Pages 1–5
2022

Predicting Solar Activity Cycles using Probabilistic Machine Learning

Guner Aygin
MSci (Hons) Physics
School of Physics and Astronomy, University of Birmingham
12/12/2022

Keywords: Gaussian Process, Linear regression, Savitzky-Golay filter, Solar Cycle

ABSTRACT

Solar activity cycles have long been shown to be periodic, leading many to believe their properties are predictable. Solar cycle research is critical for the protection of public health and forecasting of space weather, allowing necessary precautions to be taken when planning future space-related missions. Existing methods used for predictions vary from simulating the dynamo processes in the sun to developing artificial neural networks, which make predictions solely from the available data. In this report, probabilistic machine learning techniques are used to analyse Savitzky-Golay filtered daily sunspot data. The time period of the solar cycles was investigated and shown to vary from cycle to cycle, averaging $T = 10.9$ years. The relationship between amplitude and descending time was also investigated and found to have a weak negative correlation $m = -1.74^{+1.44}_{-1.36}$, but with a significant probability of there being no correlation at all. A Gaussian process regression model was used to model the shapes of the cycles and make future predictions. The model is able to accurately fit the data but is unable to produce any plausible predictions.

1 INTRODUCTION

Solar magnetic activity cycles, or solar cycles, arise from the evolution of magnetic fluctuations within the solar convection zone (Balogh et al. 2015). A hydromagnetic dynamo process is believed to be responsible for these fluctuations, which occur when a rotating convective fluid, with the ability to conduct electricity, is able to generate a sustainable magnetic field (Wang and Jiang 2014).

Solar activity includes the release of energy stored within the solar magnetic field, which manifests as solar flares or coronal mass ejections (CMEs) (Wang and Jiang 2014). During periods of high solar activity, areas of the solar surface are prohibited from convecting, due to the high magnetic fields. These areas are much cooler than the surrounding surface and appear as dark spots, commonly known as sunspots. Solar activity is measured by counting the number of sunspots visible on the solar surface.

During a solar cycle, the intensity of solar radiation and the number of solar flares varies from a minimum to a maximum (Hathaway 2015). The Earth's atmosphere is greatly affected by variations in solar activity (Hargreaves 1992). High solar radiation amplifies the harmful ultraviolet and X-ray radiation dose absorbed by astronauts, airline pilots and passengers alike (Feminella and Storini 1997). Solar flares and CMEs greatly affect space weather and have the potential to cause catastrophic damage to satellites, wasting precious research time and costing billions to repair or replace (Siscoe 2000). Predicting the features of future solar cycles will allow necessary precautions to be taken here on Earth to protect public health, and to maximise success when planning long-term space missions (Pesnell 2012).

It has been known since the 19th century that the solar cycles exhibit a nearly periodic behaviour, with an average period of 11 years; the value and consistency of which is explored in Sections 3.1 & 4.1. Since observations began there have been 24 complete cycles, with Cycle 25 currently underway (Hathaway 2015). There have been numerous attempts to predict the shapes and features of the solar cycles but with very little success (Pesnell 2012). Apart from the period, solar cycles remain poorly understood, with sunspot numbers showing little sign of any predictable pattern. Most models have focused on predicting the physical processes driving the cycles (Charbonneau 2020), but in recent years researchers have attempted to model the cycles using Machine Learning (ML) techniques (Gonçalves, Echer, and Frigo 2020), which is the primary focus of this research.

The ability to predict solar cycles goes hand-in-hand with the ability to predict stellar cycles - which occur via the same processes as that of the Sun. Exoplanet research utilising the radial velocity (RV) method is greatly affected by stellar variations (Beurs et al. 2022; Czesla et al. 2009). Determining the optimum moment to observe a star involves foreseeing when stellar activity will be minimal. Stellar activity modelling presents the challenge of containing much sparser data, which means our ML models have less information to learn from. Stellar cycle predictions can be attempted using models which can predict solar cycles.

The primary aim of this project is to investigate the ability of ML techniques to predict the shape of future solar cycles. These models are centred around probabilistic techniques, differing from those which attempt to simulate the complex physical phenomena driving the solar cycles. Linear regression was used to model certain aspects of the solar cycle. A Gaussian process was the first method used which attempts to model the entirety of the solar cycles, with a plan to expand into utilising deep learning in the form of neural networks, building upon some of the existing research utilising ML. The feasibility of this project depends on whether we are able to make satisfactory cycle predictions. This project will conclude either that solar cycles are inherently unpredictable, or they can be predicted, utilising the technique found to produce the most likely outcomes.

2 LITERATURE REVIEW

Charbonneau 2020 models the solar cycle using a physics-based approach, modelling the magnetic fluctuations as a 'hydromagnetic dynamo process', involving the simulation of the magnetohydrodynamical induction equation (see Davidson 2002). The simulations themselves are computationally expensive and offer no concrete future predictions. As there is still much which remains to be understood about the physics behind solar cycles Charbonneau 2020 concludes by suggesting simpler models will be preferred for future, long-term, activity predictions.

Kitiashvili 2016 builds upon earlier dynamo models, incorporating a technique known as data assimilation, combining data with dynamo theory to provide an improved prediction for future cycles (see Kalnay et al. 1996). One issue with this method is that it continues to place too much predictive power on the physics-based theoretical model - and as we have already mentioned, solar cycles are still not well-understood.

Others have attempted to find relationships within the sunspot data, without modelling the cycle as a whole. Z Du and S Du 2006 claims that there is a strong linear correlation between the amplitude of a cycle and the descending time three cycles earlier (using the monthly-mean sunspot number). If true, this relationship can be used to inform our ML model of what we expect the future cycle amplitudes to reach. This correlation is investigated in Section 3.2 and 4.2.

ML techniques on the other hand can simplify the problem by removing all theoretical assumptions, and solely use the data to make predictions. In the last few years, there have been developments in the use of ML for solar cycle modelling. One such ML technique which is explored in Section 3 is regression, the most basic implementation of which is linear regression (Section 3.2). Gaussian process ($\mathcal{GP}$) regression models (see Section 3.3) are slowly gaining popularity for modelling solar cycles, with one of the first attempts by Gonçalves, Echer, and Frigo 2020. Camacho, Faria, and Viana 2022 and Barros et al. 2020 use $\mathcal{GP}$'s to model stellar variations to improve exoplanet detection; comparable to methods used to predict solar cycles; the former taking the $\mathcal{GP}$ one step further, by introducing $\mathcal{GP}$ regression networks for multi-output regression, providing additional flexibility to the model.

Much like $\mathcal{GP}$s, artificial neural networks (NN) are rapidly gaining popularity, especially for solving astrophysical problems (Bloom et al. 2012). NN are composed of artificial neurons, mimicking the human brain's behaviour, a form of ML known as 'deep learning'. Beurs et al. 2022 use NN to remove stellar activity signals from RV observations. NN could also be used for predicting future solar cycles. It has been shown that Bayesian neural networks can be used to approximate a $\mathcal{GP}$ (Agrawal, Papamarkou, and Hinkle 2020), which can be investigated in future research.

3 METHODS

ML algorithms are used to build models from training data to make future predictions. Probabilistic ML uses a Bayesian approach to data analysis providing posterior distributions for all the calculated model parameters (Gelman et al. 2020).

3.1 Basic Sunspot Data Analysis

The data used for this analysis is daily sunspot number data from 1818 to 2022, which is extremely noisy and difficult to model (see Fig. 1), but provides us with the rawest form of sunspot data.

A Savitzky-Golay (SG) filter was used to smooth the data (Press and Teukolsky 1990), improving on other methods which use mean monthly sunspot number (see Section 4.1). The smoothed data was used to train the models used thus far.

As previously mentioned, solar cycles have been known to experience a period of approximately 11 years, which was investigated. The positions and sunspot numbers of the maxima/minima of each solar cycle was determined from the smoothed signals, and the period of each cycle was calculated. The time between a cycle's maxima and subsequent minima (descending time) was also found, to analyse whether a linear relationship exists between amplitude and descending time three cycles earlier (see Section 2 & 4.2).

3.2 Linear Regression

Linear regression is a supervised ML technique used to predict outputs for continuous data and can be applied to the task of predicting sunspot numbers, as well as serving as a building-block for future ML models used throughout this project.

To test the linear regression methods linearly correlated data was created, containing Gaussian noise multiplied by an additional factor to increase the spread of the data. This was done to assess how well our model was able to fit with data containing small uncertainties but large scatter (such as the data in Section 4.2).

The most basic implementation of a linear regression algorithm is a simple least-squares (LS) fitting Press, Teukolsky, et al. 2007. A gradient descent (GD) algorithm was also used to iterate through possible parameters, eventually converging on the LS solution (Bottou 2012). The GD algorithm can be used for other optimisation problems, but the pre-built python modules are able to optimise much faster and are the preferred choice. Maximum likelihood estimation (MLE) is a common Frequentist technique used to determine the optimal parameters, with all three methods showing that they are able to accurately fit noisy, linear data.

One Bayesian approach utilised for linear regression is marginalisation. Here, prior distributions are assigned to each parameter, and the posterior distribution was sampled using a Markov chain Monte Carlo (MCMC) algorithm to find the posterior probability distributions (Andrieu et al. 2003; Camacho, Faria, and Viana 2022). Another approach, which is comparable to MLE, is finding the maximum *a posteriori* probability (MAP). MLE, MAP and MCMC were used to fit the descending time data in Section 4.2.

3.3 Gaussian Process Regression

A Gaussian process is "a collection of random variables, any finite number of which have a joint Gaussian distribution" (Williams and Rasmussen 2006). They can be used to model data where the functional form is not known *a priori* (Williams and Rasmussen 2006; Duvenaud 2014).

$\mathcal{GP}$ models depend entirely on the mean and covariance function (kernel). Duvenaud 2014 provides a list of many common kernels, including the squared exponential (SE) kernel and periodic kernel, which can be combined to form the *quasi-periodic* kernel:

$$k_{i,j} = A^2 \exp\left[-\frac{(x_i - x_j)^2}{2l_1^2} - \frac{\sin^2}{l_2^2}\left(\frac{\pi(x_i - x_j)}{P}\right)\right] + \sigma^2\delta_{ij} \quad (1)$$

Where $k_{i,j}$ is the covariance matrix, A is the amplitude, P is the period, l_1 and l_2 are the length scales of the SE and periodic kernels respectively, and σ is the uncertainty. This particular kernel is used for modelling the solar cycles as the cycles are observed to be periodic with varying amplitudes. Using the cycle maxima data and the squared exponential kernel it was also possible to model the variability of the amplitudes, which could in future be used in the construction of a deep and wide neural network.

A common approach taken for finding the optimal kernel parameters is to calculate the MAP or simply use MLE (Gonçalves, Echer, and Frigo 2020). These methods only provide a single value for the parameters, with no uncertainty. Marginalisation integrates over all the data and is a much more rigorous and robust approach for finding the optimal kernel parameters. Marginalisation was used when training the $\mathcal{GP}$, which may provide more accurate predictions compared with other attempts. The calculated marginalised parameters are shown in Fig. 4.

The most difficult aspect of marginalisation is setting the appropriate parameter priors so that the posteriors peak on a single value, rather than multiple ones. This involved setting extremely wide priors to start with; where multiple peaks were detected, the priors were adjusted to only sample from regions close to the most appropriate value.

$\mathcal{GP}$s have certain drawbacks, namely the computation time scales as $\mathcal{O}(N^3)$. This may be improved upon using the library *celerite* (see Section 5).

4 RESULTS

4.1 Smoothing & Cycle Period

Three different parameters were used for the SG filter, shown in Fig. 1. As a result, three different sets of cycle periods were obtained, with

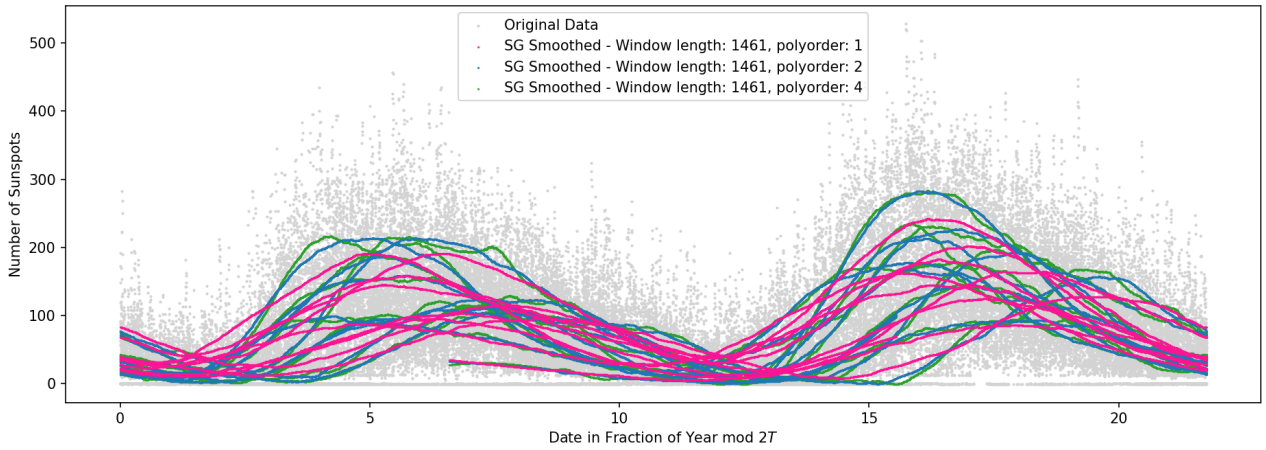

Fig. 1. Phase diagram of the three smoothed functions, with Savitzky-Golay parameters displayed. Phase was calculated by finding the modulus of each cycle's period and twice the average period (21.7 years). The figure shows how the periods vary from the expected $\sin^2(2T)$ they would display if they all had the exact same period.

periods ranging from 8.48 years to 12.77, and a mean $T = 10.9$ years, slightly less than the widely accepted 11 years. Figure 1 highlights the inconsistency of the solar cycle periods, as we can see that the cycles do not perfectly line up. It also showcases the difference in each cycle's amplitudes, which has been the main difficulty with solar cycle predictions.

4.2 Amplitude & Descending Time relationship

The smoothed signal with polyorder = 1 was used to calculate the amplitudes and descending times, as it was the smoothest function and had the most clearly defined maxima. As the data was so noisy the uncertainties were set to zero, similar to the fake data which was used for testing. The results of the linear regression can be seen in Fig. 2.

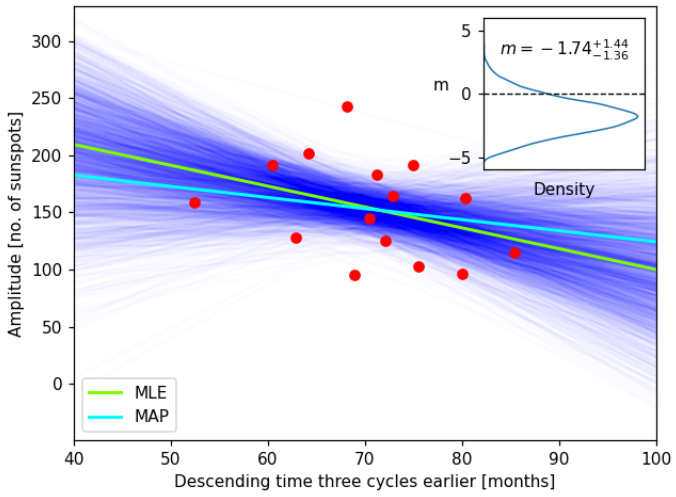

Fig. 2. Plot of Amplitude against Descending Time three cycles earlier. Data points are shown in red, with MCMC samples plotted in blue. Both MLE and MAP best fits are plotted for comparison. The subplot shows the posterior distribution of m, with maximum probability density occurring at $m = -1.74^{+1.44}_{-1.36}$. The black dotted line shows the location of $m = 0$ in our posterior distribution.

Fig. 2 mostly shows a negative correlation, however, some of the traces show a positive one. The relationship doesn't seem to be as definite as the one suggested in Z Du and S Du 2006. The subplot in Fig. 2 showing the posterior probability of m reveals that there is a non-negligible probability that the gradient is simply 0, which would imply no correlation between the two variables. As the correlation is too noisy it cannot reliably be used to predict cycle amplitudes based on past descending times.

4.3 Gaussian Process predictions

To date, the best solar cycle fitting has come from using the $\mathcal{GP}$ described in Section 3.3, with the quasi-periodic kernel. The plot of the training data and $\mathcal{GP}$ samples can be seen in Fig. 3. Only 200 data points were used, to save computation power, but the $\mathcal{GP}$ does an excellent job of modelling the cycle between 1818 and 2022.

However, when looking beyond the last training point, we notice that the predictions very quickly deviate from what one would expect the cycles to vaguely look like. We would expect the cycles to oscillate from a minimum, close to zero, to a maximum, over a period of around 11 years. This signifies that, although our $\mathcal{GP}$ models the data very well, it cannot be used for future cycle predictions just yet.

There are two hypothesised reasons for the lack of predictive power of our current $\mathcal{GP}$. The first is that l_1 is far too short. No matter how wide the prior was made, the posterior always peaked around 0.59. This indicates the model very quickly forgets the shape of previous cycles, and as a result, reverts back to the mean. This could also suggest that, intrinsically, the sunspot data may be too random to exhibit any long-term predictable pattern. An attempt was made to produce more satisfactory predictions by fixing the value of l_1 to a 100, rather than marginalising. This, however, did not yield any plausible results either.

The second hypothesis is that the mean function should be a function of time. This would allow the future predictions to oscillate due to the mean, with the rest of the $\mathcal{GP}$ accounting for the noise of the signal. This can be investigated in future work.

Similar to the smoothed signal, when a $\mathcal{GP}$ was used to model the amplitudes it too provided no concrete future predictions, again reverting back to the mean, but was able to model the data very well.

The computation time taken to train our model was greatly affected by the number of parameters in the kernel. In future, when adding more complexity to the models, one can fix specific parameters to the values previously obtained by marginalisation, only marginalising over new parameters.

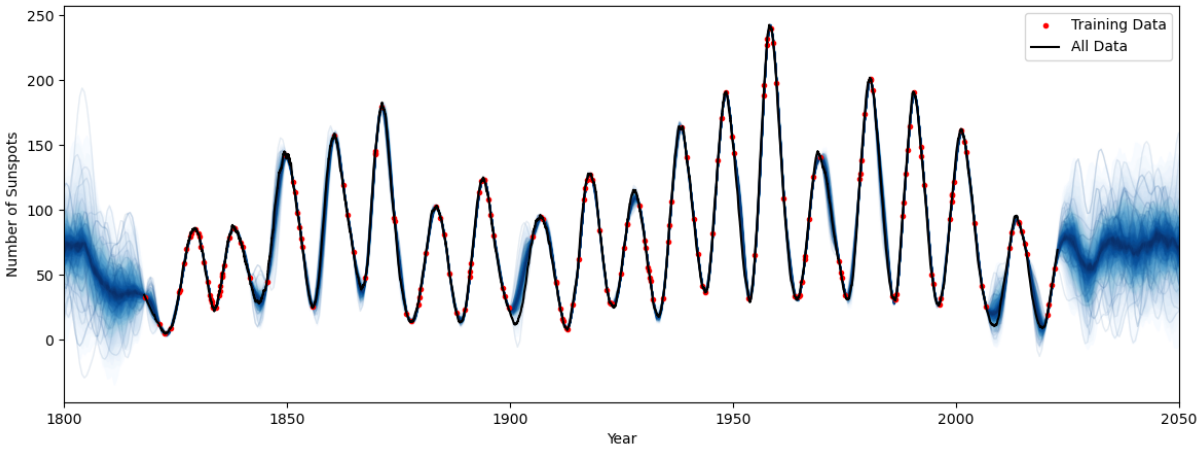

Fig. 3. $\mathcal{GP}$ regression plots (blue), trained using samples from the smoothed signal (red) and a quasi-periodic kernel (equation 1). The complete data set is plotted (black) for comparison. The $\mathcal{GP}$ makes predictions starting from 1800 to 2050, to assess how well it can predict future cycles.

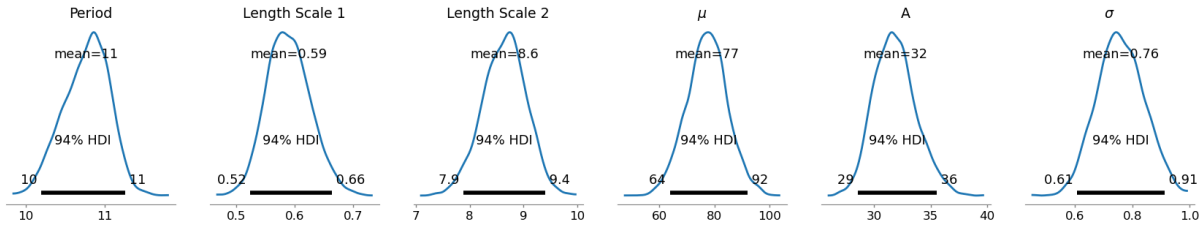

Fig. 4. Posterior probability distributions for the parameters used in the quasi-periodic kernel (equation 1), as well as the mean function μ, for the $\mathcal{GP}$ in Fig. 3. The graphs show the 94% highest density interval (HDI).

5 PROJECT PLAN

The remainder of this project will be spent improving on the existing methods outlined, as well as utilising new ML techniques, notably deep learning techniques.

The first task is the improvement of the $\mathcal{GP}$ model. As mentioned, there is an issue with the predictive power of this $\mathcal{GP}$, which can be fixed by improving the kernel and mean function such that some legitimate predictions can be obtained. This will occur in semester 2, week 1 (W1).

Once the $\mathcal{GP}$ is able to make sufficient predictions, a faster method known as celerite will be trialled in W2, which has a computation time scaling as $\mathcal{O}(N)$ with the number of data points. However, this method requires multiple simple harmonic oscillator equations to be combined so that our model has the appropriate form, which scales as $\mathcal{O}(J^2)$ with the parameters.

From here, a new approach can be taken, involving the use of NN (see Section 2). In W3 the theory of NN will be investigated, and simple models involving fake data will be created and compared with $\mathcal{GP}$s.

NN can be used to approximate a $\mathcal{GP}$, which will be the primary investigation of W4-5. The differences between NN and $\mathcal{GP}$ can be more clearly observed through this process and will give a clearer understanding of the mechanisms behind these techniques.

There are different types of NN which will be investigated, notably standard, recurrent, convolution, and generative adversarial networks. The time frame for implementing each of these methods will be at least 1 week per method, which brings takes the project to W9.

As a culmination of this project, a deep and wide neural network will be created, which can use the $\mathcal{GP}$ model to better inform the deep neural network, with the hope that this provides better predictions.

In the final few weeks different data sets, including space-weather research, will be analysed alongside the sunspot data and fed into the wide neural network. The more data our model has regarding solar cycles the better the predictions should, theoretically, be.

New sunspot data will also be used when training the final models, containing information stretching further back in time, which will give the models more information with which to train.

If one of the models is able to successfully predict the solar cycles then we can extend the project to include stellar cycle data, which is much sparser. The properties of individual stellar cycles are less familiar to the scientific community, and so a ML model is the ideal tool for modelling and predicting the shape of future cycles. With many stellar cycle predictions, an analysis can be conducted into the efficacy of our ML algorithms when the stellar cycles have evolved.

As I am undertaking this project myself, there is no division of labour in any aspect of this project. My supervisor Guy Davies has been aiding me thus far and will continue to do so throughout the next stages of this project, to maximise its success.

ACKNOWLEDGEMENTS

I would like to thank Guy Davies for all his continued support throughout this project, providing me with the resources to expand my knowledge of statistics and Gaussian processes, as well as improving my coding practices and pushing me to attempt advanced ML techniques.

REFERENCES

Agrawal, Devanshu, Theodore Papamarkou, and Jacob Hinkle (2020) "Wide neural networks with bottlenecks are deep Gaussian processes". In: *Journal of Machine Learning Research* 21.175.

Andrieu, Christophe et al. (2003) "An introduction to MCMC for machine learning". In: *Machine learning* 50.1, pp. 5–43.

Balogh, A et al. (2015) "Introduction to the solar activity cycle: Overview of causes and consequences". In: *The Solar Activity Cycle*, pp. 1–15.

Barros, SCC et al. (2020) "Improving transit characterisation with Gaussian process modelling of stellar variability". In: *Astronomy & Astrophysics* 634, A75.

Beurs, Zoe L de et al. (2022) "Identifying Exoplanets with Deep Learning. IV. Removing Stellar Activity Signals from Radial Velocity Measurements Using Neural Networks". In: *The Astronomical Journal* 164.2, p. 49.

Bloom, JS et al. (2012) "Automating discovery and classification of transients and variable stars in the synoptic survey era". In: *Publications of the Astronomical Society of the Pacific* 124.921, p. 1175.

Bottou, Léon (2012) "Stochastic gradient descent tricks". In: *Neural networks: Tricks of the trade*. Springer, pp. 421–436.

Camacho, JD JP Faria, and PTP Viana (2022) "Modelling stellar activity with Gaussian process regression networks". In: *arXiv preprint arXiv:2205.06627*.

Charbonneau, Paul (2020) "Dynamo models of the solar cycle". In: *Living Reviews in Solar Physics* 17.1, pp. 1–104.

Czesla, S et al. (2009) "How stellar activity affects the size estimates of extrasolar planets". In: *Astronomy & Astrophysics* 505.3, pp. 1277–1282.

Davidson, Peter Alan (2002) *An introduction to magnetohydrodynamics*.

Du, Zhanle and Shouyu Du (2006) "The relationship between the amplitude and descending time of a solar activity cycle". In: *Solar Physics* 238.2, pp. 431–437.

Duvenaud, David (2014) "Automatic model construction with Gaussian processes". PhD thesis. University of Cambridge.

Feminella, Francesco and Marisa Storini (1997) "Large-scale dynamical phenomena during solar activity cycles." In: *Astronomy and Astrophysics* 322, pp. 311–319.

Gelman, Andrew et al. (2020) "Bayesian workflow". In: *arXiv preprint arXiv:2011.01808*.

Gonçalves, Ítalo G Ezequiel Echer, and Everton Frigo (2020) "Sunspot cycle prediction using warped Gaussian process regression". In: *Advances in Space Research* 65.1, pp. 677–683.

Hargreaves, John Keith (1992) *The solar-terrestrial environment: an introduction to geospace-the science of the terrestrial upper atmosphere, ionosphere, and magnetosphere*. Cambridge university press.

Hathaway, David H (2015) "The solar cycle". In: *Living reviews in solar physics* 12.1, pp. 1–87.

Kalnay, Eugenia et al. (1996) "The NCEP/NCAR 40-year reanalysis project". In: *Bulletin of the American meteorological Society* 77.3, pp. 437–472.

Kitiashvili, Irina N (2016) "Data assimilation approach for forecast of solar activity cycles". In: *The Astrophysical Journal* 831.1, p. 15.

Pesnell, W Dean (2012) "Solar cycle predictions (invited review)". In: *Solar Physics* 281.1, pp. 507–532.

Press, William H and Saul A Teukolsky (1990) "Savitzky-Golay smoothing filters". In: *Computers in Physics* 4.6, pp. 669–672.

Press, William H Saul A Teukolsky, et al. (2007) *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press.

Siscoe, George (2000) "The space-weather enterprise: past, present, and future". In: *Journal of Atmospheric and Solar-Terrestrial Physics* 62.14, pp. 1223–1232.

Wang, Jingxiu and Jie Jiang (2014) "Magnetohydrodynamic process in solar activity". In: *Theoretical and Applied Mechanics Letters* 4.5, p. 052001.

Williams, Christopher KI and Carl Edward Rasmussen (2006) *Gaussian processes for machine learning*. Vol. 2. 3. MIT press Cambridge, MA