MNRAS 000, 1–13 (2025) Preprint 2 December 2025 Compiled using MNRAS LATEX style file v3.2

Nested Sampling for ARIMA Model Selection in Astronomical Time-Series Analysis

Ajinkya Naik[1]★, Will Handley[1]†
[1]Institute of Astronomy, University of Cambridge, Cambridge, CB3 0HA, UK

ABSTRACT

The upcoming era of large-scale, high-cadence astronomical surveys demands efficient and robust methods for time-series analysis. ARIMA models provide a versatile parametric description of stochastic variability in this context. However, their practical use is limited by the challenge of selecting optimal model orders while avoiding overfitting. We present a novel solution to this problem using a Bayesian framework for time-series modelling in astronomy by combining Autoregressive Integrated Moving Average (ARIMA) models with the Nested Sampling algorithm. Our method yields Bayesian evidences for model comparison and also incorporates an intrinsic Occam's penalty for unnecessary model complexity. A vectorized ARIMA–Nested Sampling framework is implemented allowing us to perform model selection across grids of Autoregressive (AR) and Moving Average (MA) orders, with efficient inference of selected model parameters. The method is validated on simulated and real astronomical time series, including the yearly sunspots number record, Kepler lightcurve data of the red giant KIC 12008916, and TESS photometry of the exoplanet host star Ross 176. In all cases, the algorithm correctly identified the true or best-fitting model while simultaneously yielding well-constrained posterior distributions for the model parameters. Our results demonstrate that Nested Sampling offers a potentially rigorous alternative to autoregressive model selection in astronomical time-series analysis.

Key words: methods: statistical, methods: data analysis, *(Sun:)* sunspots, exoplanets.

1 INTRODUCTION

Time series analysis is a crucial tool in the field of time domain astronomy. As next generation astronomical surveys and telescopes are set to yield unprecedented volumes of complex time series data, an efficient preliminary method for their analysis is necessary. Some examples of such methods are frequency domain analysis (Lomb 1976; Scargle 1982), Gaussian Processes (Foreman-Mackey et al. 2017) and Machine Learning methods (Richards et al. 2011; Baron 2019). Within this diverse landscape of methodologies, parametric autoregressive modelling offers a complementary approach to time series analysis in the form of ARIMA models (Elorrieta, Felipe et al. 2019; Akhter et al. 2020; Carruba & Aljbaae 2021).

The ARIMA (Autoregressive-Integrated-Moving Average) framework is a suite of models employed in analysing and forecasting time series data from various domains including, but not limited to, economics, finance and climate science. The Autoregressive (AR) component of ARIMA was first introduced by Yule (1927) to study the number of sunspots. These models capture the autocorrelation in a time series by linearly regressing the present value over its own past (lagged) values. Moving Average (MA) models on the other hand, operate on a similar principle by expressing the present value as a linear combination of past forecast errors or residuals. The idea of systematically combining these modelling methods can be traced back to Box & Jenkins (1976), who formalized them into ARIMA with the introduction of the Integrated (I) component to handle non-stationary time series. These are labelled as ARIMA(p, d, q) models, characterized by the respective orders p, d and q of the Autoregressive (AR), Integrated (I) and Moving Average (MA) components.

ARIMA models are seldom used in the analysis of astronomical time series data. The primary reason being that they require evenly sampled data in time, which is often in contrast to ground based astronomical datasets due to their irregular cadences. Nevertheless, ARIMA modelling can still be applied, with reasonable success, by binning light curve data with moderately irregular cadences (Feigelson et al. 2018). Upcoming facilities such as the Vera C. Rubin Observatory's Legacy Survey of Space and Time (LSST) (Ivezić et al. 2019), the Roman Space Telescope (Spergel et al. 2015; Johnson et al. 2020), and ongoing space missions like TESS (Ricker et al. 2015) and Gaia (Gaia Collaboration et al. 2016) will also produce adequately regular and high cadence time series data suitable for ARIMA modelling . The generality and flexibility of the ARIMA framework allows it to efficiently model a diverse range of astronomical time series. However, this very flexibility comes with the risk of over-parametrization and overfitting. It is non-trivial to choose the optimal (p, d, q) order of the ARIMA model. The heuristic diagnostics approach taken by the standard Box-Jenkins methodology, used for ARIMA model specification and validation, may limit robustness in some cases. Model selection on the basis of the Akaike Information Criterion (AIC) (Akaike 1974) and Bayesian Information Criterion (BIC) (Schwarz 1978) rely on maximum likelihood

★ E-mail: najinkya1313@gmail.com
† E-mail: wh260@cam.ac.uk

optimization, which may bias the model selection, especially for complex likelihood shapes.

The Nested Sampling algorithm used in Bayesian computation has shown promise in model selection problems in the context of Astrophysics (Trotta 2008). Recent advances in GPU accelerated Nested Sampling (Yallup et al. 2025) have reduced the computational costs associated to this method, allowing it to be implemented for a variety of model selection problems (Ormondroyd et al. 2025; Lovick et al. 2025; Leeney et al. 2025; Prathaban et al. 2025a; Yallup 2025). Using Nested Sampling in conjunction with ARIMA models can thus provide an alternative solution to the problem of selecting the right (p, d, q) order, with the added benefit of also returning posterior samples for the ARIMA parameters. The previously discussed risk of overfitting is avoided with an in-built Occam's penalty imparted on the Bayesian evidences of over-parametrized model fits.

In the following section, we discuss the mathematical formalism of ARIMA models and review the nested sampling algorithm. Section 3 then describes the methods and details of using the Nested Sampling algorithm for ARIMA model selection. The results of applying this framework on artificial and real astronomical time series data are presented in Section 4. We discuss potential future directions for this work and end with conclusions in Section 5.

2 BACKGROUND

Subsection 2.1 outlines the mathematical background and properties of ARIMA models. For a more comprehensive exposition of ARIMA models, readers are directed to the standard texts of Box & Jenkins (1976). In subsection 2.2, we discuss the nested sampling algorithm and its role in model selection.

2.1 ARIMA Framework

The general forecasting equation for ARIMA modelling is a combination of the Autoregressive (AR) and Moving Average (MA) model equations. An AR(p) equation to obtain the forecast $\hat{y}_t$ for an observed point y_t of the time series is

$$\hat{y}_t = c + \sum_{a=1}^{p} \phi_a y_{t-a} + \epsilon_t \tag{1}$$

It involves a linear weighted sum over p lagged values of the time series with a constant intercept term c (usually related to the long-term mean or "drift" of the time series) and a random forecasting error ϵ_t added to the forecast. The random noise ϵ_t is assumed to be homoscedastic and drawn from a normal distribution. Similarly, an MA(q) model is expressed by a linear weighted sum of q lagged forecast errors, again with a constant term c and a random error ϵ_t

$$\hat{y}_t = c + \sum_{m=1}^{q} \theta_m \epsilon_{t-m} + \epsilon_t \tag{2}$$

Equations 1 and 2 together represent an ARMA process:

$$\hat{y}_t = c + \sum_{a=1}^{p} \phi_a y_{t-a} + \sum_{m=1}^{q} \theta_m \epsilon_{t-m} + \epsilon_t \tag{3}$$

An ARMA process fundamentally assumes the time series data to be stationary. A time series is defined to be stationary if its statistical properties, such as mean, variance and autocorrelation, remain constant over time. The Integrated (I) component of ARIMA models is an added utility to deal with non-stationary time series using finite differencing. Its associated order d is the number of times the raw observations are differenced to render the series stationary. After fitting the differenced time series to an ARMA process, the forecast values $\hat{y}_t$ are integrated back to recover the original sequence trend.

To facilitate a compact representation of the general ARIMA(p, d, q) equation, the Backshift (or Lag) operator $\mathcal{B}$, with the action of shifting a quantity back by one time-step, is introduced:

$$\mathcal{B} y_t = y_{t-1} \qquad \mathcal{B}\epsilon_t = \epsilon_{t-1} \tag{4}$$

The AR and MA components of the ARIMA process can then be written as

$$(1 - \phi_1\mathcal{B} - \ldots - \phi_p\mathcal{B}^p) y_t = \Phi^p(\mathcal{B}) y_t \tag{5}$$

$$(1 + \theta_1\mathcal{B} + \ldots + \theta_q\mathcal{B}^q)\epsilon_t = \Theta^q(\mathcal{B})\epsilon_t \tag{6}$$

Here, $\Phi^p(\mathcal{B})$ and $\Theta^q(\mathcal{B})$ are the characteristic polynomials of the backshift operator for the AR and MA components. The differenced time series Y_t associated with the Integrated (I) part can also be represented in terms of $\mathcal{B}$ as

$$(1 - \mathcal{B})^d y_t \tag{7}$$

Finally, combining Equations 5, 6 and 7 condenses the forecasting equation for an ARIMA(p, d, q) model:

$$\Phi^p(\mathcal{B})(1 - \mathcal{B})^d y_t = \Theta^q(\mathcal{B})\epsilon_t + c \tag{8}$$

2.1.1 Stationarity and Invertibility Constraints

Even though the problem of non-stationarity in the time series data y_t is solved by differencing, the ARIMA model fitted to the data must itself produce a stationary and stable time series forecast. This requirement manifests in the form of two mathematical constraints on the AR and MA weights ϕ_a and θ_m - the stationarity and invertibility constraints, respectively. Mathematically, these constraints are defined by imposing the following condition: all roots of the AR and MA characteristic polynomials - $\Phi^p(\mathcal{B})$ and $\Theta^q(\mathcal{B})$, must lie outside the unit circle. Therefore an ARIMA(p, d, q) process is stationary and invertible if the absolute value of all the roots of $\Phi^p(\mathcal{B})$ and $\Theta^q(\mathcal{B})$ are greater than one.

The stationarity constraint on the AR weights ensures that the modelled time series exhibits a stable and mean-reverting behaviour. It allows for the influence of past values to eventually decay with time preventing the series from diverging to infinity. This can be demonstrated for a simple AR(1) model, for which the characteristic polynomial is

$$\Phi^1(\mathcal{B}) = (1 - \mathcal{B}\phi_1) \tag{9}$$

This polynomial has the root:

$$\mathcal{B} = 1/\phi_1 \tag{10}$$

Imposing the stationarity constraint on this root, we get:

$$|\phi_1| < 1 \tag{11}$$

The t^{th} observation of an AR(1) time-series can be expressed indefinitely in terms of the lagged values (suppressing the constant term c for simplicity):

$$y_t = \phi_1 y_{t-1} + \epsilon_t = \phi_1^2 y_{t-2} + \phi_1\epsilon_{t-1} + \epsilon_t = \ldots \tag{12}$$

Evidently, if Condition 11 is violated, the influence of past values will grow without bound, resulting in an explosive, non-stationary time series.

The invertibility constraint is the dual to the stationarity condition. To see this, first note that any Moving Average process can be represented in terms of an infinite AR(∞) process (Brockwell & Davis

2009, Sec.3.1). The invertibility condition simply ensures that the ARIMA process admits a convergent infinite representation and subsequently, the present forecast errors ϵ_t can be expressed uniquely as a linear combination of past observations. We demonstrate this again for an MA(1) model for which the condition on the roots of the characteristic polynomial $\Theta^1(\mathcal{B})$ results in the condition on the MA weight:

$$|\theta_1| < 1 \tag{13}$$

Rewriting the forecasting equation for an MA(1) model in terms of ϵ_t:

$$\epsilon_t = y_t - \theta_1\epsilon_{t-1} = y_t + \theta_1^2\epsilon_{t-2} - \theta_1 y_{t-1} = \dots \tag{14}$$

Continuing indefinitely, we get the infinite AR(∞) representation of an MA(1) process:

$$\epsilon_t = y_t - \theta_1 y_{t-1} - \theta_2^2 y_{t-2} - \theta_3^3 y_{t-3} \dots \tag{15}$$

The above series is convergent only when Condition 13 is satisfied. Moreover, the autocorrelation at lag 1 is given by:

$$\rho_1 = \frac{\theta_1}{1+\theta_1^2} \tag{16}$$

There is a degeneracy in this autocorrelation for θ_1 and $1/\theta_1$. Therefore, the constraint $|\theta_1| < 1$ also enforces a unique representation of the MA(1) model by its parameters.

2.2 Nested Sampling

The Nested Sampling algorithm, first developed by Skilling (2006), is primarily used to compute the evidence term in Bayesian inference. Given a dataset D, and a model $\mathcal{M}$ characterized by parameters θ_i, Bayesian inference updates the prior distribution on the model parameters $P(\theta_i|\mathcal{M})$ to obtain the posterior distribution $P(\theta_i|D;\mathcal{M})$ using the Bayes theorem:

$$P(\theta_i|D;\mathcal{M}) = \frac{P(D|\theta_i;\mathcal{M})P(\theta_i|\mathcal{M})}{P(D|\mathcal{M})} = \frac{\mathcal{L}(D|\theta_i)\pi(\theta_i)}{Z} \tag{17}$$

Here $\mathcal{L}(D|\theta_i)$ is the Likelihood function and the normalization constant Z is the Bayesian evidence:

$$Z = \int \mathcal{L}(D|\theta_i)\pi(\theta_i)d\theta_i \tag{18}$$

It is generally intractable to analytically evaluate this integral, especially in higher dimensional spaces. Nested Sampling is a numerical algorithm which transforms this multi-dimensional evidence integral into a one dimensional problem by introducing the prior volume - the amount of fractional prior mass contained within an iso-likelihood contour of $\mathcal{L}'$ in the prior space:

$$X(\mathcal{L}') = \int_{\mathcal{L}>\mathcal{L}'} \pi(\theta)d\theta \tag{19}$$

In terms of the prior volume, the evidence integral then becomes

$$Z(X) = \int_0^1 \mathcal{L}(X)dX \tag{20}$$

The Nested Sampling algorithm evaluates this one-dimensional integral by iteratively sampling points from the prior constrained to successively higher likelihood contours, hence compressing the prior volume at each step. The integral in Equation 20 is evaluated numerically as the weighted sum:

$$Z = \sum_i w_i\mathcal{L}_i \tag{21}$$

The weights w_i correspond to the shrinkage in prior volume ΔX_i between two successive iterations:

$$w_i = \Delta X_i = X_{i-1} - X_i \tag{22}$$

For a detailed description of the algorithm and its practical implementation, readers are referred to Ashton et al. (2022) and Buchner (2023).

2.2.1 *Sampler Output and Details*

The algorithm terminates when a user-defined convergence criterion is met. A standard practice is to define a threshold value for the maximum possible contribution of evidence from the remaining live points $Z_{\rm live}$ to the existing evidence estimate Z (Prathaban et al. 2025b). The convergence criterion is then defined such that the algorithm terminates when $Z_{\rm live}$ is smaller than this threshold. This threshold is generally expressed in terms of some fractional value of the existing evidence estimate Z.

To avoid numerical overflow, the calculations and results are typically computed in logarithmic space. The completed sampling run therefore yields an estimate of the log evidence $\log Z$ for the given model along with an estimate of the uncertainty $\sigma_{\log Z}$ associated with it. The uncertainty in the log-evidence is given by:

$$\sigma_{\log Z} = \sqrt{\frac{\mathcal{D}_{\rm KL}}{n_{\rm live}}} \tag{23}$$

$\mathcal{D}_{\rm KL}$ is the Kullback-Leibler divergence (Kullback & Leibler 1951) which represents the amount of concentration of the posterior relative to the prior:

$$\mathcal{D}_{\rm KL} = \int P(\theta|D)\log\left(\frac{P(\theta|D)}{\pi(\theta)}\right) \tag{24}$$

$n_{\rm live}$ is the number of live points sampled from the prior. The total convergence time of the nested sampling run scales directly with $\mathcal{D}_{\rm KL}$ and $n_{\rm live}$. For a fixed uncertainty $\sigma_{\log Z}$, $n_{\rm live}$ is directly proportional to $\mathcal{D}_{\rm KL}$. The number of live points $n_{\rm live}$ is therefore chosen to optimize precision and time for convergence.

The sampler also yields a chain of nested posterior samples with corresponding weights p_i, derived from the prior volume shrinkage w_i at each i^{th} iteration using

$$p_i = \frac{\mathcal{L}_i w_i}{Z} \tag{25}$$

The posterior samples are representative of the posterior probability distribution of the parameters $P(\theta_i|D)$. By evaluating the model function $f(\theta)$ for each sample, one can construct the posterior predictive distribution $P(y|x,D)$ allowing the visualization of predictive contours and credible regions.

2.2.2 *Model Selection*

Bayesian evidences are useful in the problem of model comparison and selection (Jeffreys 1961; Kass & Raftery 1995). In particular, for a set of competing models $\mathcal{M}_i$, the evidence Z_i can also be interpreted as the conditional probability: $P(D|\mathcal{M}_i)$. Using Bayes' theorem for models, the posterior probability of a model $P(\mathcal{M}_i|D)$ can be computed:

$$P(\mathcal{M}_i|D) = \frac{P(D|\mathcal{M}_i).P(\mathcal{M}_i)}{\sum_i P(D|\mathcal{M}_i).P(\mathcal{M}_i)} \tag{26}$$

Assuming a uniform prior, $P(\mathcal{M}_i) = 1/n$ across all n competing

models, the above equation reduces to:

$$P(\mathcal{M}_i|D) = \frac{P(D|\mathcal{M}_i)}{\sum_i P(D|\mathcal{M}_i)} = \frac{Z_i}{\sum_i Z_i} \tag{27}$$

The evidence Z_i associated with any model is thus a principled measure of the goodness of fit. This in turn is related to the likelihood function $\mathcal{L}(D|\theta_i)$ as well as the size of the prior space over which this fit is carried. The latter factor ultimately results in the Occam's razor that penalizes the evidence of an overfitted model.

3 METHODOLOGY

This section outlines the methodology adopted to apply the nested sampling algorithm on ARIMA models. We discuss details of the likelihood and prior functions in subsection 3.1, the nested sampler in subsection 3.2, and finally the model selection and validation procedures in subsections 3.3 and 3.4, respectively.

The following quantities are treated as the set of parameters characterizing any given ARIMA model : the respective AR and MA weights ϕ_a and θ_m, the missing initial state (D_0 and ϵ_0), the standard deviation σ associated to the present error ϵ_t and the unconditional mean of the time series μ, which can be shown to be related to the intercept term c through:

$$c = \mu(1 - \sum_a^p \phi_a) \tag{28}$$

3.1 Likelihood and Priors

A Gaussian likelihood function is chosen for the n observations of the given time series D_t :

$$\mathcal{L}(D|\theta_i) = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(D_t - \hat{\mathbf{y}}_\mathbf{t})^2}{2\sigma^2}\right) \tag{29}$$

After fitting the ARIMA model, the resulting residuals are assumed to be independent Gaussian random variables with mean zero and constant variance (σ^2). Since the noise model enters explicitly through the likelihood function, the ARIMA forecast values $\hat{\mathbf{y}}_\mathbf{t}$, used in computing the likelihood should be completely deterministic. These deterministic forecasts $\hat{\mathbf{y}}_\mathbf{t}$ are obtained from Equation 3 by setting $\epsilon_t = 0$. The missing initial values for the time series data D_0 and the forecast errors ϵ_0 are treated as nuisance parameters in our inference.

We now discuss the choice of prior distributions for the ARIMA model parameters. For σ, a reasonable choice of prior is a Half-Normal or a Half-Cauchy distribution with scale σ'. We adopt a relatively weak prior for σ by choosing a half-normal distribution truncated from below at zero and scaled by $\sigma' = 50$. The prior choice for the unconditional mean μ depends on the time series data under consideration. For a differenced or de-trended time-series, a standard normal distribution is chosen as prior for μ. In other cases, a wide normal distribution centred around the expected long-term mean μ_0 of the time-series is used. The prior distribution for the p missing initial values of D_0 is set equivalent to the prior for μ, as we expect these values to lie close to μ for a stationary series. For the initial residuals ϵ_0, a standard normal distribution is chosen. In the case of an ARIMA$(1, d, 1)$ model, a uniform distribution bounded between -1 and 1 is the most straightforward choice of priors for the coefficients ϕ_a and θ_m. The bounds are in accordance with the stationarity and invertibility conditions discussed in subsection 2.1.1. However, for higher order ARIMA models, the constraints on the individual weights are coupled and hence the choice of prior distributions for them is non-trivial.

Table 1. Prior distributions for ARIMA model parameters.

Parameter	Prior Distribution
σ	Truncated Normal$(0, \sigma')$ with $\sigma' = 50$
μ	$\mathcal{N}(\mu_0, \tau^2)$
ϕ_p	$\mathcal{N}(0, \sigma'')$, subject to stationarity
θ_p	$\mathcal{N}(0, \sigma'')$, subject to invertibility
D_0	$\mathcal{N}(\mu_0, \tau^2)$
ϵ_0	$\mathcal{N}(0, 1)$

For this paper, we assign a mathematically constrained normal prior distribution $\mathcal{N}(0, \sigma'')$ on the ARMA weights and implement a "rejection sampling" approach. The scale of the distribution σ'' is normally set to one as we do not expect the time-series under consideration to be very strongly auto-correlated. An initial set of $10^3 \times n_{\text{live}}$ points is sampled from this distribution. The points are tested for stationarity and invertibility by calculating the roots of the AR and MA characteristic polynomials, and imposing the conditions discussed in subsection 2.1.1. The parameter points passing this test are then added to a pool of valid particles. This process is continued iteratively until the desired number of valid live particles n_{live} is reached in this pool. The normal prior distributions on the weights are also constrained to the stationary and invertible regions of the parameter space.

Table 1 summarizes the prior distributions assigned to each parameter of the ARIMA model.

3.2 Sampler Configuration

We use a vectorized formulation of the nested sampling algorithm developed by Yallup et al. (2025) within `blackjax`. To ensure optimal speed and compatibility with this sampler, we also built a fully vectorized ARIMA framework in Python using `JAX` (Bradbury et al. 2018).

The sampler is initialized by specifying the model order, prior parameters, the number of live points n_{live} and a random seed for the run. The process is parallelized in this sampler through the simultaneous deletion and repletion of a batch of n_{delete} lowest-likelihood points. This parallelization parameter is set to $n_{\text{delete}} = 50$. The `blackjax` nested sampler adopts Slice Sampling (Neal 2003) as a default algorithm for the inner MCMC kernel of the sampler. The length of the MCMC chain run k for each single replacement of the points is expressed in terms of the number of dimensions $\hat{d}$ of the parameter space. This value is conventionally set to $k = 6\hat{d}$ in our work. The convergence criterion is defined, as discussed in subsection 2.2.1, using the threshold value:

$$\frac{Z_{\text{live}}}{Z} < 10^{-3} \tag{30}$$

3.3 ARIMA Model Selection and Fitting

In this paper, we focus on evaluating the model log posterior probabilities $\log P_i$ on a grid of ARIMA models formed by the AR order p and the MA order q, and a fixed d order. To ensure that the model log posterior probabilities are comparable across candidate

ARIMA specifications, each model must be formulated as a complete generative description of the same observed time series. In particular, any two ARIMA models with the same p and q orders but different d orders should be implemented without any external pre-differencing to the data. In Section 4, we demonstrate model selection across the d order for a simulated ARMA process with a trend. In all other cases, for simplicity, we adopt first order differencing if a trend is apparent. The standard stationarity checks prescribed in the Box-Jenkins methodology are also used. In particular, statistical unit root tests such as the Augmented Dickey Fuller (ADF) (Dickey & Fuller 1979) and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) (Kwiatkowski et al. 1992) Tests are used to confirm stationarity.

A grid search using $n_{\text{live}} < 500$ is first performed for identifying the best models. The results are visualized by plotting a heatmap of the calculated model logarithmic posterior probabilities $\log P_i$ on the ARIMA grid, similar to Yu (2023, Fig.7). From the heatmap, the model with the highest log posterior probability is picked for a second, high resolution nested sampling run.

Further analysis is performed by dividing the time series data into training and observed windows. The model with the highest log posterior probabilities is fit to the training dataset using a high resolution nested sampling run ($n_{\text{live}} = 500$ to 1000). Weighted posterior samples are obtained from this run which are then used to visualize the model fit and residuals. For this, we use `fgivenx` (Handley 2018) - a python package for plotting posterior line plots and predictive posteriors of functions. Using the weighted posterior samples, the posterior predictive forecasts along with their 1σ, 2σ and 3σ credible regions are plotted.

3.4 Model Validation

The grid-search results are validated on the broad bases of residual analysis and forecasting accuracy. Practically, almost all ARIMA models are able to model the training window of any given time-series by making one-step ahead, in-sample predictions. On the other hand, a more accurate metric for model validation is comparing the results of out-of-sample multi-step forecasts.

To confirm that the selected best-fit model has correctly captured the variability of the training dataset, we test the residuals from this fit for normality and autocorrelation. This is done by plotting residual histograms and autocorrelation plots along with performing the Ljung-Box statistical test (Ljung & Box 1978) for autocorrelation. The multi-step direct forecasts for sunspots number data are also tested for accuracy by evaluating the Mean Square Errors (MSE), Root Mean Square Errors (RMSE) and Mean Absolute Errors (MAE). In order to facilitate comparison with a traditional ARIMA model selection method, the Bayesian Information Criterion (BIC) heatmap is also plotted on the ARIMA grid using the ARIMA model functionality from `statsmodels`.

4 RESULTS

This section presents the results of applying the model selection procedure to simulated and real astronomical time series data.

4.1 Simulated Time Series

The framework was first tested on a simulated AR(2) time series (Figure 1). The grid search method correctly revealed ARIMA$(2,0,0)$ as the model with the highest logarithmic posterior probability of $\log P_{\text{max}} = -1.2 \pm 0.5$ (Figure 2). It suffices to compare the posteriors of the parameters to their true values for the validation of our method in this case. A high resolution nested sampling run of the ARIMA$(2,0,0)$ model for the time series was able to produce posteriors samples close to the true values of the parameters (Figure 3).

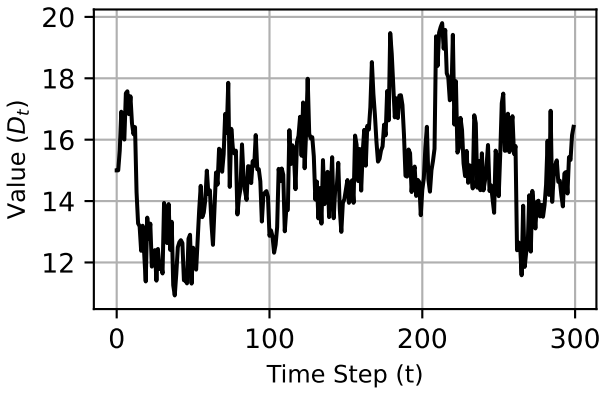

Figure 1. Artificially generated AR(2) time-series of 300 data points with $\phi_1 = 0.6$ and $\phi_2 = 0.3$. A constant intercept term of $c = 1.5$ and a standard deviation of $\sigma = 1.0$ associated to ϵ_t was used.

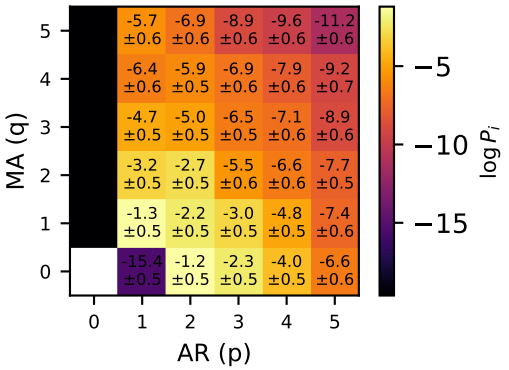

Figure 2. Heatmap of the model log posterior probabilities P_i for simulated AR(2) time-series (Figure 1). A hot-spot is observed at ARIMA$(2,0,0)$. The log posterior probabilities level off for higher orders as expected due to the action of Occam's penalty factor.

To demonstrate model selection across different d orders of the ARIMA model, we test our framework on a simulated ARMA$(1,1)$ process with a linear trend (Figure 4). The data was generated using the equation

$$D(t) = c + \beta t + \phi_1 y_{t-1} + \theta_1 \epsilon_{t-1} + \epsilon_t \tag{31}$$

with the βt term producing the linear trend. Nested sampling runs were initiated on this data for ARIMA$(1,d,1)$ models, with d ranging from 0 to 4. The calculated model log posterior probabilities $\log P_i$ along with their uncertainties are outlined in Table 2. The posterior probability is highest for ARIMA$(1,1,1)$ as expected, since first order of differencing is required to eliminate the linear trend in the data.

4.2 Sunspots Number Data

Sunspots are darker and cooler spots observed on the solar photosphere. The evolution of the number of sunspots during the sunspot

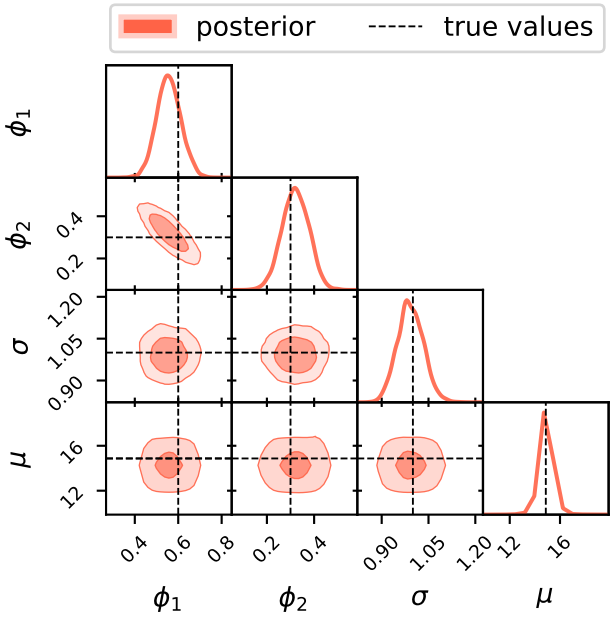

Figure 3. Posterior distributions of the AR(2) model parameters inferred from the simulated AR(2) time-series (Figure 1). The 1-D kernel density estimates show well-constrained posterior densities centred near the true parameter values (indicated by the black dashed lines), thus demonstrating good recovery of the underlying process dynamics.

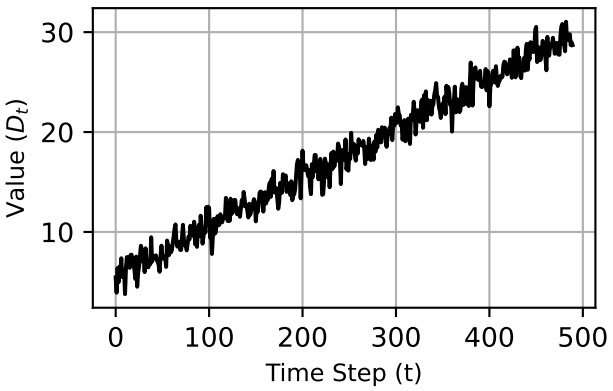

Figure 4. Artificially generated ARMA(1, 1) process of 490 data points with a linear trend. The ARMA coefficients are chosen to be $\phi_1 = 0.6$ and $\theta = -0.4$. The constant intercept term and standard deviation are $c = 2$ and $\sigma = 1$, respectively.

Table 2. Model log-posterior probabilities $\log P_i$ and their uncertainties σ for ARIMA(1, d, 1) fit to the data in Figure 4

d	$\log P_i$	$\sigma_{\log P_i}$
0	−19.79499	0.20408
1	−0.01814	0.25429
2	−4.03641	0.20032
3	−8.12772	0.21359
4	−10.83405	0.24954

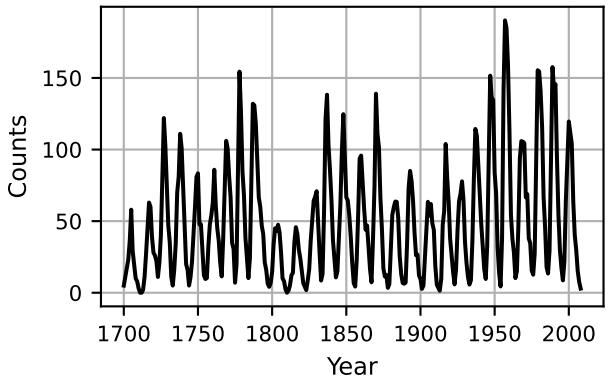

Figure 5. Yearly Sunspots Number Data from 1700 to 2008

Table 3. Results of the Stationarity Tests (ADF and KPSS) on yearly sunspots number data.

Statistic	**ADF Test**	**KPSS Test**
Test Statistic	-2.931083	0.124768
p-value	0.041851	0.100000
Lags Used	8	7
Number of Observations Used	246	247
Critical Value (1%)	-3.457215	0.739000
Critical Value (5%)	-2.873362	0.463000
Critical Value (10%)	-2.573070	0.347000

cycle is found to be directly associated with the solar activity (Hathaway 2015). It is therefore essential to analyse and forecast the sunspots number for predicting space weather and mitigating its impact on Earth.

We use the yearly sunspots number data from the year of 1700 to 2008 for our analysis. The first 255 data points corresponding to the years from 1700 to 1954 are used as the training dataset. The results of ADF and KPSS stationarity tests, on this training data, are outlined in Table 3. From the p-value and the critical value at 5% significance level of both tests, it can be concluded that the time-series data is stationary. Therefore, no differencing is required and the d order is set to zero for the ARIMA grid search. A grid search was carried out for ARIMA orders up-to ten (Figure 6), which selected ARIMA(9, 0, 1) model with the highest log posterior probability of $\log P_{\max} = -1.3 \pm 0.5$. On the contrary, ARIMA(3, 0, 3) is chosen as the best-fit model on the basis of the Bayesian Information Criterion (BIC), as shown in Figure 7. The maximum likelihood estimation method adopted by the ARIMA model function in `statsmodels` failed to converge for multiple higher order ARIMA models.

ARIMA(9, 0, 1) is chosen for fitting the data using a high resolution ($n_{\text{live}} = 1000$) nested sampling run. The marginalized posterior distributions of ARIMA parameters obtained from this fit are shown in Figure 8. The model fit and residuals are shown in Figure 9. The model occasionally predicts a negative sunspots number count due to the assumption of normally distributed errors $\epsilon_t \sim \mathcal{N}(0, \sigma')$. We regard these predictions as unphysical. It can be seen that the residual sequence symmetrically fluctuates around zero. A histogram of the pooled residuals from different posterior samples is shown in Figure 10, depicting the residuals to be almost normally distributed. We use time-ordered residuals of the fit, obtained from the posterior means of the parameters, to further check for any autocorrelation in the residu-

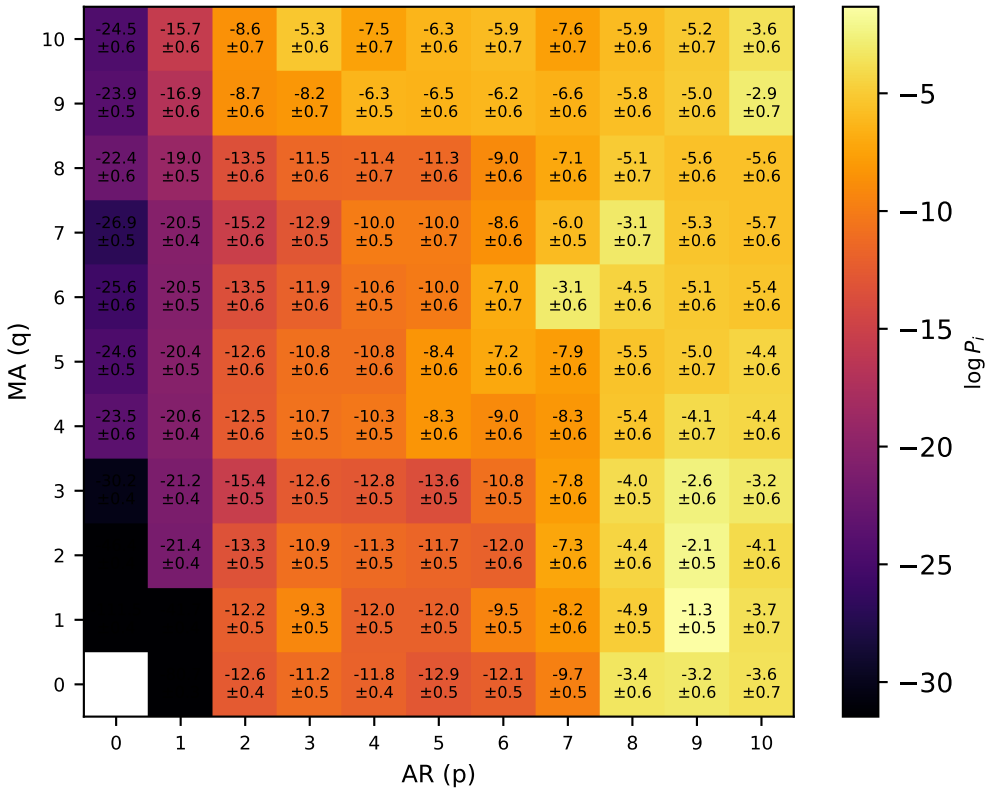

Figure 6. Heatmap of the model log posterior probabilities obtained from the nested sampling runs on yearly sunspots number data (Figure 5). The grid is annotated with model log posterior probabilities log P_i and their estimated uncertainties $\sigma_{\log P_i}$ for reference. A clear statistical preference for higher ARIMA orders, and particularly for a higher AR order p can be seen, with ARIMA$(9, 0, 1)$ having the highest model posterior probability. .

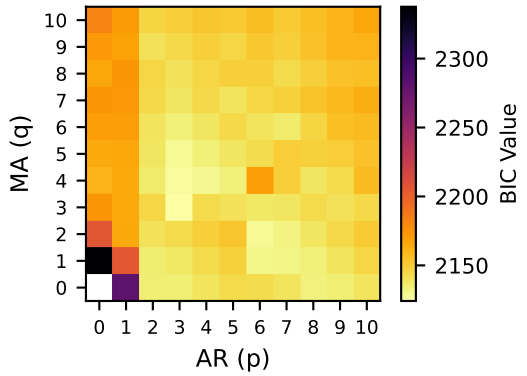

Figure 7. Heatmap of the Bayesian Information Criterion (BIC) values for ARIMA model fit on yearly sunspots number data. The lowest value of BIC is observed for ARIMA$(3, 0, 3)$. There is no clear preference seen for higher ARIMA orders suggesting that this method of model selection has penalized complex models more strongly than nested sampling.

Table 4. Ljung–Box(LB) test p-values for residual autocorrelation of sunspots number data.

Lag	p-value
1	0.841791
2	0.978972
3	0.728602
4	0.819598
5	0.905064
6	0.785299
7	0.624773
8	0.655938
9	0.746816
10	0.535520

als. The Autocorrelation function (ACF) and Partial Autocorrelation function (PACF) plots (Figure 11) suggest no significant autocorrelation in the residual time-series. This is further confirmed from the results of the Ljung-Box test, up to lag 10, on the time-ordered residuals (Table 4).

The results show that all p-values exceed 0.05 across lags 1-10. Hence, we fail to reject the null hypothesis of no serial correlation in the residuals at the 5% significance level. The residuals are therefore characteristic of a white noise sequence, indicating that the ARIMA$(9, 0, 1)$ fit have effectively captured the temporal structure of our training data.

Finally, we show in Figure 12, the results of direct multi-step forecasting of the sunspots number data from the years of 1954 to 2008, using an ARIMA$(9, 0, 1)$ fit. In Appendix A, the forecasting performance of ARIMA$(9, 0, 1)$ is compared to ARIMA$(3, 0, 3)$. Although the latter model is favoured by the Bayesian Information Criterion (BIC) (Figure 7), it yields a significantly lower model log posterior probability compared to ARIMA$(9, 0, 1)$.

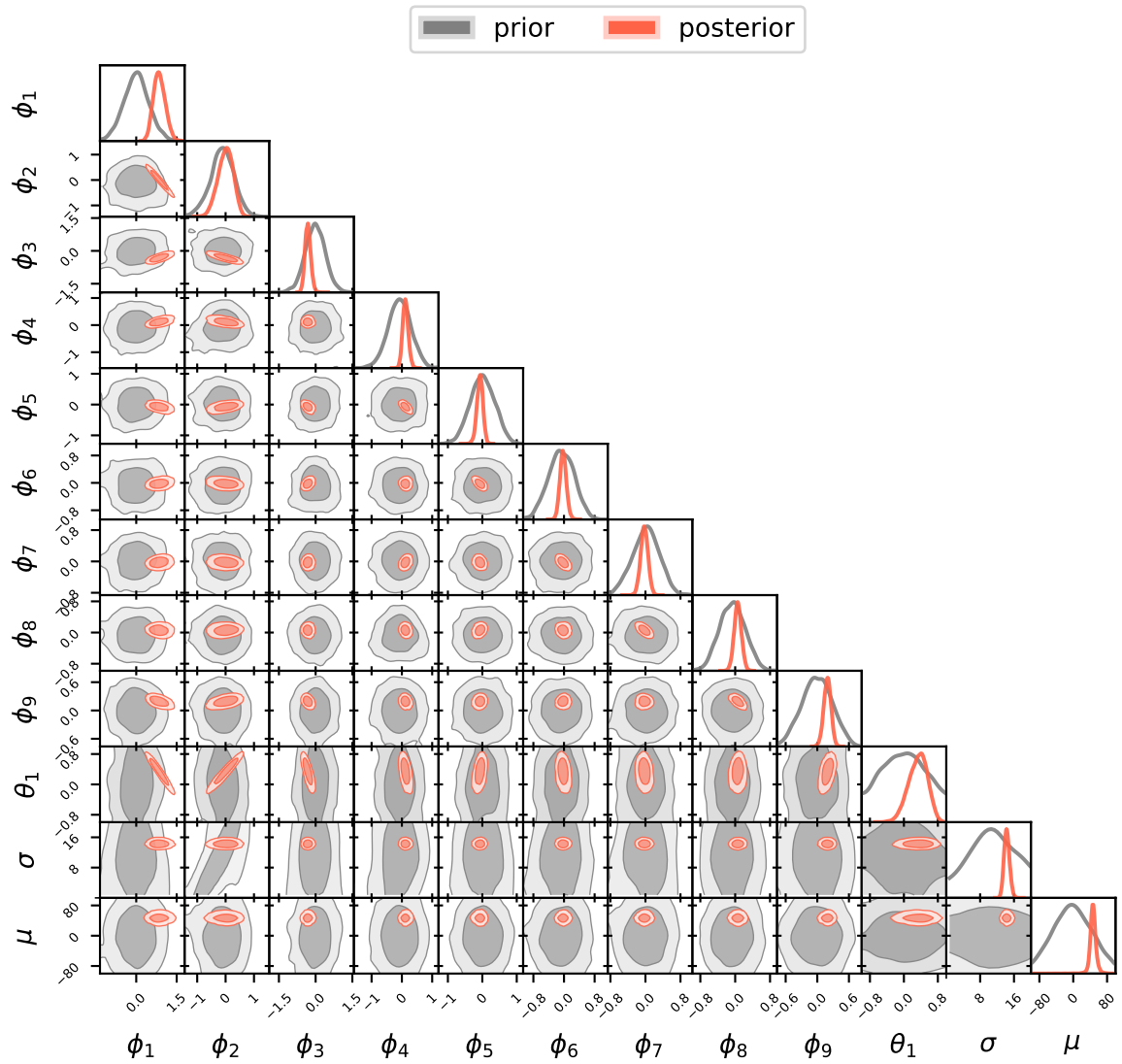

Figure 8. Posterior corner plot of ARIMA(9, 0, 1) model parameters obtained from the nested sampling run on the yearly sunspots number data (Figure 5). Orange curves represent the marginalized posterior distributions, while the gray curves show the corresponding prior distributions (Table 1) adopted for each parameter. The sharp posterior peaks indicate well-constrained AR and MA coefficients, with most of the posterior mass concentrated within the stationary and invertible regions of parameter space. The information gain between prior and posterior, measured by the Kullback–Leibler divergence, is $D_{KL} = 19$ nats, demonstrating a substantial concentration of probability mass due to the data.

4.3 Kepler and TESS lightcurves

The Kepler and TESS (Transiting Exoplanet Survey Satellite) photometric lightcurves act as ideal datasets for testing our ARIMA-Nested Sampling framework owing to their long and regular cadences with minimal systematics noise. In this subsection, we present the results of applying our model selection methodology to photometric lightcurves of two targets observed by Kepler and TESS - KIC 12008916 and Ross 176, respectively.

KIC 12008916

KIC 12008916 is a low-luminosity red giant in the original Kepler field and is one of the benchmark stars in the Kepler astroseismic sample (Davies & Miglio 2016). The star has a well-resolved, solar-like oscillation spectrum. The lightcurve exhibits stochastic variability which is stable and stationary over longer timescales, providing an ideal testbed for ARIMA modelling.

We use lightcurve data from the Kepler Quarter 00, corresponding to the year of 2009 (Figure 13). The Pre-search Data Conditioning Search Aperture Photometry flux (PDCSAP) flux was used for this analysis, which was found to be significantly autocorrelated, as seen from the ACF/PACF plots in the top panel of Figure 15. The results of the ADF and KPSS stationarity tests confirm the time series data to be stationary. An ARIMA grid search, with $d = 0$ fixed, revealed ARIMA$(0, 0, 1)$ as the best fit model ($\log P_{\max} = -0.0014 \pm 0.76$) despite the significant autocorrelation present beyond first lag in the data. The results of the grid search are shown in Figure 14.

The ARIMA$(0, 0, 1)$ model was fit to the data using a high resolution ($n_{\text{live}} = 500$) nested sampling run and the means of the posterior samples were used to obtain time-ordered residuals. We observe no significant autocorrelation in these time-ordered resid-

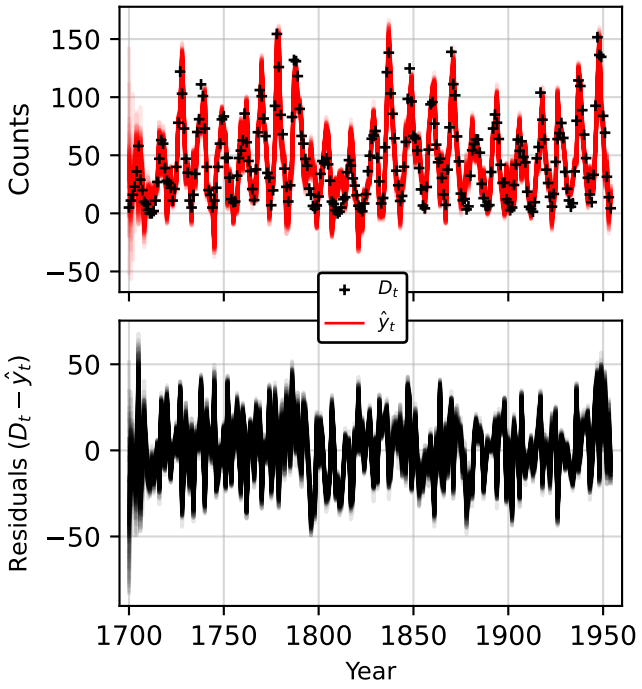

Figure 9. Results of ARIMA$(9, 0, 1)$ model fit to the training window, corresponding to the years of from 1700 to 1954, of the yearly sunspots number data (Figure 5). The upper panel shows the observed sunspot counts (in black) and the model fit curves (in red) obtained using 500 weighted posterior samples of the ARIMA parameters, while the lower panel displays the corresponding residuals. The residuals fluctuate around zero with no strong temporal structure, indicating that the fitted model captures most of the systematic variability in the data.

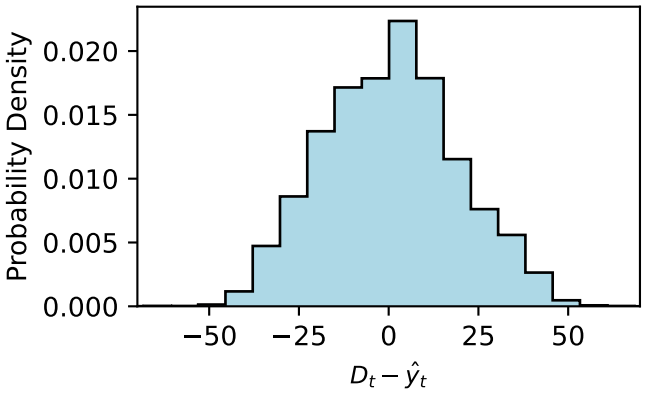

Figure 10. A histogram of the pooled residuals from the ARIMA$(9, 0, 1)$ fit in Figure 9 to yearly sunspots number data. The residuals appear to be normally distributed around 0.

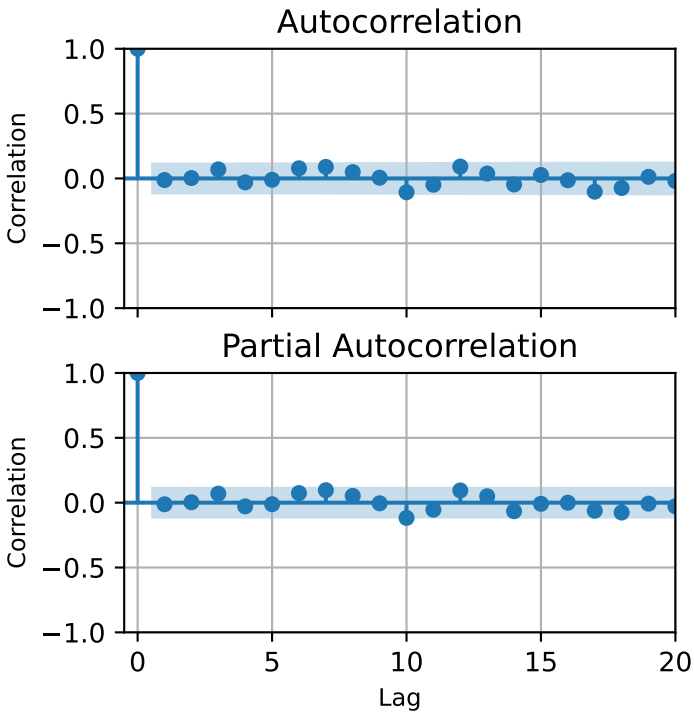

Figure 11. Autocorrelation (top) and Partial Autocorrelation (bottom) function plots of the mean residuals from the ARIMA$(9, 0, 1)$ fit to the yearly sunspots number data (Figure 9). Both functions lie within the 95% confidence bounds across all lags, indicating that the residuals are effectively uncorrelated.

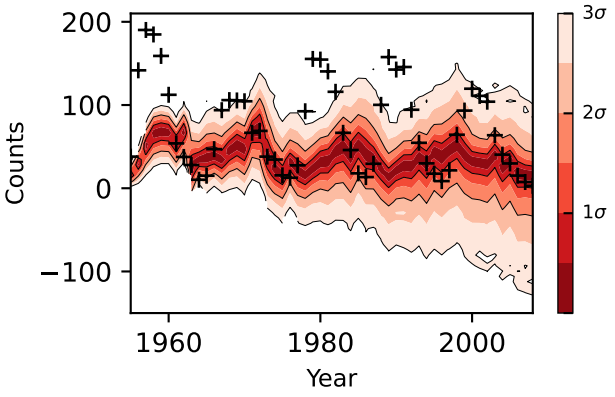

Figure 12. Posterior predictive forecasts of the yearly sunspots number for the years from 1954 to 2008. The forecasts are obtained using 5000 weighted posterior samples from the ARIMA$(9, 0, 1)$ fit (Figure 9). Shaded contours denote the 1σ, 2σ and 3σ credible regions of the predictive posterior distribution $P(\hat{y}_t | t, D_t)$.

uals, as shown by their ACF/PACF plots from the bottom panel of Figure 15. The ARIMA$(0, 0, 1)$ model has therefore successfully captured the stochastic variability in the original lightcurve.

Ross 176

The detection of transiting exoplanets is sometimes impeded due to the presence of autocorrelated noise in the lightcurves (Pont et al. 2006). This autocorrelated variability may arise from activity of the host star, instrumental artefacts, or other unexplained effects. Autoregressive modelling in such cases is effective in removing the correlated noise and enhancing the signal-to-noise ratio of transit detections (Melton et al. 2024). We demonstrate the results of applying the ARIMA-Nested Sampling framework to the photometric lightcurve data of Ross 176 - a late K-type main-sequence star. The star was recently confirmed to host a transiting super-earth exoplanet

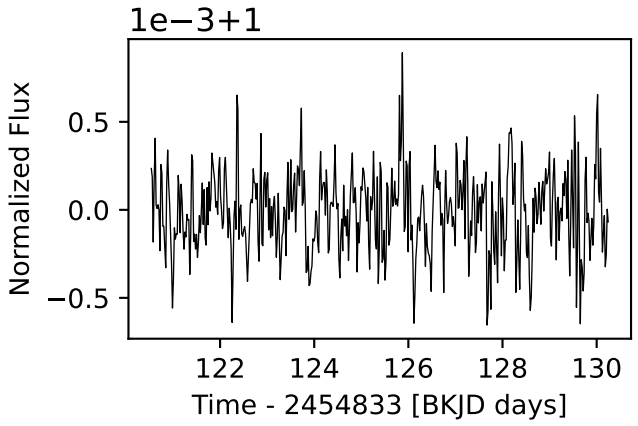

Figure 13. Kepler long-cadence light curve of KIC 12008916, normalized and shown over a 10-day segment. The star exhibits the characteristic stochastic, solar-like oscillations of a low-luminosity red giant.

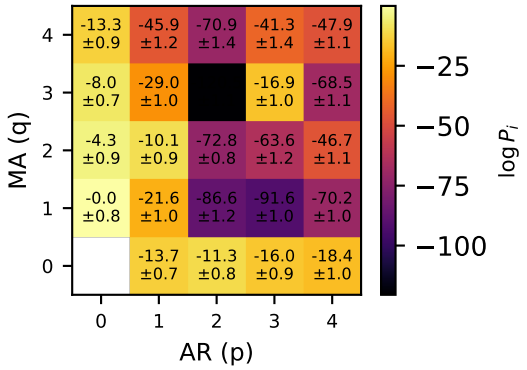

Figure 14. Heatmap of the ARIMA models' log posterior probabilities obtained from the nested sampling runs on KIC 12008916 data (Figure 13).

- Ross 176b (Geraldía-González, S. et al. 2025), with a period of approximately 5 days. For this analysis, we used the PDCSAP flux of the TESS photometric lightcurve, corresponding to the sector 83 and year 2024 (Figure 16). The lightcurve was obtained using the MIT quick-look pipeline (Huang et al. 2020a,b).

The data was partitioned into five training datasets corresponding to the presence of five significant gaps in the total time series data. The fourth and fifth training datasets exhibit an obvious trend. This was further confirmed by the ADF and KPSS tests performed on these datasets, which indicated non-stationarity. They are therefore subjected to first order differencing ($d = 1$) before further analysis. All datasets showed some level of autocorrelation, with datasets (1),(4) and (5) depicting significant autocorrelation, even after differencing (see Figure 17).

We performed an ARIMA grid search on the five training datasets, limiting the maximum ARMA orders to $p = q = 4$. For the last two training datasets, the sampler failed to converge for higher orders. We present in Table 5 , the model log posterior probabilities for these datasets calculated up to the maximum possible ARMA orders before the sampler failed to converge. From the results of ARIMA grid search and Table 5, the best fit ARIMA models for the training datasets (1), (2), (3), (4) and (5) were

Table 5. Log posterior probabilities and their associated uncertainties for ARIMA$(p, 1, q)$ model fit to training datasets (4) and (5) in Figure 16.

(p, 1, q)	Dataset (4)		Dataset (5)	
	$\log P_i$	$\sigma_{\log P_i}$	$\log P_i$	$\sigma_{\log P_i}$
(0,1,1)	-7.7228×10^{-5}	0.7672	-1.7548×10^{-6}	0.7179
(0,1,2)	-9.4965×10^{0}	0.8216	-1.4139×10^{1}	0.8151
(0,1,3)	-1.3768×10^{1}	0.8518	-1.5256×10^{1}	0.9782
(0,1,4)	-1.7617×10^{1}	0.8893	-1.7444×10^{1}	0.9257
(1,1,0)	-4.7834×10^{2}	0.6988	-5.3737×10^{2}	0.7782
(1,1,1)	-1.7541×10^{1}	0.9768	-2.2808×10^{1}	0.9227
(1,1,2)	-2.2764×10^{1}	0.9711	-3.7526×10^{1}	0.9353
(1,1,3)	-1.6884×10^{1}	1.0118	-2.1302×10^{1}	1.0024
(1,1,4)	-3.5210×10^{1}	0.9806	-2.1779×10^{1}	0.9704
(2,1,0)	-3.3926×10^{2}	0.8071	-3.4895×10^{2}	0.8189
(2,1,1)	-1.5017×10^{1}	0.8937	-1.4081×10^{1}	0.9881
(2,1,2)	-1.4214×10^{1}	0.9279	-9.4836×10^{1}	1.2746
(2,1,3)	-4.5073×10^{1}	1.2798	-3.4356×10^{1}	0.9419
(2,1,4)	-3.4459×10^{1}	0.7911	—	—
(3,1,0)	-2.8067×10^{2}	0.8213	—	—
(3,1,1)	-3.1983×10^{1}	1.0630	—	—
(3,1,2)	-3.9954×10^{2}	1.0939	—	—
(3,1,3)	-4.7549×10^{1}	1.0840	—	—

identified to be ARIMA$(1, 0, 2)$, ARIMA$(1, 0, 0)$, ARIMA$(2, 0, 2)$, ARIMA$(0, 0, 1)$ and ARIMA$(0, 0, 1)$, respectively. These models were fit to the training dataset using a high resolution nested sampling run. The posterior means were used to generate forecasts for the training datasets and obtain the time ordered residuals. The right panel (b) of Figure 17 show the autocorrelation and partial autocorrelation plots of the residuals for datasets (1), (4) and (5). The lack of significant autocorrelation in these residuals, as seen from the plots, show that the best fit ARIMA models have successfully captured the variability of the datasets.

5 CONCLUSIONS

In this paper, we presented a novel method for ARIMA model selection using a fully vectorized nested sampling algorithm. The framework was first validated on an artificial AR(2) and ARIMA$(1, 1, 1)$ process, where it successfully recovered the true underlying model order, and subsequently, posteriors centred on the true values of the model parameters. Its application to sunspots number data indicated preference for higher order ARIMA models. This can be explained by the long-memory behaviour and variability across multiple timescales of the solar magnetic activity. These additional autocorrelation structures cannot be captured by low-order models, and the Bayesian evidence therefore justifies the increased complexity. This highlights the ability of our Nested Sampling framework to identify genuinely multi-scale stochastic processes without overfitting. The data was fit to an ARIMA$(9, 0, 1)$ model which yielded accurate multi-step future forecasts using the predictive posterior functions. Finally, the framework was applied for analysing photometric lightcurves of KIC 12008916 and Ross 176 using Kepler and TESS data, respectively. The ARIMA models, selected from the nested sampling runs, were able to completely capture the autocorrelation present in both datasets.

A significant computational bottleneck in this work was sampling valid initial points for higher order ARMA coefficients ϕ_a and θ_m using the process discussed in Subsection 3.1. If very strong autocor-

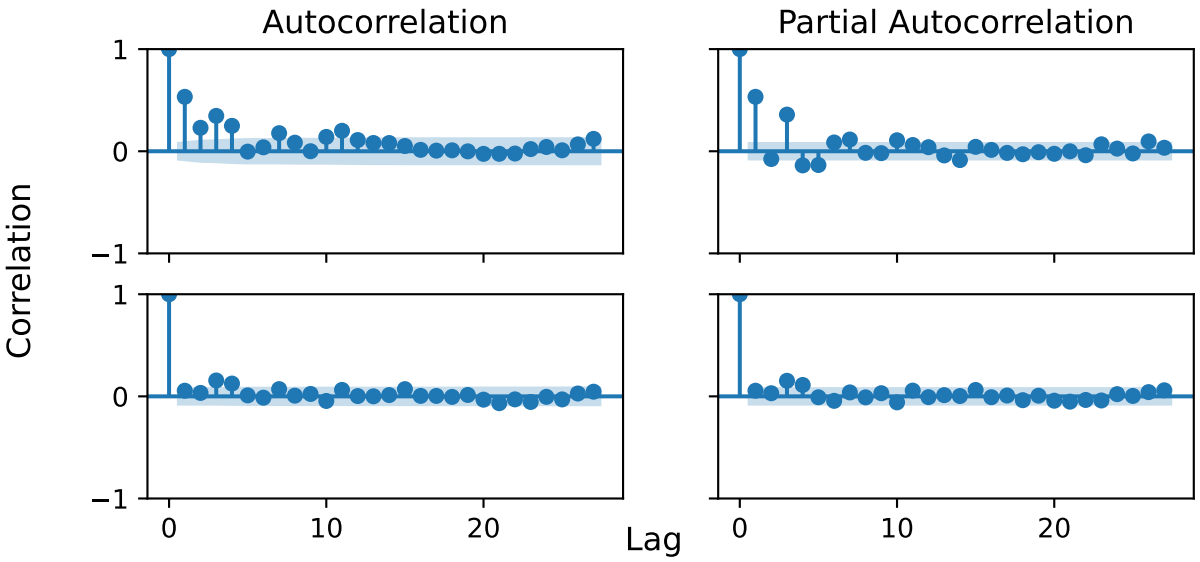

Figure 15. Autocorrelation and Partial autocorrelation plots for the lightcurve of KIC 12008916 (top) and the residuals (bottom) obtained after subtracting the best fit ARIMA$(0, 0, 1)$ model from the lightcurve data. The model fit has been able to capture most of the autocorrelation in the lightcurve data.

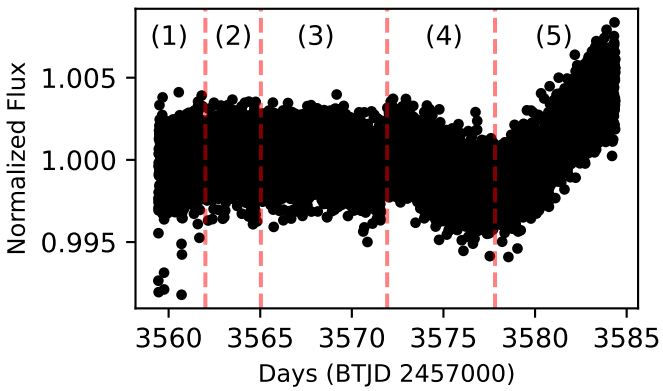

Figure 16. Normalized TESS photometric lightcurve of Ross 176 from Sector-83, processed using the MIT Quick Look Pipeline (QLP). Dashed red lines indicate gaps in the data where we perform the partition into the corresponding training datasets and label them numerically.

relation in the time-series is not expected *a priori*, then the "rejection sampling" process can be sped up by regularizing and constricting the priors using $\sigma'' < 1$ (Table 1). A more powerful approach is offered by reparametrization of the ARMA(p, q) coefficients to a set of $p + q$ partial autocorrelation coefficients (Barndorff-Nielsen & Schou 1973). These coefficients are naturally constrained between -1 and 1, and therefore a simple uniform prior, bounded between these two values, can be implemented.

We note that future improvements to the multi-step forecasts could be made by adopting a rolling window, which is a more accurate (although computationally expensive in our case) way of using ARIMA forecasting (Ndungi & Stanislavovich 2025). Additionally, a potential forecasting refinement within a Bayesian framework involves a short time-step rolling forecast and using the obtained posterior distributions as priors for the next forecast iteration. Our analysis in this work, though limited to classic ARIMA models, can also be effectively applied to its hybrid extensions, such as - Seasonal ARIMA (SARIMA), Seasonal ARIMA with exogenous variables (SARIMAX), Continous ARIMA (CARIMA) and Autoregressive Fractionally Integrated Moving Average (ARFIMA) models (Feigelson et al. 2018). These extensions can be more effective in modelling both periodic and stochastic variability exhibited in many astronomical time series, such as those of transiting exoplanets, eclipsing binaries or Active Galactic Nuclei (AGN).

The investigation of potential physical interpretations behind the use of autoregressive modelling in time domain astronomy forms another compelling future direction of work. This question could be explored by applying our model selection procedure to different time-varying astronomical systems. If systems of a particular class tend to favour similar model structures then the ARIMA model may be representative of an underlying physical process rather than being mere statistical descriptions of the data.

ACKNOWLEDGEMENTS

This material is based upon work supported by the Google Cloud research credits program, with the award GCP397499138. The authors were supported by the research environment and infrastructure of the Handley Lab at the University of Cambridge. AN was funded through the Cambridge Mathematics Placement (CMP) programme and the Institute of Astronomy (IoA) summer research programme at the University of Cambridge. WH was supported by a Royal Society University Research Fellowship.

DATA AVAILABILITY

The sunspots time series data used in this work were obtained from the *sunspots* dataset provided in the `statsmodels` library (credit: SIDC, RWC Belgium, World Data Center for the Sunspot Index, Royal Observatory of Belgium, 1700-2008). The Kepler and TESS photometric lightcurves were obtained and processed using the `lightkurve` (Lightkurve Collaboration et al. 2018) Python package.

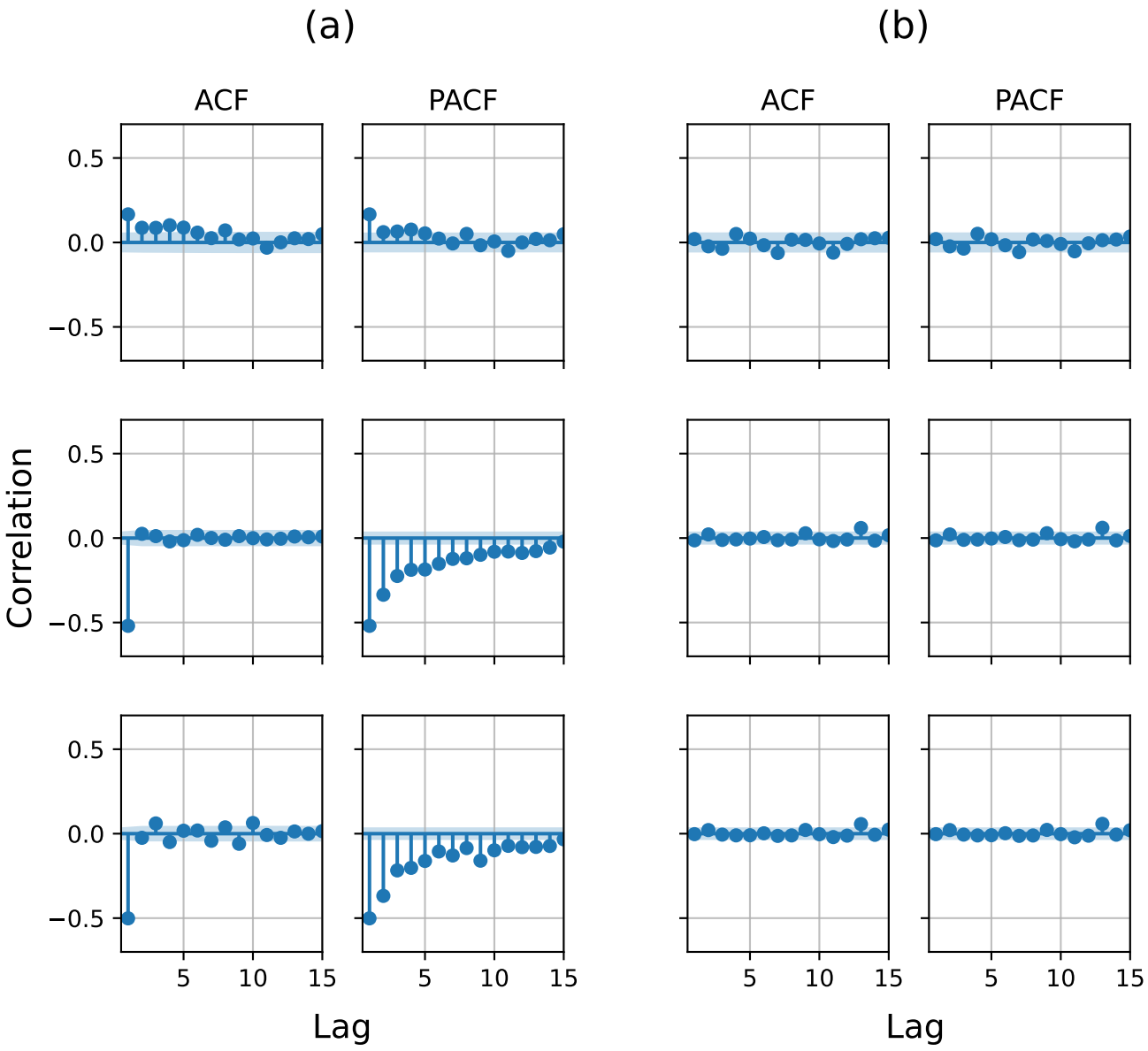

Figure 17. Autocorrelation function (ACF) and Partial Autocorrelation function (PACF) plots for the training datasets (1), (4), and (5) (from top to bottom) of Ross 176 (see Figure 16) in the left column (a), and for the residuals, obtained from the best fit ARIMA model, in the right column (b). The lightcurves exhibit strong autocorrelation, which is effectively removed using ARIMA modelling, as demonstrated by the near-white ACF/PACF structure in the residuals.

All the data used in this analysis, including the relevant nested sampling chains, is available at (Naik 2025). We also include a python notebook to generate the plots and figures presented in this paper.

REFERENCES

Akaike H., 1974, IEEE Transactions on Automatic Control, 19, 716

Akhter M., Hassan D., Abbas S., 2020, Astronomy and Computing, 32, 100403

Ashton G., et al., 2022, Nature Reviews Methods Primers, 2, 39

Barndorff-Nielsen O., Schou G., 1973, Journal of Multivariate Analysis, 3, 408

Baron D., 2019, Machine Learning in Astronomy: a practical overview (arXiv:1904.07248), https://arxiv.org/abs/1904.07248

Box G. E. P., Jenkins G. M., 1976, Time Series Analysis: Forecasting and Control, 2nd edn. Holden-Day, San Francisco

Bradbury J., et al., 2018, JAX: composable transformations of Python+NumPy programs, http://github.com/google/jax

Brockwell P. J., Davis R. A., 2009, Time Series: Theory and Methods, 3rd edn. Springer Series in Statistics, Springer, New York

Buchner J., 2023, Statistics Surveys, 17, 169

Carruba V., Aljbaae S., 2021, in European Planetary Science Congress. pp EPSC2021–36, doi:10.5194/epsc2021-36

Davies G. R., Miglio A., 2016, Astronomische Nachrichten, 337, 774

Dickey D. A., Fuller W. A., 1979, Journal of the American Statistical Association, 74, 427

Elorrieta, Felipe Eyheramendy, Susana Palma, Wilfredo 2019, A&A, 627, A120

Feigelson E. D., Babu G. J., Caceres G. A., 2018, Frontiers in Physics, Volume 6 - 2018

Foreman-Mackey D., Agol E., Ambikasaran S., Angus R., 2017, The Astronomical Journal, 154, 220
Gaia Collaboration Prusti T., de Bruijne J. H. J., Brown A. G. A., Vallenari A., Babusiaux C., Bailer-Jones C. A. L., et al. 2016, Astronomy & Astrophysics, 595, A1
Geraldía-González, S. et al., 2025, A&A, 700, A216
Handley W., 2018, The Journal of Open Source Software, 3
Hathaway D. H., 2015, Living Reviews in Solar Physics, 12, 4
Huang C. X., et al., 2020a, Research Notes of the AAS, 4, 204
Huang C. X., et al., 2020b, Research Notes of the AAS, 4, 206
Ivezić Ž., et al., 2019, ApJ, 873, 111
Jeffreys H., 1961, Theory of Probability, 3rd edn. Oxford University Press, Oxford
Johnson S. A., Penny M. T., Gaudi B. S., Kerins E., Rattenbury N. J., Robin A. C., Calchi Novati S., Poleski R., 2020, The Astronomical Journal, 160, 184
Kass R. E., Raftery A. E., 1995, Journal of the American Statistical Association, 90, 773
Kullback S., Leibler R. A., 1951, The Annals of Mathematical Statistics, 22, 79
Kwiatkowski D., Phillips P. C., Schmidt P., Shin Y., 1992, Journal of Econometrics, 54, 159
Leeney S. A. K., Handley W. J., Bevins H. T. J., de Lera Acedo E., 2025, Bayesian Anomaly Detection for Ia Cosmology: Automating SALT3 Data Curation (arXiv:2509.13394), https://arxiv.org/abs/2509.13394
Lightkurve Collaboration et al., 2018, Lightkurve: Kepler and TESS time series analysis in Python, Astrophysics Source Code Library (ascl:1812.013)
Ljung G. M., Box G. E. P., 1978, Biometrika, 65, 297
Lomb N. R., 1976, Astrophysics and Space Science, 39, 447
Lovick T., Yallup D., Piras D., Mancini A. S., Handley W., 2025, High-Dimensional Bayesian Model Comparison in Cosmology with GPU-accelerated Nested Sampling and Neural Emulators (arXiv:2509.13307), https://arxiv.org/abs/2509.13307
Melton E. J., Feigelson E. D., Montalto M., Caceres G. A., Rosenswie A. W., Abelson C. S., 2024, The Astronomical Journal, 167, 203
Naik A., 2025, Nested Sampling for ARIMA Model Selection, doi:10.5281/zenodo.17771974, https://doi.org/10.5281/zenodo.17771974
Ndungi R., Stanislavovich L. I., 2025, in Kahraman C., Cebi S., Oztaysi B., Cevik Onar S., Tolga C., Ucal Sari I., Otay İ., eds, Intelligent and Fuzzy Systems. Springer Nature Switzerland, Cham, pp 294–302
Neal R. M., 2003, The Annals of Statistics, 31, 705
Ormondroyd A. N., Handley W. J., Hobson M. P., Lasenby A. N., Yallup D., 2025, Dynamic or Systematic? Bayesian model selection between dark energy and supernova biases (arXiv:2509.13220), https://arxiv.org/abs/2509.13220
Pont F., Zucker S., Queloz D., 2006, Monthly Notices of the Royal Astronomical Society, 373, 231
Prathaban M., Yallup D., Alvey J., Yang M., Templeton W., Handley W., 2025a, Gravitational-wave inference at GPU speed: A bilby-like nested sampling kernel within blackjax-ns (arXiv:2509.04336), https://arxiv.org/abs/2509.04336
Prathaban M., Bevins H., Handley W., 2025b, Monthly Notices of the Royal Astronomical Society, 541, 200
Richards J. W., et al., 2011, The Astrophysical Journal, 733, 10
Ricker G. R., et al., 2015, Journal of Astronomical Telescopes, Instruments, and Systems, 1, 014003
Scargle J. D., 1982, The Astrophysical Journal, 263, 835
Schwarz G., 1978, The Annals of Statistics, 6, 461
Skilling J., 2006, Bayesian Analysis, 1, 833
Spergel D., et al., 2015, Wide-Field InfrarRed Survey Telescope-Astrophysics Focused Telescope Assets WFIRST-AFTA Final Report (arXiv:1503.03757)
Trotta R., 2008, Contemporary Physics, 49, 71
Yallup D., 2025, Particle Monte Carlo methods for Lattice Field Theory (arXiv:2511.15196), https://arxiv.org/abs/2511.15196

Table A1. Predictive performance metrics for the ARIMA(9, 0, 1) and ARIMA(3, 0, 3) models.

Metric	**ARIMA**(9, 0, 1)	**ARIMA**(3, 0, 3)
Mean Squared Error (MSE)	4103.68	5592.04
Root Mean Squared Error (RMSE)	64.06	74.78
Mean Absolute Error (MAE)	46.58	55.45

Yallup D., Kroupa N., Handley W., 2025, in Frontiers in Probabilistic Inference: Learning meets Sampling. https://openreview.net/forum?id=ekbkMSuPo4
Yu X., 2023. https://api.semanticscholar.org/CorpusID:265295307
Yule G. U., 1927, Philosophical Transactions of the Royal Society of London. Series A, 226, 267

APPENDIX A: COMPARISON OF FORECASTING PERFORMANCE

The forecasting performance of ARIMA(9, 1) and ARIMA(3, 3) are compared in Figure A1. ARIMA(9, 1) achieves a markedly better overall forecasting performance as compared to ARIMA(3, 3). It shows the expected growth of predictive uncertainty with time, which is characteristic of long-term ARIMA forecasting. ARIMA(9, 1) forecasts also yielded much lower values for the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) (see Table A1), thereby confirming and validating the results of our model selection methodology.

This paper has been typeset from a TEX/LATEX file prepared by the author.

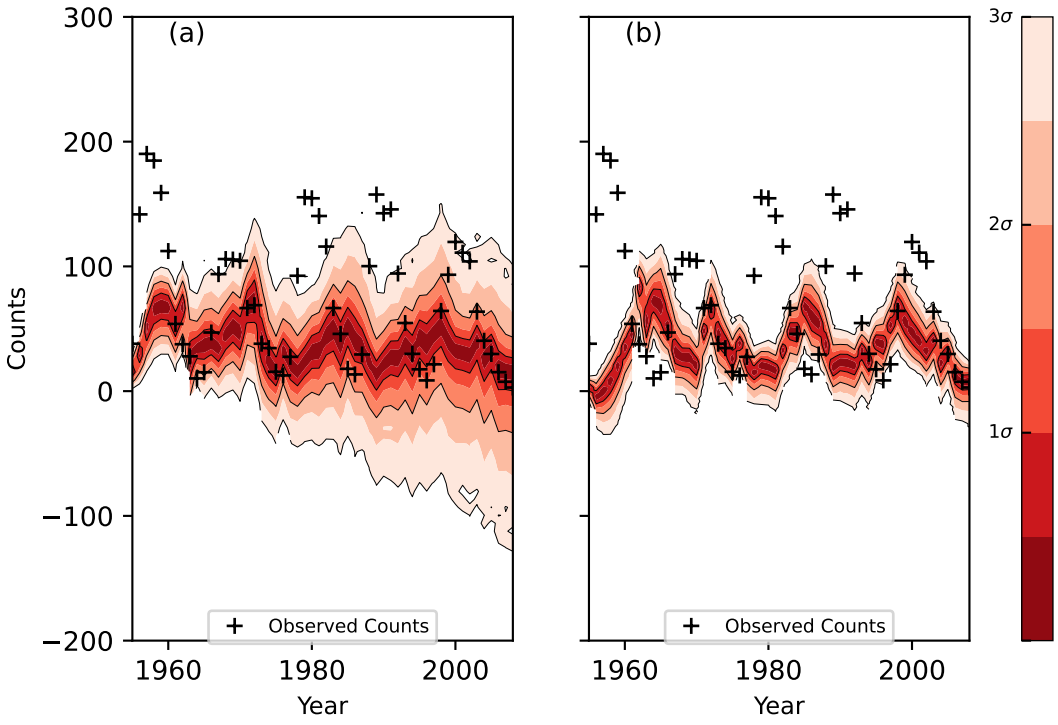

Figure A1. Comparison of posterior predictive forecasts for two ARIMA models applied to the yearly sunspots-number time series. Panels (a) and (b) show the posterior predictive distributions of the ARIMA$(9, 1)$ and ARIMA$(3, 3)$ models, respectively with shaded contours denoting the 1σ,2σ and 3σ credible regions of the predictive posterior distribution $P(\hat{y}_t|t, D_t)$.