Journal of Econometrics and Statistics
Vol. 5, Issue 1, 2025, pp. 107-127

A Novel ARFIMA-ANN Hybrid Model for Forecasting Time Series - and its Role in Explainable AI

David L. Dowe[a,c], Shelton Peiris[b] and Eric Kim[a]

[a] *Department of Data Science and AI, Faculty of IT, Monash University, Clayton, Vic., Australia*
[b] *School of Mathematics and Statistics, University of Sydney, Sydney, Australia*

ABSTRACT
Autoregressive Fractionally Integrated Moving Average (ARFIMA) has been successfully applied in modelling and forecasting with linear economic time series with long memory components. In order to capture additional complex nonlinear economic relationships with many unknown patterns, another popular approach known as the Artificial Neural Network (ANN) can be used. It has been recognised that a combination of both ARFIMA and ANN can be used to capture both the linear and nonlinear components of a time series. This paper proposes an alternative hybrid model, which is distinctive in integrating both the linear and nonlinear components of applied time series with long memory - and in considering both additive and multiplicative models. A simulation study has been carried out to investigate the properties of this ARFIMA-ANN hybrid modelling and forecasting. We justify the usefulness of this proposed hybrid model in practice using empirical data sets from various domains - financial, environmental (pollution), climate (El Niño) and energy (electricity load) - and compare the accuracy of forecasting with existing models. We have shown that, in general, these hybrid models will often produce more accurate forecast values than other - individual - models. We also discuss explainability and interpretability.

KEYWORDS
Long memory, Fractional difference, Heteroscedastic, ARFIMA, Forecasting, neural net, ANN, Hybrid, Explainable AI, XAI, Interpretable.

1. Introduction

The concept of long-range dependency has become popular in applications in economic time series since the work of Granger and Joyeux (1980) [28]. These authors proposed the family of autoregressive fractionally integrated moving average (ARFIMA) models, which essentially replace the traditional integer degree of differencing in ARIMA structure by a fractional degree of differencing in the open interval $(0, 1/2)$. Since then, ARFIMA modelling and analysis has been an important research topic, especially in economics and finance. Analysis of stock index and returns reveals an apparent long memory component in the time series - see, e.g., [6, 35, 39, 41]. These stock market

CONTACT Author[c]. Email: david.dowe@monash.edu

Article History
Received : 27 November 2024; Revised : 28 December 2024; Accepted : 02 January 2025; Published : 08 January 2025

To cite this paper

David L. Dowe, Shelton Peiris & Eric Kim (2025). A Novel ARFIMA-ANN Hybrid Model for Forecasting Time Series – and its Role in Explainable AI. *Journal of Econometrics and Statistics.* 5(1), 107-127.

indices have been analysed using ARFIMA theory. Franco et al. (2018) [26] used the bootstrap approach in developing prediction (or forecast) intervals in the ARFIMA model.

In recent years, the parameter estimation of ARFIMA models has been developed through the state space approach by Palma and Chan (2005) [45] and Dissanayake et al. (2016) [10]. During the past few decades, an approach supported by applied artificial intelligence systems and, in particular, artificial neural networks (ANNs) has been used in time series forecasting. A main reason for this is clear from the fact that fitting a linear model for a given data set is not always satisfactory as any unknown association of the data may typically be nonlinear. A promising approach of capturing such unknown nonlinear relationships is identified as the use of a suitable ANN - see for example, [20]. In their papers, Wang et al. (2013) [66], Khandelwal et al. (2015) [33] and Fang, Dowe et al. (2021) [21] have developed and implemented ARMA-ANN Hybrid modelling with applications for stationary (long) short-memory time series. A different kind of hybrid model is given in [69].

When a time series is homogeneous non-stationary or contains seasonal components, the corresponding ARIMA or seasonal ARIMA (SARIMA) has been used with ANN by (Hermansah et al., 2019) [29] to handle unknown nonlinear relationships. As the concept of long memory is useful in modelling many applied financial time series, this paper extends the existing work and develops a new approach based on ARFIMA-ANN.

The works [21, 26, 39, 41] all have financial applications. Fang, Dowe et al. (2021) [21] uses the Bayesian information-theoretic minimum message length (MML) principle (see, e.g., [13, 49, 62–64] and also has environmental pollution applications - and Fang, Xie et al. (2021) [22] has discussed a pollution application. MML can be thought of in terms of Ockham's razor - see, e.g., [4, 42]. We have financial and environmental (pollution) applications in sec. 5.1 and sec. 5.2 respectively. Many other application areas for this current area of work include (e.g.) climate (see, e.g., sec. 5.3), energy (see, e.g., sec. 5.4), epidemiology (including COVID and other pandemics), etc. and a variety of other areas (see, e.g., sec. 5.5).

There is currently something of a tension within artificial intelligence (AI) between (on the one hand) interpretable, explainable, relatively simple models (upon which we can often do at least some sort of sensitivity analysis) and (on the other hand, opaque black-box) deep learning (seemingly with at most little viable option for a sensitivity analysis). Indeed, the field of explainable AI (XAI) is currently a growing area - due partly to the increasingly deployment of AI systems and the increasing responsibilities delegated to them, and surely also at least some consideration of potential dangers of AI [53] [14, secs. 1.1, 3 and 4.8] [27] (Grace et al., 2023) [7], delegating control [43] [37], in addition to legal considerations including but not limited to responsibility (and law and regulation). In the specific context of large language models (LLMs) and generative AI (without commenting here on ChatGPT or any other specific cases), the notion of understanding [17, sec. 5.1][18, secs. 1, 2.1 and 6][19, sec. 2] is similarly relevant and important in regard to explainable AI and to at least some of the issues from the immediately preceding sentence. One can even raise the issue of getting machines to judge the Turing test (equivalently, administer the Turing test or imitation game) [17, sec. 5][18, sec. 5.2][19, sec. 5][48][30] as a version of machines being able to assess the merits - and interpretability - of other machines (including their arguments and explanations).

Bearing in mind matters from the last paragraph re explainable AI, we build upon our earlier work in showing the merits of having a transparent explainable statistical model as a form of XAI and then extending that with deep learning. More specifically, we extend the approach of earlier work of (Fang, Dowe et al., 2021) [21] by again doing an additive model (or a deep learning model on the residuals, or difference between the true data and the statistical model). We also introduce anew a multiplicative model (or a deep learning model on the ratio between the true data and the statistical model).

Our earlier (additive model) and current (additive and multiplicative models) work shows the merits of having a transparent explainable statistical model as a form of XAI - with the additive model entailing doing deep learning on the residuals, i.e., on the differences between the true raw data and our model. The fact that the residuals are typically smaller in magnitude than the original raw data suggests the promise we have found in this approach. For the multiplicative model (which we can do with univariate data and other cases), the ratio being modelled by deep learning will often approximately equal the value 1. Hybridising conventional statistical modelling with deep learning in this way not only gives an explainable underlying basis to deep learning, but it also often improves the resultant predictions.

With the above in mind, section 2 reviews the theory of autoregressive fractionally integrated moving average (ARFIMA) modelling together with a number of basic results for later reference - and the end of section 2 outlines the rest of the paper. Sec. 3 outlines neural network approaches, and sec. 4 is on hybrid methods in time series analysis. In sec. 5 on applications (abovementioned sec. 5.1 on financial data, sec. 5.2 on pollution data, sec. 5.3 on climate data and sec. 5.4 on energy data), we point out the merits of the hybrid approach. In sec. 6, we summarise, discuss future directions of (research) work, and mention the relevance of this work to explainable artificial intelligence (or explainable AI, or XAI, or interpretable AI).

2. ARFIMA model and basic results

A stationary time series is identified as having long-range dependency or long memory or is highly persistent if

- the autocorrelation function (acf) ρ_k decays very slowly at a hyperbolic rate such that

$$\rho_k \sim k^{2d-1}, \text{ where } 0 < d < 0.5,$$

- the spectrum or spectral density function (sdf) $f(\lambda)$ is unbounded near the origin and follows the power law when $0 < d < 0.5$,

$$f(\lambda) \sim \lambda^{-2d} \text{ as } \lambda \to 0^+.$$

From the above, it is interesting to note that when $d < 0$, the sdf $f(\lambda) \to 0$ as $\lambda \to 0$. In this case, the time series is said to have intermediate memory. It is known that, the time series described above is invertible when $d > -0.5$. These two properties (above) of a stationary long memory time series are clearly different from short memory time series, which are as described below:

- the acf ρ_k decays exponentially such that $\rho_k \sim |\delta|^k$, where $|\delta| < 1$,
- the sdf is bounded and is finite at the origin.

There is significant evidence to show that long memory processes exist in many practical time series analysis problems, especially in economics and finance. It has been shown that a long memory time series $\{X_t\}$ superimposed with a short memory component can be modelled by a suitable member from a family with a fractional degree of differencing $d \in (0, 0.5)$, known as an autoregressive fractionally integrated moving average model of order (p, d, q) or ARFIMA(p, d, q) given by

$$\phi(B)(1-B)^d X_t = \theta(B)\epsilon_t, \tag{1}$$

where B is the backshift operator such that $B^j X_t = X_{t-j}$; and $\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$ and $\theta(B) = 1 + \theta_1 B + \ldots + \theta_q B^q$ are polynomials of orders p and q respectively such that the roots of $\phi(\ell) = 0$ and $\theta(\ell) = 0$ are outside the unit circle (which are known as stationary and invertible conditions respectively); $0 < d < 0.5$ and ϵ_t is a sequence of uncorrelated random variables with mean 0 and a constant variance σ^2, which are not necessarily independent. This sequence ϵ_t (of uncorrelated random variables with mean 0 and a constant variance, σ^2, which are not necessarily independent) is also known as a white noise (WN) sequence and is written as $\epsilon_t \sim WN(0, \sigma^2)$. This family in (1) is well suited for modelling time series with long memory and has become a popular tool in econometrics. Under the stationarity (or AR regularity) conditions, it is easy to show that the Wold representation of (1) is

$$X_t = \psi(B)\epsilon_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}, \tag{2}$$

where $\psi(B) = [\phi(B)]^{-1}(1-B)^{-d}[\theta(B)] = \sum_{j=0}^{\infty} \psi_j B^j$.

It is easy to show that $\psi_j = a_j + a_{j-1}b_1 + \ldots + a_1 b_{j-1} + b_j$ for all $j \geq 1$ with $[\phi(B)]^{-1}[\theta(B)] = \sum_{j=0}^{\infty} a_j B^j$ and $(1-B)^{-d} = \sum_{j=0}^{\infty} b_j B^j$, $b_j = \frac{\Gamma(d+j)}{\Gamma(j+1)\Gamma(d)}$. As these results are distinguishable from a stationary ARMA family (i.e., exponential decay of the autocorrelation function (acf) and a bounded spectrum at the origin), we illustrate them below.

A family of stationary and invertible ARMA(p,q) for short memory time series can be obtained from (1) when $d = 0$. When $d = 1, 2, \ldots$, this becomes an ARIMA(p,d,q) and is used to model non-stationary time series. Carefully investigating the above acf and sdf, a clear distinction between the short memory ARMA and long memory ARFIMA processes can be used in applications.

For illustration, we take the following two models:

- ARMA(1,1): $(1 - 0.6B)X_t = (1 + 0.7B)\epsilon_t$, and
- ARFIMA(1,0.4,1): $(1 - 0.6B)(1 - B)^{0.4} X_t = (1 + 0.7B)\epsilon_t$ models,

where, in each case, ϵ_t is taken from the standard Normal distribution, $N(0, 1)$. The acf plot of ARMA(1,1) shows that the values drop-out very quickly and the spectrum is bounded at the origin unlike in the ARFIMA(1,0.4,1) case (see figure 1).

We have just discussed ARFIMA modelling. As per the end of sec. 1, we now discuss models from neural networks approaches (in sec. 3), then hybrid ARFIMA-LSTM (or hybrid ARFIMA-ANN) models (with some simulated data) in sec. 4, and then comparisons on real-world data-sets in sec. 5. We return at the end of the paper in sec. 6 to discuss, among other things, the possibility of trying to get a further interpretable approximation to the (hybrid) deep learning approach.

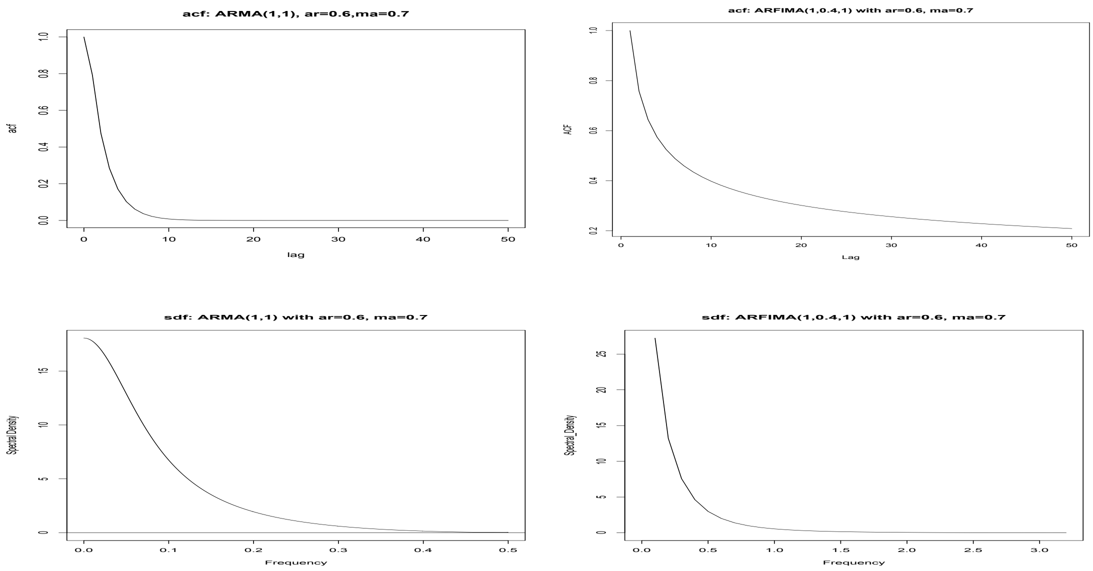

Figure 1. Plot of ACF and SDF of ARIMA and ARFIMA models

3. Neural network approaches

The concept underpinning neural networks and their applications in various applied problems have been around for more than seven decades. However, it has recently been extensively applied in different areas of studies such as medical science, computer science and data science as computing power has been increasing dramatically over recent decades. Neural network research is a renowned branch of machine learning, and the approach is based on a resemblance to the activities of biological neural networks in the body, especially in the brain. The next section, sec. 3.1, is, therefore, devoted to give a brief review of neural networks, known as Artificial Neural Network (ANNs), in order to combine them with long memory time series analysis.

3.1. *ANNs and Time Series Analysis*

As we have seen, the ARFIMA family is used to model a process using a suitable linear approximation following the properties of the acf, pacf (partial auto-correlation function) and the sdf. However, this linear approximation may not be satisfactory in certain applications. Therefore a nonlinear technique has been developed for modelling over a very wide range of time series and related applications. This is a more flexible and a general approach in terms of architecture. The architecture of this method bears a high similarity to the neurons in the brain, hence the name “artificial neural network” or “ANN”. As the name suggests, ANNs use artificial neurons connected in layers to simulate the human synapse or neuronal junction. Therefore, the basic elements of an ANN are interconnected in adjacent layers. A typical simple ANN consists of three types of layers consisting of an input layer, hidden layers each with a nonlinear function, and an output layer. Such a single hidden layer feedforward network is widely used in many applications related to time series modelling and forecasting. This network has simple processing units which are connected by acyclic links as shown in the diagram in figure 2, which shows a feed-forward ANN diagram (where figure 2 only shows one hidden layer, but an outline of more detailed LSTM modelling is given in [21, figs. 1 and 2]). Note that there is a connection from each node in the input layer with the nodes in the hidden layer and from each hidden layer node with the nodes of the output layer (and, if there is more than one hidden layer, there are connections between each layer and the next). All connections carry some (possibly unknown, to be inferred or estimated) weights which play a role in determining the output.

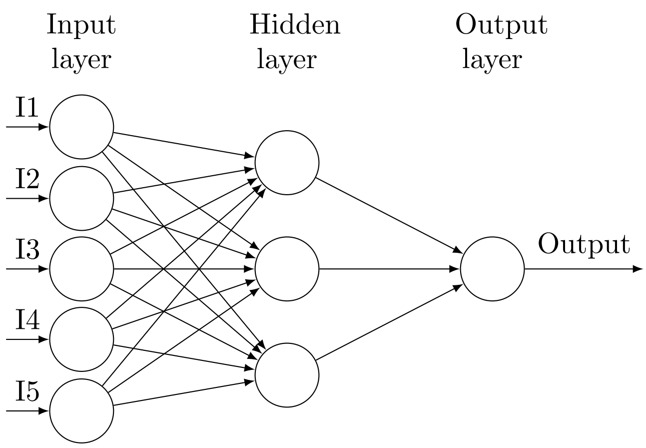

Figure 2. A feed-forward ANN diagram

Suppose that $\{X_t\}$ is a stationary time series. In ANN modelling, a nonlinear functional mapping from the p past observations $X_{t-1}, \ldots, X_{t-p}$ to the current value of X_t is to be obtained. That is, select the optimal p such that the X_t is expressed as the best nonlinear combination satisfying

$$X_t = c + \sum_{j=1}^{q} \beta_j g(f_j(X_{t-1}, \ldots, X_{t-p})) + \epsilon_t, \tag{3}$$

with a simple mathematical functional relationship for $f_j(.)$ such that

$$f_j(.) = \alpha_{0j} + \sum_{i=1}^{p} \alpha_{ij} X_{t-i}, \tag{4}$$

where $g(.)$ is a suitably chosen nonlinear function for all j which is determined by the network, c is a constant and $\epsilon_t \sim WN(0, \sigma^2)$ (where WN is defined as in sec. 2).

The full model is obtained by combining the equations 3 and (4) with $1 + 2q + pq$ parameters c, β_j $(j = 1, 2, \cdots, q)$ and α_{ij} $(i = 0, 1, 2, \cdots, p; j = 1, 2, \cdots, q)$. The transfer function $g(.)$ of the hidden layer is generally taken as a sigmoid function and that of the output layer is a linear function. The sigmoid function (a.k.a. logistic function) is mostly picked up as the activation function in ANN, because its derivative is easy to evaluate and its functional form is given by

$$g(\theta) = \frac{1}{1 + \exp(-\theta)} \tag{5}$$

Let $h(.)$ be the nonlinear functional form of (3) together with (4) and (5) and, for simplicity, write this as

$$X_t = h(X_{t-1}, \ldots, X_{t-p}, \boldsymbol{\delta}) + \epsilon_t, \tag{6}$$

where $\boldsymbol{\delta}$ is the vector of all parameters in the model. (An alternative way of doing this ANN modelling would be as in [36]. We note that (Allen et al., 2023) [3] argued how Generative Adversarial Neural Nets (GANs) can be implemented to improve modelling and forecasting of time series. We also mention the survey of recurrent neural network - or RNN - methods for time series forecasting in [31].)

The model coefficients in the final ANN are the weights of each link. It is clear that both these two popular approaches of ARFIMA and ANN are rich classes from different models. As usual a large sample of data is required in order to build satisfactory models - and, in ARFIMA, the principle of parsimony is often used in choosing the best possible model for forecasting. Let the forecast from equation (6) be $\hat{X}_t$ and the corresponding forecast error $\epsilon_t = X_t - \hat{X}_t$. The next section considers a hybrid methodology by combining ARFIMA and ANN in time series forecasting. We will discuss the degree to which such a combination can be thought of as - or might contribute to - a form of explainable AI (or XAI).

4. Hybrid methods in long memory time series analysis: a simulation study

In long memory time series analysis, one finds the best approximate member of the ARFIMA family to a given set of data. However, the true structure of the data is unknown and it may be superimposed with an unknown complex nonlinearity. Ignoring such complexity and using a linear approximation through a member of the ARFIMA family runs the risk of producing relatively inaccurate forecast values. As described in sec. 3, the use of ANN in time series analysis will typically improve the forecasting accuracy when it contains undetectable nonlinear or complex components. Therefore, one can develop a suitable hybrid methodology in order to improve the forecast accuracy by combining both of these approaches of ARFIMA (linear) and ANN (nonlinear).

Suppose that an observed long memory time series Y_t consists of both the linear and nonlinear components denoted by L_t and N_t respectively at time t. There are two main potential cases in practice, as given below:

- Additive components: In this case the time series is $Y_t = L_t + N_t$.
- Multiplicative components: This is given by $Y_t = L_t \times N_t$.

The linear component L_t can be modelled by equation (1) and letting $\hat{L}_t$ be the corresponding forecast value at t. In this study, we include ARFIMA with additive components to develop the ARFIMA-ANN approach. For the ARFIMA simulation, we consider $ARFIMA(1, d, 1)$ and use the *arfima.sim* function from the *arfima* package in R to simulate the time series data.

Figure 3 shows an example of simulated data - more specifically, a time series plot of a set of 3000 data simulated through the model ARFIMA(1,0.4,1) given by $(1 - 0.6B)(1 - B)^{0.4}X_t = (1 + 0.7B)\epsilon_t$. We provide figure 3 as a visualisation aid, although it is not directly relevant to the small set of simulation experiments described below.

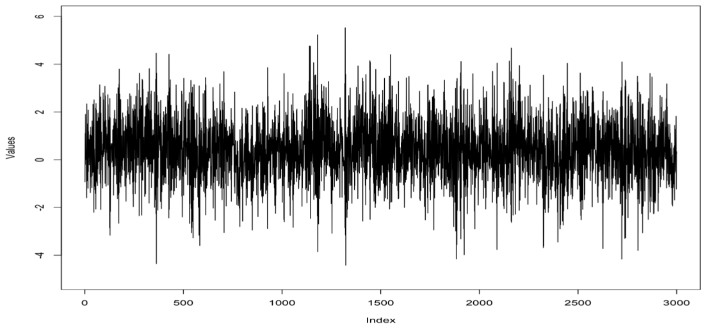

Figure 3. Simulated data example

For the simulation experiments here in sec. 4 leading to tables 1, 2, 3 and 4, we used uniform random numbers in the open interval range $(-1, 1)$ to generate two values of ϕ and also used uniform random numbers in the range $(0.0001, 0.4999)$ to generate five values of d. On the other hand, we consider the ANN model with one hidden layer with three neurons and the sigmoid function in both the hidden layer and in the output layers. For the simulation study of ANN series, we consider a mean of 2 different values, i.e., 1 and 2 and a standard deviation of 5 different values, namely $1, 2, 3, 4, 5$ in the Gaussian function to generate $2 \times 5 = 10$ different sets of weights in the ANN model - and the final simulated time series data generated is, in turn, the summation and the multiplication respectively of ANN and ARFIMA simulated parts. In each simulation, we simulated a sample of 1,000 time-series data with length 3,000.

We then (or further) estimate the ARFIMA$(1, d, 1)$ and the hybrid additive and multiplicative ARFIMA$(1, d, 1)$-ANN models respectively from the relevant generated data. The estimation parts have been done using the *numpy* library in Python. Tables 1, 2, 3 and 4 show the results of the average of mean-squared-error (MSE) on the relevant generated data in 1,000 samples as a comparison between ARFIMA and ARFIMA-ANN hybrid models, including the comparison of the Akaike Information Criterion (AIC) and Schwarz's BIC. AIC is given by [1]

$$argmin_{\vec{\theta}} \;\; L + k \tag{1}$$

where $\vec{\theta}$ is the set of parameters to be estimated, k is the number of parameters in $\vec{\theta}$, L is the negative log-likelihood, and N is the amount of data; and Schwarz's Bayesian Information Criterion (BIC) is given by [25, 50]

$$argmin_{\vec{\theta}} \;\; L + (k/2)\log(N) \tag{2}$$

We use I_1 to I_{10} to indicate the ten $(2 \times 5 = 10)$ rows of different simulation data sets in tables 1, 2, 3 and 4. Real-world data-sets are considered in sec. 5.

4.1. *Possible extensions to the additive and multiplicative models*

This section discusses a possible extension which is not carried out in the paper. It can safely be skipped without loss of continuity.

By way of notation, $\ln = \log_e$, and $exp()$ will sometimes be used to denote $e^{()}$ or e

Table 1. MSE for ten simulation data sets - 1% out sample forecasting

	Average of MSE (& Standard deviation)					
	ARFIMA		ARFIMA-ANN-Additive		ARFIMA-ANN-Multiplicative	
	AIC	BIC	AIC	BIC	AIC	BIC
I_1	0.2461 (0.038)	0.42 (0.0751)	0.2681 (0.0452)	0.2667 (0.0159)	0.117 (0.0927)	0.0974 (3.388)
I_2	0.258 (0.0357)	0.3928 (0.0594)	0.3077 (0.0456)	0.3128 (0.0104)	0.0976 (0.072)	0.0866 (3.3721)
I_3	0.2621 (0.0412)	0.4124 (0.0786)	0.3304 (0.0456)	0.3367 (0.0338)	0.1253 (0.0343)	0.1431 (3.4174)
I_4	0.2379 (0.0364)	0.4144 (0.081)	0.3319 (0.0446)	0.3276 (0.0195)	0.1153 (0.065)	0.0879 (3.3959)
I_5	0.3319 (0.0539)	0.3168 (0.0803)	0.3753 (0.0638)	0.3783 (0.0274)	0.1362 (0.1113)	0.1742 (3.4229)
I_6	0.2228 (0.0443)	0.4201 (0.0461)	0.2846 (0.0538)	0.2854 (0.015)	0.0795 (0.0364)	0.1326 (3.3962)
I_7	0.2338 (0.0413)	0.2867 (0.0853)	0.2921 (0.0459)	0.2872 (0.0392)	0.0971 (0.0719)	0.1475 (3.4122)
I_8	0.3782 (0.1246)	0.7309 (0.1645)	0.4148 (0.1302)	0.4118 (0.1119)	0.1333 (0.1412)	0.1446 (3.4362)
I_9	0.2405 (0.0492)	0.1937 (0.0793)	0.2638 (0.0521)	0.2692 (0.0368)	0.1163 (0.0378)	0.0717 (3.4078)
I_{10}	0.1429 (0.0148)	0.2518 (0.0164)	0.1971 (0.0156)	0.2035 (0.0147)	0.0682 (0.0224)	0.1246 (3.408)

Table 2. MSE for ten simulation data sets - 5% out sample forecasting

	Average of MSE (& Standard deviation)					
	ARFIMA		ARFIMA-ANN-Additive		ARFIMA-ANN-Multiplicative	
	AIC	BIC	AIC	BIC	AIC	BIC
I_1	0.9754 (0.1784)	1.1493 (0.2155)	0.9974 (0.1857)	0.996 (0.1563)	0.305 (0.1629)	0.2853 (3.4583)
I_2	1.2474 (0.175)	1.3821 (0.1987)	1.2971 (0.1849)	1.3021 (0.1497)	0.3525 (0.1417)	0.3416 (3.4417)
I_3	1.4431 (0.1964)	1.5934 (0.2337)	1.5114 (0.2007)	1.5177 (0.1889)	0.4296 (0.1118)	0.4475 (3.4949)
I_4	0.952 (0.1997)	1.1285 (0.2443)	1.0461 (0.2079)	1.0417 (0.1828)	0.2993 (0.1466)	0.2719 (3.4775)
I_5	1.3825 (0.2403)	1.3674 (0.2667)	1.4259 (0.2502)	1.4289 (0.2138)	0.407 (0.2045)	0.445 (3.5161)
I_6	0.9129 (0.2085)	1.1102 (0.2102)	0.9747 (0.2179)	0.9756 (0.1792)	0.2574 (0.1185)	0.3105 (3.4783)
I_7	1.0295 (0.1589)	1.0824 (0.2029)	1.0878 (0.1636)	1.0829 (0.1569)	0.3022 (0.1308)	0.3525 (3.471)
I_8	1.848 (0.4869)	2.2008 (0.5268)	1.8847 (0.4926)	1.8817 (0.4742)	0.5122 (0.3224)	0.5234 (3.6174)
I_9	1.0886 (0.2543)	1.0417 (0.2844)	1.1119 (0.2572)	1.1172 (0.2419)	0.3349 (0.1404)	0.2903 (3.5104)
I_{10}	0.5516 (0.0823)	0.6605 (0.0839)	0.6058 (0.0831)	0.6123 (0.0822)	0.1735 (0.0562)	0.2299 (3.4418)

Table 3. MSE for ten simulation data sets - 10% out sample forecasting

	Average of MSE (& Standard deviation)					
	ARFIMA		ARFIMA-ANN-Additive		ARFIMA-ANN-Multiplicative	
	AIC	BIC	AIC	BIC	AIC	BIC
I_1	1.9017 (0.2786)	1.9327 (0.306)	1.9579 (0.2808)	1.963 (0.2777)	0.5184 (0.2045)	0.5826 (3.5565)
I_2	1.8797 (0.2591)	2.0072 (0.2792)	1.9113 (0.262)	1.9111 (0.2467)	0.5178 (0.1321)	0.5544 (3.5416)
I_3	1.8502 (0.3016)	1.9819 (0.3285)	1.9427 (0.3079)	1.9498 (0.2855)	0.5206 (0.217)	0.5279 (3.5296)
I_4	1.8328 (0.2576)	2.2569 (0.2738)	1.8525 (0.2662)	1.8567 (0.2394)	0.4884 (0.2071)	0.5139 (3.5406)
I_5	2.5234 (0.4012)	2.4781 (0.4022)	2.6187 (0.4044)	2.6214 (0.3898)	0.6771 (0.2471)	0.702 (3.5752)
I_6	1.7028 (0.3251)	1.6877 (0.3516)	1.7463 (0.3351)	1.7493 (0.2986)	0.4896 (0.2469)	0.5275 (3.5585)
I_7	1.7475 (0.3219)	1.833 (0.3494)	1.7997 (0.33)	1.7997 (0.3298)	0.4864 (0.2436)	0.4738 (3.5068)
I_8	2.7473 (0.9676)	3.0957 (0.9929)	2.8001 (0.9766)	2.7981 (0.9416)	0.7425 (0.5311)	0.7217 (3.8309)
I_9	1.7839 (0.3521)	1.9573 (0.3788)	1.7921 (0.3558)	1.7963 (0.3561)	0.4992 (0.2444)	0.5137 (3.5519)
I_{10}	1.1139 (0.1042)	1.1224 (0.1283)	1.1611 (0.1047)	1.1623 (0.0941)	0.3261 (0.1179)	0.3041 (3.444)

Table 4. MSE for ten simulation data sets - 20% out sample forecasting

	Average of MSE (& Standard deviation)					
	ARFIMA		ARFIMA-ANN-Additive		ARFIMA-ANN-Multiplicative	
	AIC	BIC	AIC	BIC	AIC	BIC
I_1	2.4296 (0.248)	2.6035 (0.2851)	2.4516 (0.2553)	2.4502 (0.2259)	0.6798 (0.1977)	0.6601 (3.4931)
I_2	1.5693 (0.2095)	1.704 (0.2332)	1.619 (0.2194)	1.624 (0.1842)	0.4355 (0.1589)	0.4246 (3.459)
I_3	1.129 (0.2605)	1.2793 (0.2978)	1.1973 (0.2648)	1.2036 (0.253)	0.3487 (0.1439)	0.3665 (3.527)
I_4	2.2966 (0.1591)	2.4731 (0.2037)	2.3906 (0.1673)	2.3862 (0.1421)	0.6459 (0.1263)	0.6185 (3.4572)
I_5	2.9102 (0.4001)	2.8951 (0.4265)	2.9536 (0.41)	2.9567 (0.3736)	0.8008 (0.2844)	0.8387 (3.596)
I_6	2.032 (0.2886)	2.2293 (0.2903)	2.0937 (0.298)	2.0946 (0.2593)	0.5458 (0.1586)	0.5989 (3.5184)
I_7	1.7921 (0.4358)	1.8451 (0.4798)	1.8505 (0.4404)	1.8456 (0.4337)	0.4988 (0.2692)	0.5491 (3.6094)
I_8	2.238 (1.2726)	2.5907 (1.3124)	2.2746 (1.2782)	2.2716 (1.2598)	0.6126 (0.7152)	0.6239 (4.0102)
I_9	1.7247 (0.2509)	1.6779 (0.2811)	1.748 (0.2539)	1.7534 (0.2386)	0.4988 (0.1387)	0.4543 (3.5087)
I_{10}	1.5008 (0.0619)	1.6097 (0.0636)	1.555 (0.0627)	1.5614 (0.0618)	0.4182 (0.046)	0.4746 (3.4316)

to the power of.

Noting that the logarithm of a product is the sum of the logarithms, we can generalise and continue the notion of additive model and multiplicative model by noting the following and proceeding iteratively:

$$(A + a_0) - (B + b_0) = (A - B) \text{ when } a_0 = b_0 = 0$$
$$\text{(which corresponds to the additive model),}$$

$$e^{(\ln(A+a_0)+a_1)-(\ln(B+b_0)+b_1)} = e^{\ln(A)-\ln(B)} = A/B \text{ when } a_1 = b_1 = a_0 = b_0 = 0$$
$$\text{(which corresponds to the multiplicative model),}$$

$$\begin{aligned} & exp(exp((\ln(\ln(A + a_0) + a_1)) + a_2 - (\ln(\ln(B + b_0) + b_1)) + b_2)) \\ = \; & exp(exp(ln(ln(A)) - ln(ln(B)))) \quad (\text{when } a_2 = b_2 = a_1 = b_1 = a_0 = b_0 = 0) \\ = \; & exp(ln(A)/ln(B)) = A^{1/ln(B)}, \end{aligned}$$

$$\begin{aligned} & exp(exp(exp(\ln(\ln(\ln(A + a_0) + a_1)) + a_2) + a_3 \\ & - \ln((\ln(\ln(B + b_0) + b_1)) + b_2) + b_3))) \\ = \; & exp(exp(exp(\ln(\ln(\ln(A))) - \ln(\ln(\ln(B)))))) \quad (\text{when } a_i = b_i = 0, 0 \le i \le 3) \\ = \; & exp(exp((lnlnA)/(lnlnB))) \\ = \; & exp((lnA)^{1/(lnln(B))}), \end{aligned}$$

etc., with the definition of this class of models being able to be continued iteratively. And, here, A can be thought of as the long memory time series Y_t and B can be thought of as the linear model L_t, both mentioned at the start of sec. 4. Recalling N_t from the start of sec. 4, we then have that $N_t = (A - B) = (Y_t - L_t)$ for the additive model, and $N_t = (A/B) = (Y_t/L_t)$ for the multiplicative model, and so on, etc.

The purpose of the terms $a_0, b_0, a_1, b_1, a_2, b_2, a_3, b_3, ...,$ etc. (if they are used at all) is primarily to ensure that we do not take the logarithm of a non-positive number -

but (if the relevant term is not 0) each such term also has the potential to be used to assist in fitting data.

5. Some applications

In discussing simulations in sec. 4 and the applications in the current section (sec. 5), we sample from LNPPP space [11, sec. 0.2.7][21, sec. 6.1] to do our experiments. We argue elsewhere [11, footnotes 75 and 76][12][13, sec. 3][14, sec. 4.1][21, sec. 6] about various uniqueness and invariance properties of log-loss (or logarithm loss). Squared error is a popular method which is also a variant of log-loss (under the assumption of Gaussian or Normal errors).

Further to the simulations in sec. 4, we applied different approaches to analysing some real-world data-sets. In [21, secs. 6.1, 6.2 and 6.3], we respectively considered simulation data, financial data and environmental pollution data. We follow now in similar vein, with additional application areas. Our hybrid methods throughout applications in the sec. 5 (the current section) below will use the additive model from sec. 4 (as this seemed to perform better than the multiplicative model).

More specifically, we consider simulation data (with both additive and multiplicative hybrid models) in sec. 4. We then (with the only hybrid model being the additive model from sec. 4) consider applications to financial data (in sec. 5.1), environmental pollution data (PM10 in sec. 5.2), climate data (El Niño Southern Oscillation, or ENSO, data in sec. 5.3), and energy data (electricity load data in sec. 5.4).

We present results in the tables to follow, including the coefficient of determination, R^2. Throughout, all of the rolling forecasts were just 1 step forecast at each time point - so (e.g.) the R^2 score for $T = 150$ is how accurate the 1 step forecasts (rolling) were for the testing data of 150 time intervals. Also throughout, for the average MSE for non-rolling forecast for a forecast window of T, it was just a single forecast of T steps.

In sec. 5.5, we mention some other possible application areas.

5.1. *Financial data*

As per [21, sec. 6.2], "The stock prices were selected from the components of the Dow Jones Industrial Average, including Apple (APPL), Boeing (BA), Cisco System (CSCO), Goldman Sachs (GS), IBM, Intel (INTC), Johnson & Johnson (JNJ), JP-Morgan Chase (JPM), Coca-Cola (KO), and 3 M (MMM). The data selected start at 23 September 2016 and finish at 22 September 2021, with a total of 1258 trading days.". The values of R^2 for rolling forecast window of $T = 150$ are given in Table 5.

Table 6 shows the average mean squared error (MSE) for non-rolling forecasts on different sizes of financial data forecast windows.

Performances on the 10 individual data-sets (Apple (APPL), Boeing (BA), Cisco System (CSCO), Goldman Sachs (GS), IBM, Intel (INTC), Johnson & Johnson (JNJ), JPMorgan Chase (JPM), Coca-Cola (KO), and 3 M (MMM)) vary, but - on average - it appears that the LSTM is the worst method considered but the Hybrid ARFIMA-LSTM is perhaps best, with BIC possibly slightly better than AIC.

Table 5. R^2 Score for Rolling Forecast Window of T = 150 on Financial Data

Stock	LSTM	AIC ARFIMA-LSTM Hybrid	BIC ARFIMA-LSTM Hybrid	AIC ARFIMA	BIC ARFIMA
AAPL	0.5907	0.9617	0.9537	0.9449	0.9449
BA	0.7548	0.8377	0.8243	0.8407	0.8407
CSCO	0.8645	0.9558	0.9439	0.9613	0.9613
GS	0.9097	0.9612	0.959	0.9601	0.9599
IBM	0.9161	0.9538	0.9488	0.9526	0.9526
INTC	0.9136	0.9359	0.9311	0.9348	0.9364
JNJ	0.7936	0.9268	0.9292	0.9278	0.9286
JPM	-0.5108	0.8092	0.8144	0.8172	0.8172
KO	0.9294	0.9519	0.9511	0.952	0.951
MMM	0.8917	0.9247	0.9271	0.9213	0.9213
Average R^2 Score	0.7053	0.9219	0.9183	0.9213	0.9214

Table 6. Average of MSE for Non-Rolling Forecast on Financial Data

T (size of Forecast Window)	AIC ARFIMA	BIC ARFIMA	LSTM	AIC ARFIMA-LSTM Hybrid	BIC ARFIMA-LSTM Hybrid
3	6.266	6.041	26.110	8.293	5.698
6	20.576	20.090	47.269	18.950	20.014
9	26.725	26.023	94.411	24.609	25.033
12	35.938	34.910	106.947	34.918	35.529
15	53.768	52.588	102.767	52.087	52.928

5.2. *Environmental Pollution PM10 data*

Further to the environmental pollution data in [21, sec. 6.3], we now consider data from 01/03/2012 to 28/02/2017 from the study in [70]. We use the dataset PRSA_Data_Gucheng_20130301-20170228.csv, PM10.

Table 7 gives the R^2 score for rolling forecast window of $T = 150$ on PM10 pollution data, and table 8 gives the mean squared error (MSE) for rolling forecast window of $T = 150$ on this PM10 pollution data.

Table 7. R^2 Score for Rolling Forecast Window of T = 150 on PM10 Pollution Data (Gucheng)

LSTM	AIC ARFIMA-LSTM Hybrid	BIC ARFIMA-LSTM Hybrid	AIC ARFIMA	BIC ARFIMA
0.718953	0.740877	0.741298	0.739319	0.739319

Table 8. MSE for Rolling Forecast on PM10 Pollution Data (Gucheng)

T (size of Forecast Window)	AIC ARFIMA	BIC ARFIMA	LSTM	AIC ARFIMA-LSTM Hybrid	BIC ARFIMA-LSTM Hybrid
3	47.831	47.831	75.035	40.931	16.645
5	76.258	76.258	99.409	69.011	42.968
10	92.071	92.071	124.203	83.815	52.705
25	240.256	240.256	231.804	236.247	227.860
50	275.724	275.724	280.637	271.582	262.360
75	240.333	240.333	251.360	236.177	226.569
100	270.719	270.719	280.848	268.197	267.162
125	360.949	360.949	374.452	360.233	367.911
150	411.071	411.071	443.187	408.614	407.951

From tables 7 and 8, best again seems to be the ARFIMA-LSTM hybrid, especially BIC ARFIMA-LSTM hybrid.

However, on separate static non-rolling forecasts on this data, LSTM seemed to perform best.

5.3. *Climate - El Niño Southern Oscillation (ENSO) data*

Many papers have been written about climate change and climate science. One application comparing methods including Akaike's AIC and Schwarz's BIC is [61]. Recalling sec. 5.1, one application discussing both finance and climate is [22].

In our current paper here, our El Niño - Southern Oscillation (ENSO) data is from 1951 to 2023, taken from [2].

Table 9 (R^2) shows LSTM worst and the other methods comparable. Table 10 (mean squared error, or MSE) again shows LSTM worst - with BIC ARFIMA-Hybrid LSTM being generally best.

Table 9. R^2 Score for Rolling Forecast Window of T = 150 on ENSO data

AIC ARFIMA	BIC ARFIMA	LSTM	AIC ARFIMA-LSTM Hybrid	BIC ARFIMA-LSTM Hybrid
0.5034	0.5053	0.4346	0.4894	0.4961

Table 10. MSE for Non-Rolling Forecast on ENSO data

T (size of Forecast Window)	AIC ARFIMA	BIC ARFIMA	LSTM	AIC ARFIMA-LSTM Hybrid	BIC ARFIMA-LSTM Hybrid
3	1.0099	1.0009	2.3797	0.7129	0.5189
6	1.7758	1.9980	2.9730	1.3715	1.1139
9	2.4484	2.8266	4.0110	1.9879	1.9893
12	2.9438	3.3896	4.6210	2.4881	2.2093
15	2.4001	2.7681	3.6704	2.0269	1.8404

5.4. *Energy - and electricity load data*

Energy, clean energy, renewables and related areas continue to draw increasing amounts of research and attention. One of many such works in this growing area is [5]. Another is on smart grids and power data [67]. The energy data we consider here is electricity load data.

Our electricity data - or electricity load data - was from [44]. We used timeseries 30 min.csv in the package and selected GB_GBN_load_actual_entsoe_transparency (Total load in Great Britain in MW as published on ENTSO-E Transparency Platform) as the variable to do the modelling on. We used the most recent 25000 data points, corresponding to the period from 30/4/2019 to 30/9/2020.

Table 11. R^2 Score for Rolling Forecast Window of T = 150 on Electricity Load data

LSTM	AIC ARFIMA-LSTM Hybrid	BIC ARFIMA-LSTM Hybrid	AIC ARFIMA	BIC ARFIMA
0.9713	0.9665	0.9660	0.9657	0.9657

Table 11 (R^2) and table 12 (mean squared error, or MSE) show LSTM perhaps performing best of the methods considered.

5.5. *Some further possible applications*

In addition to the simulations in sec. 4 and the applications in secs. 5.1, 5.2, 5.3, and 5.4, some further possible applications include data pertaining to (e.g.) adverse drug reaction detection [9], COVID-19 and epidemiology [71], and cytology [38], etc.

As we note below in sec. 6, results here in sec. 5 (secs. 5.1, 5.2, 5.3, and 5.4) were mixed, with the hybrid ARFIMA-LSTM - perhaps the BIC hybrid ARFIMA-LSTM - arguably performing best.

Table 12. MSE for Non-Rolling Forecast on Electricity Load data

T (size of Forecast Window)	AIC ARFIMA	BIC ARFIMA	LSTM	AIC ARFIMA-LSTM Hybrid	BIC ARFIMA-LSTM Hybrid
3	1330580	1330580	1284138	2183350	2117115
6	4947127	4947127	3986888	6137384	4970944
9	5213951	5213951	6912276	6261560	2204099
12	4629543	4629543	2624674	18873781	6485042
15	4809024	4809024	4839648	3536746	14530578

6. Discussions, conclusions and future work

We highlight that an advantage of doing more conventional simple modelling first (e.g., [21]) is that the residuals (the differences between the true data and the modelled values) will typically be smaller in magnitude than the original data-set - and we have referred to this as the additive model. An alternative which we have newly presented was the multiplicative model in sec. 4, where we also did some brief comparisons on simulated data.

In sec. 5, the only hybrid model considered was the additive model. Results were mixed, with the hybrid ARFIMA-LSTM - perhaps the BIC hybrid ARFIMA-LSTM - arguably performing best.

As such, deep learning can be used to polish, refine or otherwise modify what has already been done by more conventional modelling (such as Akaike's AIC, Schwarz's BIC, minimum message length - or MML [62–64] - etc.). In the (autoregressive or) AR model in [25] and in the ARIMA model in [21], (as well as in a separate problem also on sequential data [23]), we used AIC, BIC and MML (as is also done in, e.g., [49]). With the more general ARFIMA model in this new current paper, we are using Akaike's AIC and Schwarz's BIC.

We also make some observations about explainable artificial intelligence, or explainable AI (XAI). First, the more conventional model is relatively simple - and so already gives us a form of XAI, as well as being amenable to sensitivity analysis. And, second, for those who like to do XAI by taking the model arising from deep learning and then trying to approximate it with an interpretable explainable model (such as, e.g., a decision tree or classification tree), this can also still be done here along similar lines to our original earlier approach of first doing a more conventional modelling and then doing deep learning on the residuals - here, we would simply apply such a process (of interpretable approximation) on our deep learning model (obtained from the residuals). To paraphrase, in our earlier work [21], we have done conventional statistical modelling (as a form of XAI) and then deep learning on the (additive) residuals - and it would be possible to extend this by applying a different XAI approach on the residuals from the new deep learning models in this paper (whether being used additively or - at least for univariate data - multiplicatively or an approach from sec. 4.1) or quite possibly from any deep learning model. As outlined largely in sec. 1, advantages of explainable and interpretable models include sensitivity analysis (typically) and improvements in matters pertaining to (e.g.) understanding, trust, law, responsibility and legal liability, safety and thwarting (potentially grave) dangers, etc.

In our paper here in sec. 4 above, we have built upon our earlier work and used

AIC and BIC on ARFIMA and done some brief simulations with the ARFIMA-ANN-Additive and ARFIMA-ANN-Multiplicative hybrid approaches. Future work could, among other things, re-visit the simulations of sec. 4 more thoroughly.

Given the generality of real-word data sources, our paper and its results suggest merit in doing such appropriate hybridisation of the ARFIMA approach with ANNs. In terms of explainable artificial intelligence (or explainable AI, or XAI), we have an underlying statistical model (ARFIMA) which is (transparent or) relatively easy to interpret (or explain) and then that is enhanced by incorporating a machine learning model (ANNs) which is oqaque (or black-box, or more difficult to interpret). As discussed elsewhere in this section (sec. 6), the work could be extended by seeking further and better interpretable explainable approximation(s) to the relevant resultant (deep learning) ANNs (possibly including Kolmogorov-Arnold networks, or KANs).

We now discuss future work - including passing mention of generalising beyond additive and multiplicative (as per sec. 4.1), attempting to loosely explain the success of deep learning, and also saying something about flat minima, generalisability and robustness. On the issue of explaining the success of deep learning, Bayesian theory and algorithmic information theory appear to tell us that - in some sense, MML gives the single best inference ([64][62, chap. 2][16]) and - Solomonoff posterior-weighted prediction [51, 52, 54] gives us the optimal predictor. Deep learning typically does not have anything like the simplicity of AIC, BIC and MML but rather seems to do something like making hybrid variables with its early hidden layers and inter-layer connections and do some sort of weighted prediction with its later hidden layers and inter-layer connections. The generation of new variables in the early layers of ANNs is in the spirit of algorithmic information theory (AIT) [51, 52, 54, 64][62, chap. 2]. The combination of values in later layers of ANNs is in the spirit of Solomonoff's algorithmic probability [51, 52, 54]. With ALP denoting algorithmic probability, it could be contended that deep learning is a way to approximate Solomonoff's notions of "resource limited ALP" [55] - or resource bounded probability (RBP) [56] - in time-limited optimization problems [57][14, secs. 3.2 and 3.4]. We note that some variants of MML - e.g., MMLD or I_{1D} from [11, sec. 0.2.2], [12, p. 451, eqn (4)], [24], [62, secs. 4.10 and 4.12.2 and chap. 8, p. 360] and [13, sec. 6.3] - have at least the potential to sometimes select more than one (local) optimum in the message length, and this could be thought of as advocating a (weighted) combination of the relevant models.

On the issues of discovering flat minima [32] and the related issue of relative flatness and generalization [47], MML would appear to have much to say about relative flatness, sensitivity and generalizabilty. The Wallace and Freeman (1987) form of MML (sometimes since called MML87) [65][64, sec. 6.1.2][23, sec. 4, eq. (10)][62, chap. 5][12, p. 451, eqn (5)][13, sec. 6.3, expression (20)][21, equation (8)] considers the value of a continuous Bayesian prior density over a region of volume proportional to the reciprocal of the square root of the expected Fisher information as a major contributor to the length of the first part of the MML message. If the Fisher information is small (e.g., the amount of data, N, is small) then there will be two consequences. First, because the Fisher information is a version of the second derivative, the peak will be relatively flat. And, second, the first part of the message will be shorter (because of the small Fisher information), in turn making this model a better candidate to be the MML model. Our experience with MML models is that they tend to be relatively simple and to generalise well (e.g., [64, sec. 9][13, 59, 62]). In short, flatness tends to imply small Fisher information, which tends to imply better (prior) probability, which tends to imply short length of first part of MML message, which tends to imply better candidacy for being the MML model, which in turn tends to imply better generalisability.

The informal argument just given immediately above says something of the robustness of the MML model. In combining several models (in the spirit, e.g., of Solomonoff) rather than relying on the simple best model, predictions will typically be more robust. There is room for further discussion on robustness of models and networks in the case when variables and/or network nodes are removed.

In addition to intended generalisations mentioned earlier in this section (sec. 6), certain Bayesian approaches could use boosting priors ([60, sec. 3.4][11, sec. 0.2.6][13, sec. 6.9]) and horseshoe priors [8] for cases of sparsity and shrinkage. Or, additionally and alternatively, such approaches could use Bayesian priors of different motivation [13, sec. 7.1][14, sec. 4.3]. Further generalisations intended for future work include one or more of (e.g.) further multidimensional multivariate models, SARMA, βSARMA [34], the approaches from [46], images, spatio-temporal (image) models, etc. and addressing at least some of these with MML.

Conflict of Interest

The authors confirm that this article content has no conflict of interest.

Acknowledgements

The authors extend their sincere gratitude to the editor and the three anonymous referees for their insightful comments and constructive suggestions - and these have greatly enhanced the quality and readability of this paper. The authors also thank and acknowledge the support of - and many fruitful discussions with - Zheng Fang and Dedi Rosadi in the lead-up to this paper. The first author thanks Mark Carman and Michael Kamp for preliminary discussions of relative flatness and MML (as mentioned in sec. 6). The second author completed this work during his tenure at UC Davis in Fall 2024 and gratefully acknowledges the warm hospitality and support provided by the Department of Statistics during his visit. We also thank Puti Dai for some initial communications regarding this work.

References

[1] Akaike, Hirotugu. A new look at the statistical model identification. *IEEE transactions on automatic control* **19**, no. **6**: 716-723, 1974.

[2] Albers, S. *rsoi: Import Various Northern and Southern Hemisphere Climate Indices.* https://github.com/boshek/rsoi/ , https://boshek.github.io/rsoi/ , 2022 (accessed 2023).

[3] Allen, D E, L Mushunje and S. Peiris: GANs through the looking glass: How real is the fake financial data created by Generative Adversarial Neural Nets?, *Proc. 25th International Congress on Modelling and Simulation (MODSIM2023)*, Vaze J, Chilcott C., Hutley L and Cuddy S M (eds.), Modelling and Simulation Society of Australia and New Zealand Inc. Australia, (2023), 29–35. ISBN 978-0-9872143-0-0

[4] Allison, L., Coding Ockham's Razor, *Springer*, 2018.

[5] Andrew, Lachlan L H and Pop, Draga Doncila and Razzaghi, Reza and Dowe, David L, Identifying Flexible Pool Pumps Suitable for Distributed Demand Response Schemes, *IOP Conference Series: Earth and Environmental Science*, volume **322**, number **1**, pages 012022, IOP Publishing, 2019.

[6] Baillie, Richard T., Tim Bollerslev, and Hans Ole Mikkelsen. Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics,* **74(1),** 3-30, 1996.

[7] AI 'godfather' Geoffrey Hinton warns of dangers as he quits Google. *British Broadcasting Corporation.* https://www.bbc.com/news/world-us-canada-65452940 (accessed 10/Jan/2024). 2 May 2023.

[8] Carvalho, Carlos M., Nicholas G. Polson, and James G. Scott. Handling sparsity via the horseshoe. In *Artificial intelligence and statistics*, pp. 73-80. PMLR, 2009.

[9] Dey, Arijit, Jitendra Nath Shrivastava, and Chandan Kumar. Graph Based Zero Shot Adverse Drug Reaction Detection from Social Media Reviews Using GPT-Neo. In *Doctoral Symposium on Human Centered Computing*, pp. 235-245. Singapore: Springer Nature Singapore, 2023.

[10] Dissanayake, G.S., Shelton Peiris, and Tommaso Proietti. State space modeling of Gegenbauer processes with long memory. *Computational Statistics & Data Analysis,* **100,** 115-130, 2016.

[11] Dowe, D. L., Foreword re C S Wallace. *Computer Journal*, **51 (5)**:523–560, 2008a.

[12] Dowe, D. L., Minimum Message Length and Statistically Consistent Invariant (Objective?) Bayesian Probabilistic Inference—From (Medical) "Evidence". *Social Epistemology*, **22 (4)**:433–460, 2008b.

[13] Dowe, D.L., MML, hybrid Bayesian network graphical models, statistical consistency, invariance and uniqueness, *Handbook of the Philosophy of Science - (HPS Volume 7) Philosophy of Statistics*, P.S. Bandyopadhyay and M.R. Forster (eds.), Elsevier. [ISBN: 978-0-444-51862-0 ISBN 10: 0-444-51542-9 / ISBN 13: 978-0-444-51862-0], pp. 901-982, 2011.

[14] Dowe, D. L. Introduction to Ray Solomonoff 85th memorial conference. In *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence. Lecture Notes in Artificial Intelligence (LNAI)* **7070**, pages 1–36. Springer, 2013.

[15] Dowe, D. L., Is Stephen Hawking right? Could AI lead to the end of humankind? https://theconversation.com/is-stephen-hawking-right-could-ai-lead-to-the-end-of-humankind-34967 *The Conversation*, 4 December 2014.

[16] Dowe, D. L., Gardner, S. and Oppy, G R. Bayes not Bust! Why Simplicity is no Problem for Bayesians. *British J for the Philosophy of Science.* vol. **58**, no. **4**, pp 709-754, *University of Chicago Press*, 2007.

[17] Dowe, David L., and Alan R. Hajek. A computational extension to the Turing Test. In *Proceedings of the 4th conference of the Australasian cognitive science society*, University of Newcastle, NSW, Australia, vol. **1**, no. **3**, 6pp. September, 1997a.

[18] Dowe, David L., and Alan R. Hajek. A computational extension to the Turing Test. *Technical Report #97/322, Dept of Computer Science, Monash University*, Clayton (Melbourne), Vic., Australia, 9pp, September, 1997b.

[19] Dowe, David L., and Alan R. Hajek. A non-behavioural, computational extension to the Turing Test. In *Intl. Conf. on Computational Intelligence & multimedia applications (ICCIMA'98).* Gippsland, Australia, pp. 101-106. 1998.

[20] Duan, L. and L. Xu, Business Intelligence for Enterprise Systems: A Survey, *IEEE Transactions on Industrial Informatics,* **8,** pp. 679-687, 2012, doi: 10.1109/TII.2012.2188804.

[21] Fang, Zheng, David L. Dowe, Shelton Peiris, and Dedi Rosadi. Minimum message length in hybrid ARMA and LSTM model forecasting. *Entropy,* **23(12),** 1601, 2021.

[22] Fang, Z., Jianying Xie, Ruiming Peng, and Sheng Wang. Climate finance: Mapping air pollution and finance market in time series. *Econometrics,* **9(4),** 43, 2021.

[23] Fitzgibbon, Leigh J., Lloyd Allison, and David L. Dowe. Minimum message length grouping of ordered data. In *Proc. Algorithmic Learning Theory: 11th International Conference (ALT 2000)*, Sydney, Australia, December 11–13, 2000, **11**, pp. 56-70. Springer Berlin Heidelberg, 2000.

[24] Fitzgibbon, L. J., D. L. Dowe, and L. Allison. Change-point estimation using new minimum message length approximations. In *PRICAI 2002: Trends in Artificial Intelligence:*

7th Pacific Rim International Conference on Artificial Intelligence Tokyo, Japan, August 18–22, 2002 Proceedings 7, pages 244–254. Springer, 2002.

[25] Fitzgibbon, L. J., D. L. Dowe, and F. Vahid. Minimum message length autoregressive model order selection. In *Proc. International Conference on Intelligent Sensing and Information Processing*, pp. 439-444. *IEEE*, 2004.

[26] Franco, Glaura C., Gustavo C. Lana, and Valderio A. Reisen. Prediction intervals in the ARFIMA model using Bootstrap G. *Financial Statistical J*, **1(3)**, 1-8, 2018.

[27] Grace, Katja, Harlan Stewart, Julia Fabienne Sandkühler, Stephen Thomas, Ben Weinstein-Raun, and Jan Brauner. Thousands of AI authors on the future of AI. https://aiimpacts.org/wp-content/uploads/2023/04/Thousands_of_AI_authors_on_the_future_of_AI.pdf, *AI Impacts*, January 2024, accessed 10/Jan/2024.

[28] Granger, C. W. J. and R. Joyeux. An introduction to long-memory time series models and fractional differencing. *J Time Series Analysis*, **1(1)**, 15-29, 1980.

[29] Hermansah, Dedi Rosadi, Herni Utami, Abdurakhman, and Gumgum Darmawan. Hybrid MODWT-FFNN model for time series data forecasting. *AIP Conference Proceedings,* vol. **2192**, page 090005. AIP Publishing LLC, 2019.

[30] Hernández-Orallo, José. Twenty years beyond the Turing test: moving beyond the human judges too. *Minds and machines.* **30**, no. **4**: 533-562, 2020.

[31] Hewamalage, H., Bergmeir, C., & Bandara, K. (2021). Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, **37**(**1**), 388-427

[32] Hochreiter, S. and J. Schmidhuber. Simplifying neural nets by discovering flat minima. *Advances in neural information processing systems*, 7, 1994.

[33] Khandelwal, I., Ratnadip Adhikari, and Ghanshyam Verma. Time series forecasting using hybrid ARIMA and ANN models based on DWT decomposition. *Procedia Computer Science,* **48,** 173-179, 2015.

[34] Kumar, Bhupendra, Sunil, and Neha Yadav. A novel hybrid model combining βSARMA and LSTM for time series forecasting. *Applied Soft Computing*, **134**, 2023, 110019.

[35] Ling, S., and Wai Keung Li. On fractionally integrated autoregressive moving-average time series models with conditional heteroscedasticity. *Journal of the American Statistical Association,* **92(439),** 1184-1194, 1997.

[36] Makalic, E., L. Allison, and D. L. Dowe. MML inference of single-layer neural networks. In *Proc. 3rd IASTED Int. Conf. Artificial Intelligence and Applications, September*, pages 636–642, 2003.

[37] Marques-Silva, Joao, and Alexey Ignatiev. Delivering Trustworthy AI through formal XAI. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. **36**, no. **11**, pp. 12342-12350. 2022

[38] Mathew, Mariam T., Melanie Babcock, Ying-Chen Claire Hou, Jesse M. Hunter, Marco L. Leung, Hui Mei, Kathleen Schieffer, and Yassmine Akkari. *Clinical Cytogenetics: Current Practices and Beyond.* The Journal of Applied Laboratory Medicine **9**, no. **1**, 2024: 61-75.

[39] Mikosch, T. and Catalin Starica. Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects. *Review of Economics and Statistics,* **86(1),** 378-390, 2004.

[40] Neath, Andrew and Joseph E. Cavanaugh. The Bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics,* **4(2),** 199-203, 2012.

[41] Necula, C. and Alina-Nicoleta Radu. Long memory in eastern European financial markets returns. *Economic Research-Ekonomska istrazivanja,* **25(2),** 316-377, 2012.

[42] Needham, S. and D. L. Dowe, Message length as an effective Ockham's razor in decision tree induction, *International Workshop on Artificial Intelligence and Statistics,* pp. 216-223, 2001.

[43] O'Sullivan, M. The untold story of QF72: What happens when 'psycho' automation leaves pilots powerless? *Sydney Morning Herald.* https://www.SMH.com.au/lifestyle/the-

untold-story-of-qf72-what-happens-when-psycho-automation-leaves-pilots-powerless-20170511-gw26ae.html (accessed 4/July/2023), 12 April 2017.

[44] Open Power System Data. 2020. Data Package Time series. Version 2020-10-06, accessed 2023. https://doi.org/10.25832/time_series/2020-10-06. Used timeseries 30 min.csv in the package and selected GB_GBN_load_actual_entsoe_transparency (Total load in Great Britain in MW as published on ENTSO-E Transparency Platform) as the variable to do the modelling on.

[45] Palma, W. and Ngai Hang Chan. Efficient estimation of seasonal long-range-dependent processes. *Journal of Time Series Analysis,* **26(6),** 863-892, 2005.

[46] Papageorgiou, I. and I. Kontoyiannis. Context-tree weighting for real-valued time series: Bayesian inference with hierarchical mixture models. *2023 IEEE International Symposium on Information Theory (ISIT).* Taipei, Taiwan, pp. 2464-2469, 2023.

[47] Petzka, H., M. Kamp, L. Adilova, C. Sminchisescu, and M. Boley. Relative flatness and generalization. *Advances in neural information processing systems*, **34**:18420–18432, 2021.

[48] Sanghi, Pritika and D. L. Dowe. A computer program capable of passing IQ tests. In *4th Intl. Conf. on Cognitive Science (ICCS'03)*, Sydney, pp. 570-575. 2003.

[49] Schmidt, D. F., Minimum message length inference of autoregressive moving average models (PhD. thesis, 2008). Faculty of IT, Monash University.

[50] Schwarz, Gideon. Estimating the dimension of a model. *The annals of statistics*: 461-464, 1978.

[51] Solomonoff, R. J. A formal theory of inductive inference. Part I. *Information and control*, **7 (1)**:1–22, 1964a.

[52] Solomonoff, R. J. A formal theory of inductive inference. Part II. *Information and control*, **7 (2)**:224–254, 1964b.

[53] Solomonoff, R. J. Inductive inference research: status, *RTB 154, Rockford Research, Inc.*, 140 1/2 Mt. Auburn St., Cambridge, Mass. 02138, USA, Spring 1967.

[54] Solomonoff, R. J. Complexity-based induction systems: comparisons and convergence theorems. *IEEE Trans. on Information Theory*, **24 (4)**:422–432, 1978.

[55] Solomonoff, R. J. Does algorithmic probability solve the problem of induction? In *Proc Information, Statistics and Induction in Science Conference, Melbourne, Australia, August 1996*, pages 7–8. World Scientific. ISBN 981-02-2824-4. 1996.

[56] Solomonoff, R. J. Does algorithmic probability solve the problem of induction? *Oxbridge Research, POB*, 391887, 1997.

[57] Solomonoff, R. J. Progress in incremental machine learning. In *NIPS Workshop on Universal Learning Algorithms and Optimal Search, Whistler, BC.* 2002. (Also Technical Report IDSIA-16-03, IDSIA, Lugano, Switzerland, 2003.)

[58] Srivastava, Muni and Tatsuya Kubokawa. Akaike information criterion for selecting components of the mean vector in high dimensional data with fewer observations. *Journal of the Japan Statistical Society,* **38(2),** 259-283, 2008.

[59] Tan, P. J. and D. L. Dowe. MML inference of decision graphs with multi-way joins and dynamic attributes. In *Australian Conference on Artificial Intelligence*, pages 269–281, 2003.

[60] Tan, Peter J., and David L. Dowe. Decision forests with oblique decision trees. In *Mexican International Conference on Artificial Intelligence*, pp. 593-603. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.

[61] Visser, Gerhard, David L. Dowe, and Petteri Uotila. Enhancing MML clustering using context data with climate applications. In *Australasian Joint Conference on Artificial Intelligence*, pp. 350-359. *Springer*, Berlin, Heidelberg, 2009.

[62] Wallace, C. S., *Statistical and inductive inference by minimum message length,* Springer, New York, 2005.

[63] Wallace, C. S. and D. M. Boulton, An information measure for classification, *The Computer Journal,* **11(2),** 185-194, 1968.

[64] Wallace, C. S. and D. L. Dowe, Minimum message length and Kolmogorov complexity, *The Computer Journal,* **42(4),** 270-283, 1999a.

[65] Wallace, C. S. and P. R. Freeman. Estimation and inference by compact coding. *J Royal Statistical Society: Series B (Methodological)*, **49 (3)**:240–252, 1987.

[66] Wang, Xiaoyue, Abdullah Mueen, Hui Ding, Goce Trajcevski, Peter Scheuermann, and Eamonn Keogh. Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery,* **26(2),** 275-309, 2013.

[67] Xie, Wei, Yongjin Zhu, Weiqing Cao, Jinfang Pan, Baotai Wu, Shangdong Liu, and Yimu Ji. PCA-LSTM Anomaly Detection and Prediction Method Based on Time Series Power Data. In *2022 China Automation Congress (CAC)*, pp. 5537-5542. *IEEE*, 2022.

[68] Xu, M., G. R. Tynan, P. H. Diamond, P. Manz, C. Holland, N. Fedorczak, S. Chakraborty Thakur, J. H. Yu, K. J. Zhao and J. Q. Dong. Frequency-resolved nonlinear turbulent energy transfer into zonal flows in strongly heated l-mode plasmas in the HL-2A Tokamak. *Physical Review Letters*, **108(24)**, 245001, 2012.

[69] Yang, Y., Chong Jun Fang & HongLin Xiong, A novel general-purpose hybrid model for time series forecasting, *Applied Intelligence,* **52,** pp 2212-2223 (2022) https://doi.org/10.1007/s10489-021-02442-y

[70] Zhang, S., Guo, B., Dong, A., He, J., Xu, Z. & Chen, S.X. (2017). Cautionary Tales on Air-Quality Improvement in Beijing. *Proceedings of the Royal Society A*, Volume **473**, No. 2205, Pages 20170457. Original Link (But broken): https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data alternative link: Beijing Multi-Site Air-Quality Data Set, Kaggle used the file PRSA_Data_Gucheng_20130301-20170228.csv, PM10

[71] Zhou, Luyu, Chun Zhao, Ning Liu, Xingduo Yao, and Zewei Cheng. Improved LSTM-based deep learning model for COVID-19 prediction using optimized approach. *Engineering applications of artificial intelligence* **122**, 2023, 106157