Flowering date prediction for bulbous perennials

Dániel Nagy[1],
Imre M. Jánosi[2]

[1]Institute for Physics, Eötvös Loránd University, H-1117, Pázmány Péter sétány 1/A. Budapest, Hungary
[2]Department of Physics of Complex Systems, Eötvös Loránd University, H-1117, Pázmány Péter sétány 1/A. Budapest, Hungary

November 27, 2019

Abstract

The goal of this project is to efficiently predict the first flowering date of bulbous perennials and to identify, which meteorological parameters affect the flowering dates. The LSCD model is briefly presented, and the hyperparameters of the simpler LSC model are fitted using Gaussian process optimization. A neural network approach for the prediction of flowering dates is presented.

Introduction

Many observations prove that plants have a complex sense of climate, and that many of them developed the ability to prevent premature flowering. In order to uncover the relationship between meteorological data and flowering time, a long-term data acquisition was needed. The data available to my project is a `csv` file that is containing the first flowering dates of 329 bulbous perennial plants in the period 1968-2001. Further meteorological data is downloaded from the internet. The most significant meteorological parameter is the temperature of the soil at a depth of around 10cm. The temperature data is shown on 1.

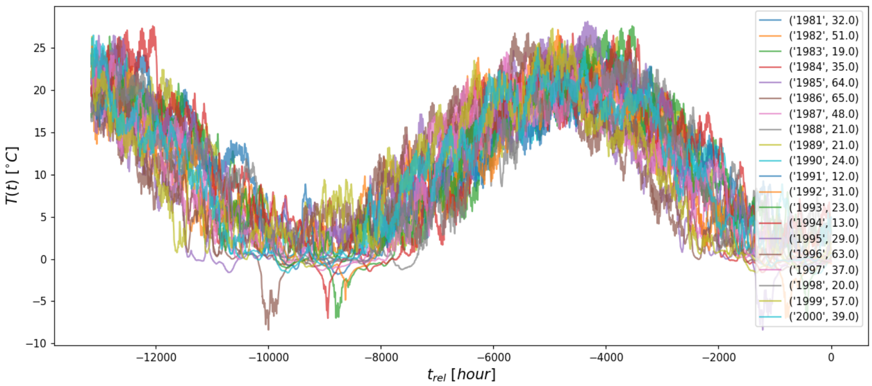

Figure 1: The soil temperature at 10cm depth across 22 years, plotted yearly at a 1-hour resolution. The x axis shows the hours before flowering date starting from 0.

The LSCD model

The LSCD model is built upon the observation that the flowering time of plants is highly correlated with the concentration of the Vernalization insensitive3 (VIN3) protein. The VIN3 mRNA levels are slowly rising with increasing weeks of cold exposure but rapidly decrease in the warm. Equations 1 and 2 describe the dynamics of spliced and unspliced VIN3 mRNA concentration.

$$\frac{\mathrm{d}\nu}{\mathrm{d}t} = p_\nu(L, S, C, D) - s_\nu \nu \tag{1}$$

$$\frac{\mathrm{d}V}{\mathrm{d}t} = s_\nu \nu - d_V V \tag{2}$$

In the equations above, $p_\nu(L, S, C, D) = L \cdot S \cdot C \cdot D$ is the productive transcription, s_ν is the splicing rate, and d_V is the degradation rate of the spliced VIN3.

According to the model, the VIN3 concentration is determined by 4 different intermediators called the Long-term, the Short-term, the Current and the Diurnal intermediator proteins. The concentration of these intermediators are described by equations 3-6.

$$\frac{\mathrm{d}L}{\mathrm{d}t} = \begin{cases} 1 - d_L L & T < T_L \\ -d_L L & T \geq T_L \end{cases} \tag{3}$$

$$C(T) = \begin{cases} p_{c1} & T \leq T_{c1} \\ p_{c1} - p_{c2}\frac{T-T_{c1}}{T_{c2}-T_{c1}} & T_{c1} < T < T_{c2} \\ p_{c2} & T \geq T_{c2} \end{cases} \tag{4}$$

$$S(T_m) = \begin{cases} 1, & T < T_S \\ S_1, & T \geq T_S \end{cases} \tag{5}$$

$$D(t) = \left[p_D + \sin\left(2\pi\left(t - \frac{t_m - 1}{24}\right)\right)\right]^2 \tag{6}$$

T_m is the maximum temperature since the last resetting, which was chosen to occur each day at 4pm. t_m is the time at dawn.

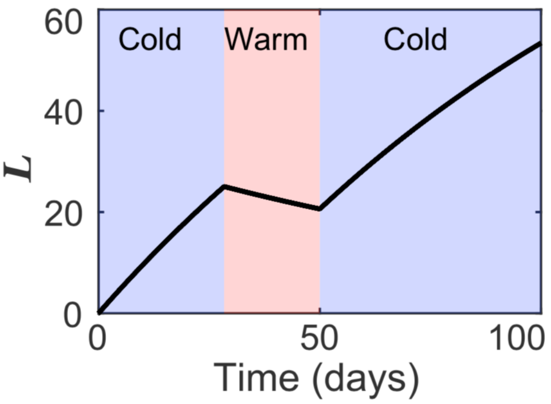

Figure 2: Illustration of change of L as a function of time and temperature.

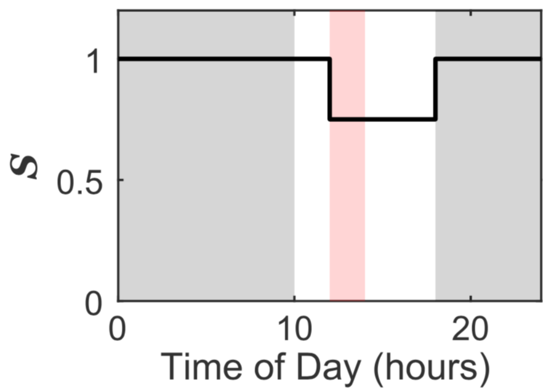

Figure 3: Illustration of change of S as a function of time and temperature.

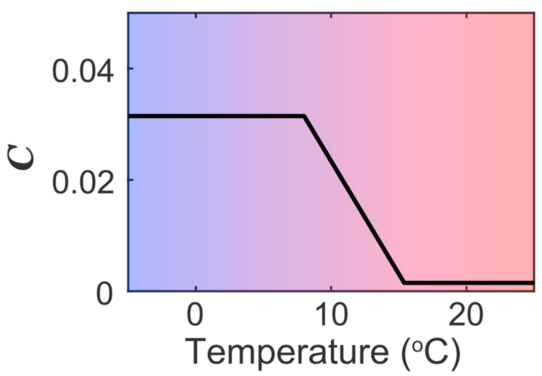

Figure 4: Illustration of change of C as a function of time and temperature.

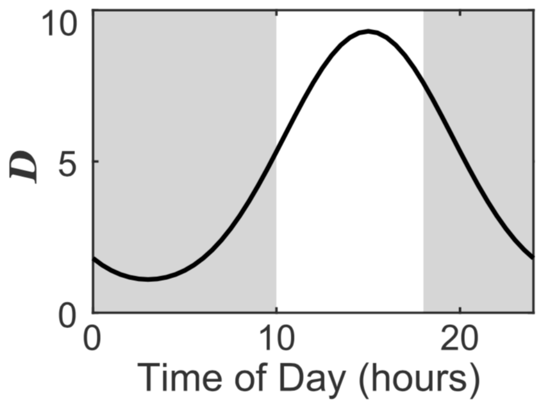

Figure 5: Illustration of change of D as a function of time and temperature.

Figures 2-5 show how the intermediator levels are changing with the temperature and time.

Results with the LSCD model

I investigated the LSCD model for a single type of bulbous perennial for a 20-year dataset.

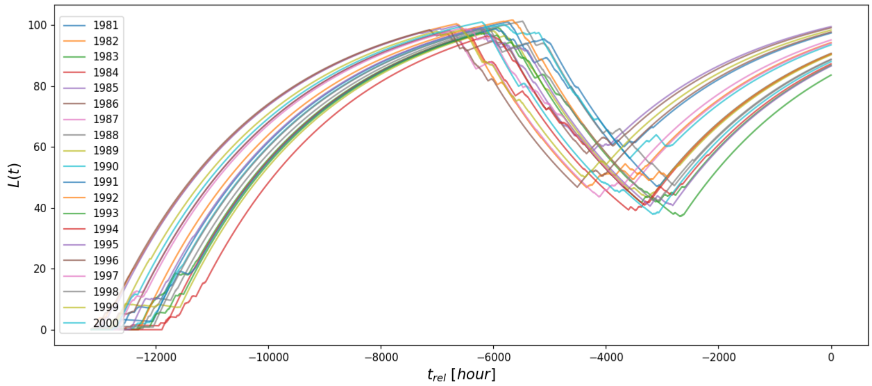

Figure 6: L intermediator for a one-year period temperature data across 20 years.

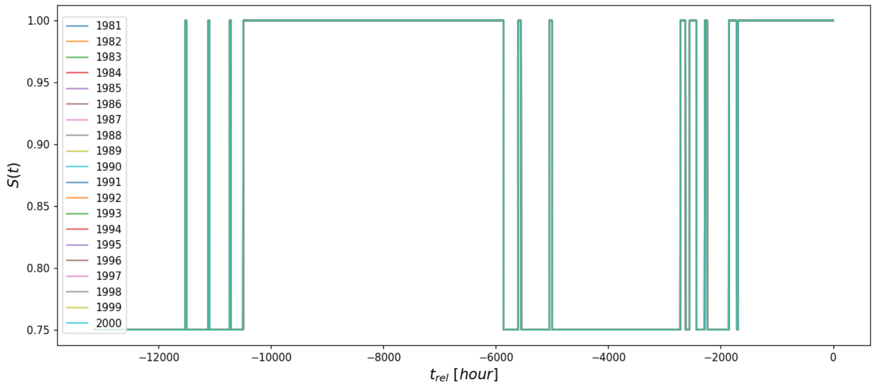

Figure 7: S intermediator for a one-year period temperature data across 20 years.

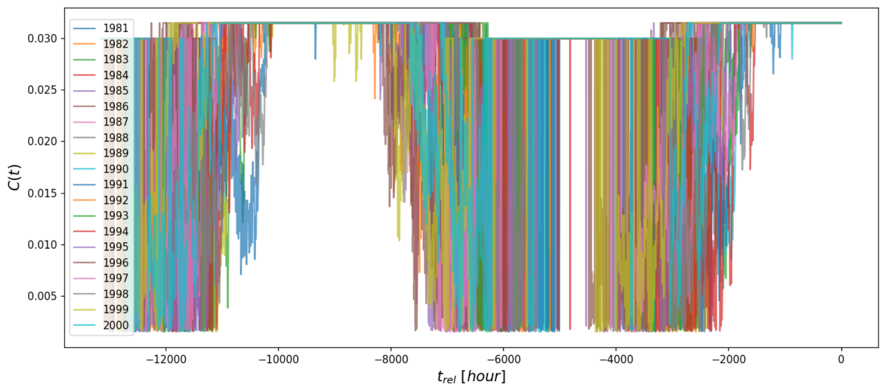

Figure 8: C intermediator for a one-year period temperature data across 20 years.

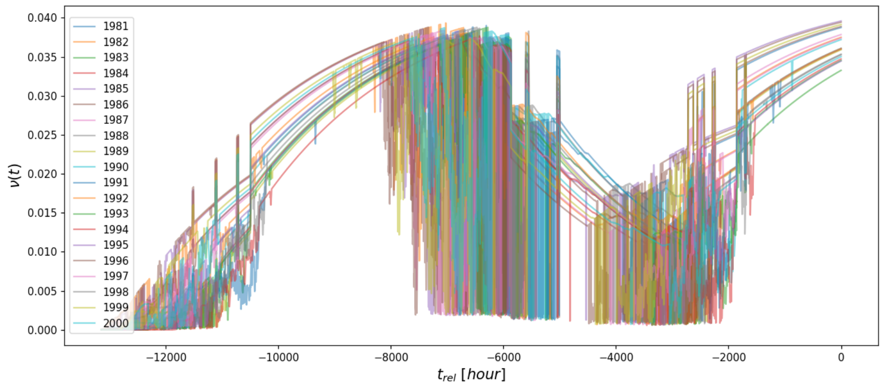

Figure 9: Spliced VIN3 concentration for a one-year period temperature data across 20 years.

Gaussian process optimization

The hyperparamteres of the simpler LSC model were searched with Gaussian process optimization using the `scikit-optimize` python package. The results of the fit are shown on figures 11 and 10.

param	dimension	initial value	fitted value
d_V	day^{-1}	18	19.7069
s_ν	day^{-1}	79.2	70.655
S_1	1	0.75	2.1732
T_L	°C	17	31.945
d_L	day^{-1}	0.009	0.074249
T_{c1}	°C	8	10.1501
T_{c2}	°C	15.4	14.4630
p_{c1}	1	0.0315	0.01038
p_{c2}	1	0.03	0.0309
p_D	1	2.05	-
T_S	°C	15	17.164

Table 1: Values of the parameters in the literature and values after applying GP optimizer on the LSC model.

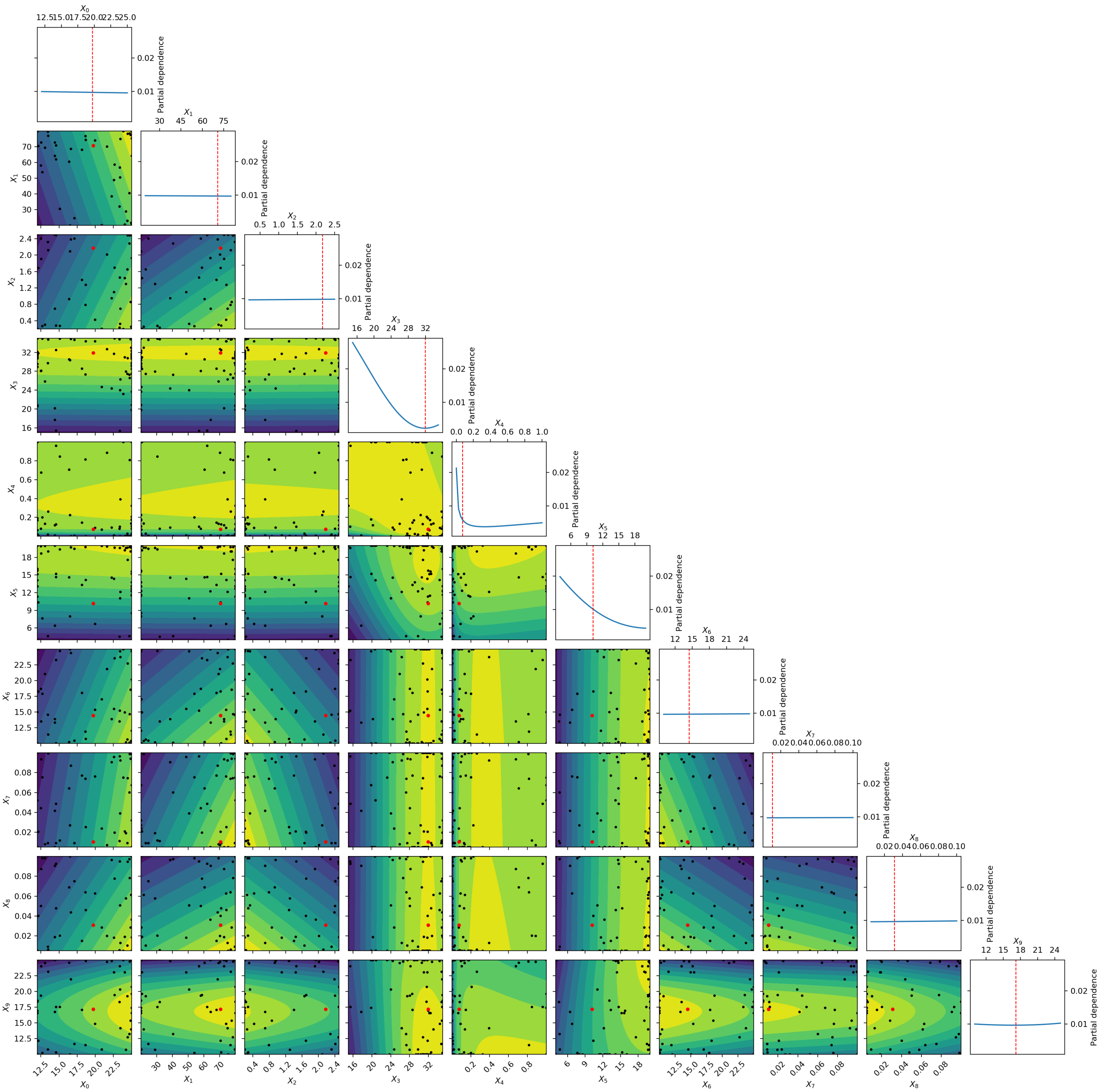

Figure 10: Results of objectives of the LSC model while fitting parameters with the GP optimizer.

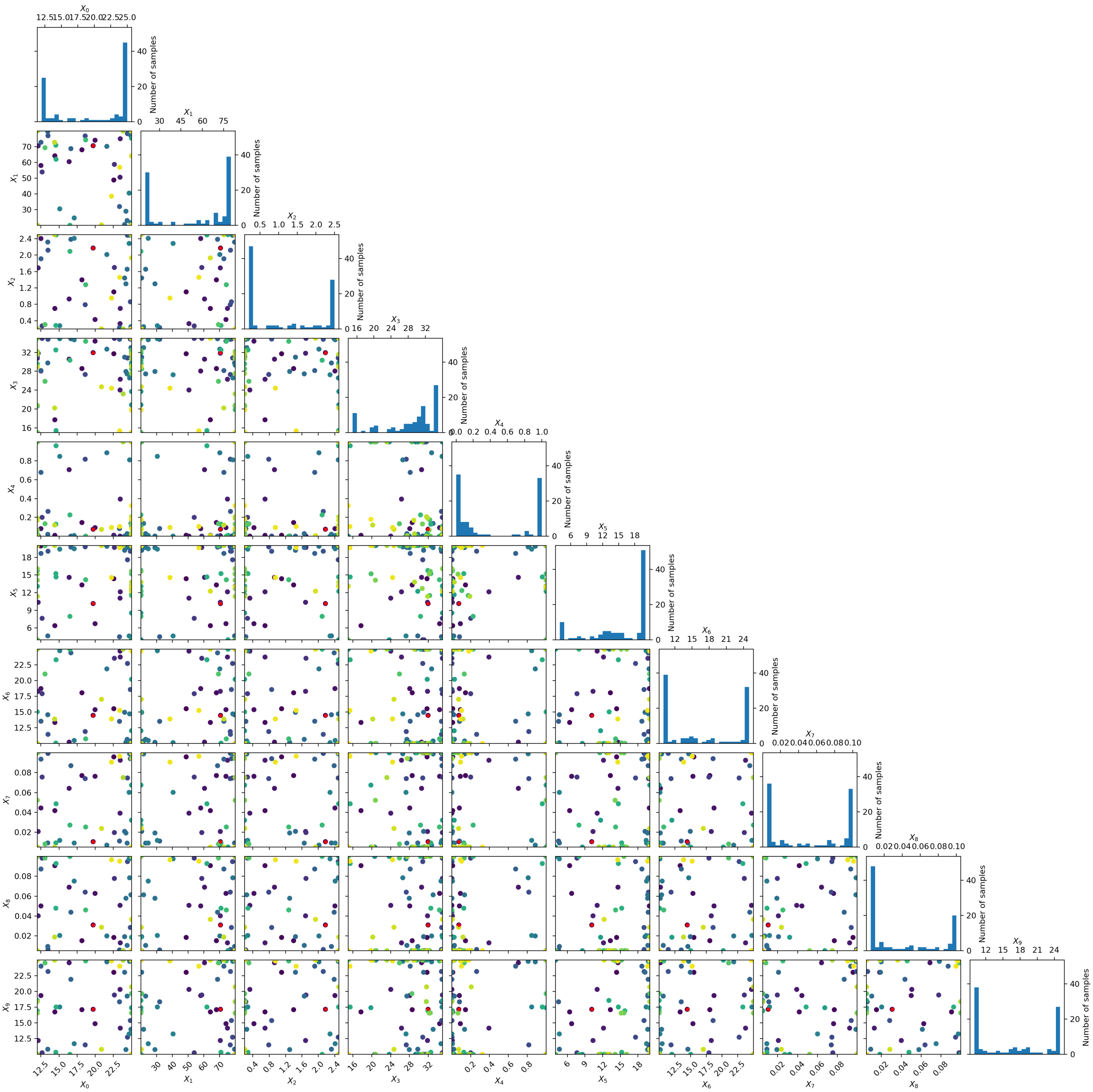

Figure 11: Results of evaluations of the LSC model with the GP optimizer.

Neural network approach

A neural network was implemented aiming to predict the time until the next flowering date bassed on a sequence of temperature measurement data. The network was feeded with a dataset created by systematically resampling the existing temperature series and labeling these sequances. The architecture is based on LSTM cells and was a single layer recursive neural network. The schema of this architecture is seen on figure 12.

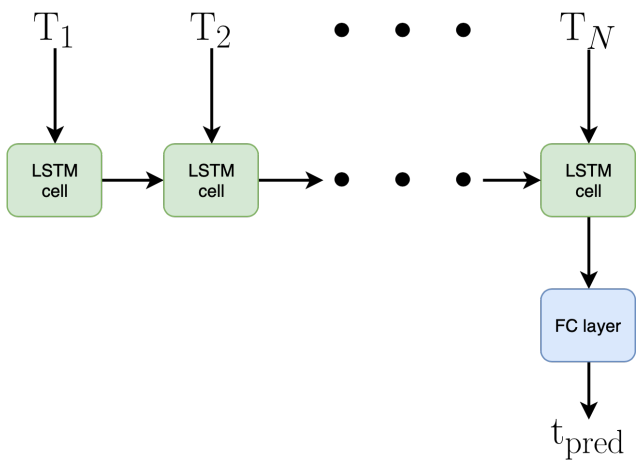

Figure 12: LSTM architecture for predicting flowering time from temperature measurement series.

The LSTM cells have cell state of 16 float values and also a hidden state of 16 floats. At the end there is a fully connected layer equipped with ReLU activation to predict a single float value.

Results of the neural network approach

The network was trained on a part of the available data with batch size 8. The model was trained for 300 epochs with Adam optimizer and the learning rate was set to 0.03. Figure 13 show the loss during training.

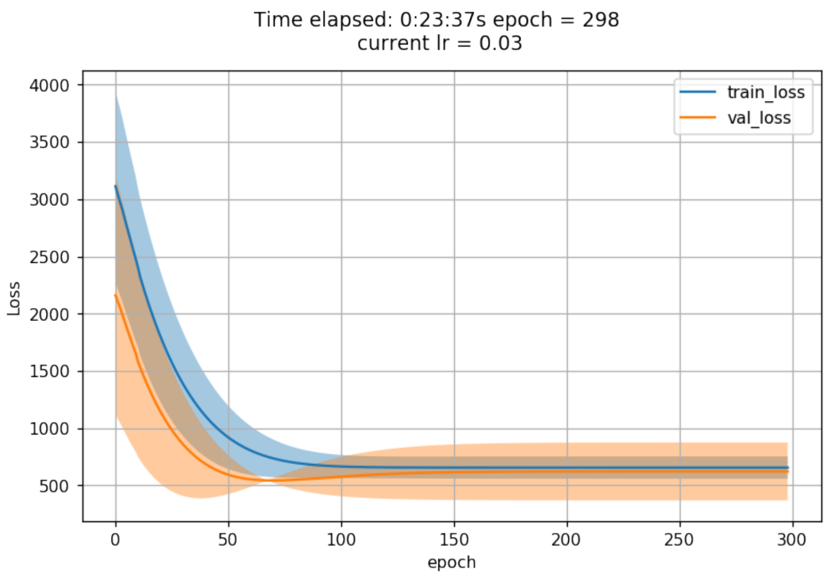

Figure 13: Loss functions on training and validation datasets across 300 epochs.

Altough the model was converging, the mean squared error is still large, meaning that it is plenty of room for improvement. The fact that this simple model was converging even for a small subset of the data, without overfitting, gives hope that significant improvement could be achieved in the future.