,M), we can construct the matrices:

$$F = \begin{bmatrix} -f_1 & -f_2 & \dots & -f_n \\ 1 & 0 & 0 & 0 \\ \dots & \dots & 0 & 0 \\ 0 & \dots & 1 & 0 \end{bmatrix} \qquad H = [1 \;\; h_1 \;\; \dots \;\; h_{n-1}] \qquad V_1 = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \dots & \dots & 0 & 0 \\ 0 & \dots & 0 & 0 \end{bmatrix}$$

We end up with the State Space representation:

$$\begin{cases} h_t = Fh_{t-1} + w_t \quad w_t \sim WN(0, V_1) \\ X_t = Hh_t \end{cases}$$

The model can then be calibrated by minimizing some loss function such as the log-likelihood or a mean square error.
I decided to minimize the mean square error, in particular the error between the one step predictor obtained by the Kalman Filter and the actual values.



SARIMAX for the Electricity Production

After fitting a GARCH model on the SARIMA residuals we get a complete description of the dynamics of X.

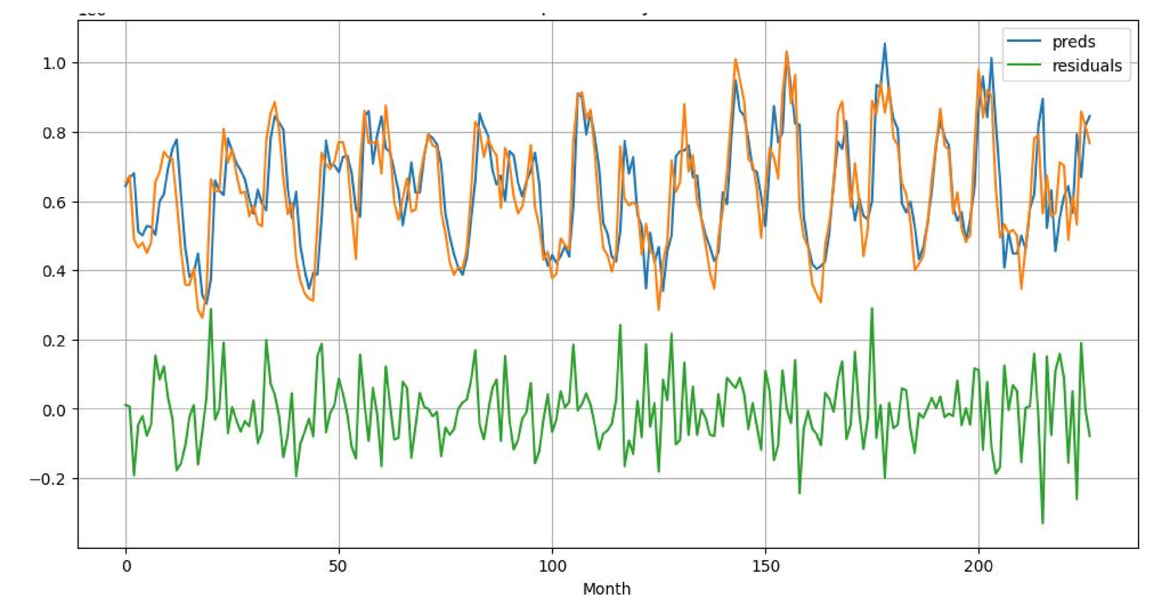
What remains to do is modeling the relationship between X and Y.
To do that I extended the SARIMA model to include an exogenous input.

$$\Phi(B^s)\phi(B)\nabla_s^d\nabla^d Y_t = \Theta(B^s)\theta(B)\epsilon_t + \sum_{i=1}^{k} u_j X_{t-j}$$

The state space representation changes slightly as we have to take into account an additional factor dt on the observation equation.

$$\begin{cases} h_t = Fh_{t-1} + w_t \quad w_t \sim WN(0, V_1) \\ X_t = Hh_t + d_t \end{cases} \qquad d_t = \sum_{i=1}^{k} u_j X_{t-j}$$

By applying again the Kalman filter given and minimizing over the the one step prediction error we can fit the model.
I provided a plot of the one step predictions obtained during the calibration of the model along with the plot of the residuals..



Forecasting

Ultimately, we can try to forecast electricity production data on longer time horizons incorporating our knowledge about the dynamics of the precipitations.
My approach consisted of chosing a time horizon T and iteratively applying the following two steps:

Algorithm: For t=1, t<=T,repeat:
Step 1: One-step prediction of the precipitations X using SARIMA-GARCH

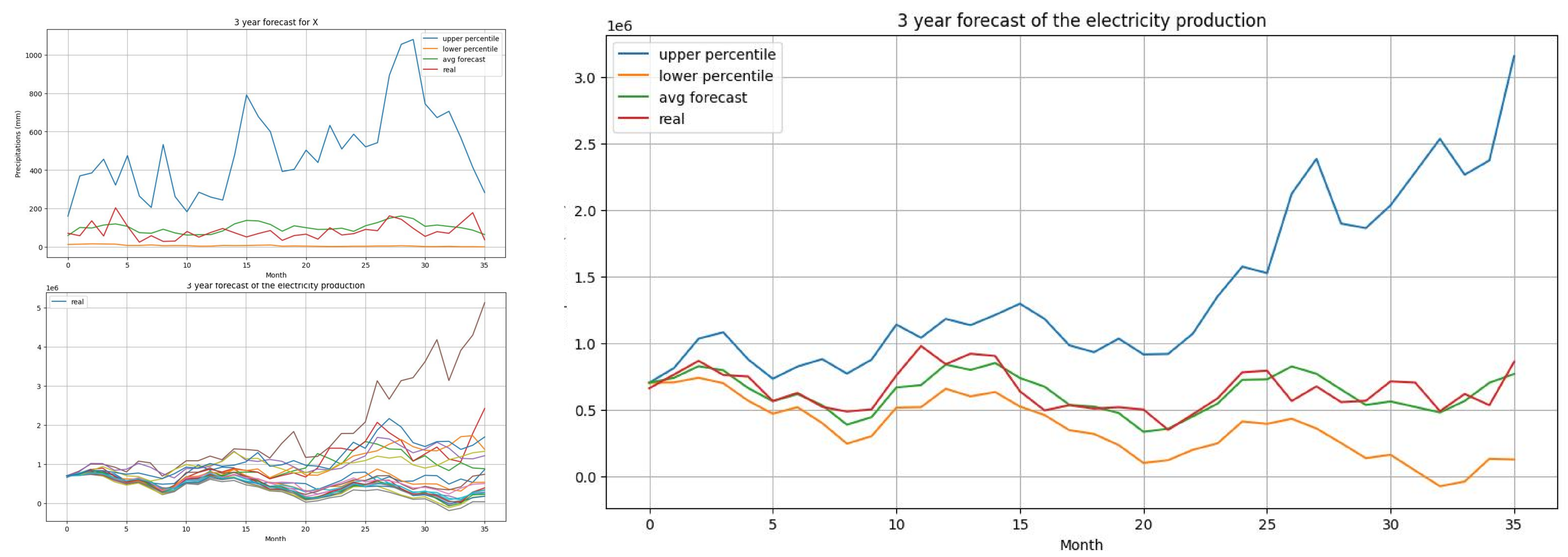
First we apply the Kalman filter to generate the mean prediction for the SARIMA model and we simulate a random error with our calibrated GARCH model.
Then we add the two contributions and take the exp() function (to invert our initial transformation) to get a simulation for X(t+1) given the past X data.

Step 2: One-step prediction of the electricity production Y using SARIMAX

Given the simulated X value we can use it as an input for the SARIMAX model,adding it to the current data.

In this way we can generate an arbitrary number of trajectories for Y which can be then used to compute confidence intervals.

Below I plotted forecast performance for both X and Y along with 95% confidence intervals.



Andrea Meschieri. 15/11/2024