ARTICLE IN PRESS

International Journal of Forecasting xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Deep Probabilistic Koopman: Long-term time-series forecasting under periodic uncertainties

Alex T. Mallen [a,*], Henning Lange [b], J. Nathan Kutz [b]

[a] Paul G. Allen School of CSE, University of Washington, 185 E Stevens Way NE, Seattle, WA 98195, United States
[b] Department of Applied Mathematics, University of Washington, 4182 E Stevens Way NE, Seattle, WA 98195, United States

ARTICLE INFO

Keywords:
Koopman theory
Long term forecasting
Neural networks
Probability forecasting
GEFCom
Atmospheric chemistry forecasting
Energy forecasting
Exploratory data analysis
Seasonality
Electricity

ABSTRACT

This paper introduces general mathematical techniques for stable long-term forecasts with calibrated uncertainty measures. For most time series models, the difficulty of obtaining accurate probabilistic future time step predictions increases with the prediction horizon. We propose a surprisingly simple class of models that characterizes time-varying distributions and enables reasonably accurate predictions thousands of time steps into the future. This technique, called Deep Probabilistic Koopman (DPK), is based on recent advances in linear Koopman operator theory and does not require time stepping for future time predictions. We demonstrate the long-term forecasting performance of these models on a diversity of domains, including electricity demand forecasting, atmospheric chemistry, and neuroscience. Our domain-agnostic technique outperforms all 177 domain-specific competitors in the most recent Global Energy Forecasting Competition for electricity demand modelling.

Published by Elsevier B.V. on behalf of International Institute of Forecasters.

1. Introduction

For many disciplines, such as meteorology, atmospheric chemistry, seismology, epidemiology, power systems, etc., long-term and probabilistic predictions of complex phenomena are paramount. In this paper, we leverage recent advances in operator theory, specifically Koopman theory (Brunton, Budišić, Kaiser, & Kutz, 2021; Mezić, 2005; Mezic, 2013) to introduce a surprisingly simple class of models that allows for accurate predictions thousands of time steps into the future. Koopman's theory establishes that any non-linear dynamical system can be lifted by means of a non-linear operator into a space (usually referred to as the observable space) in which its time evolution is linear (Koopman, 1931; Koopman & Neumann, 1932; Nathan Kutz, Proctor, & Brunton, 2018; Williams, Kevrekidis, & Rowley, 2015). It can be understood as the time-dependent analogue to Cover's theorem (Cover, 1965), arguably the theoretical underpinning of deep learning (LeCun, Bengio, & Hinton, 2015) and Kernel techniques (Scholkopf & Smola, 2001).

Let $f(t)$ be a measurement of a non-linear dynamical system at discrete time t. Koopman theory postulates that there always exists an infinite dimensional linear operator $\mathcal{K}$ and a non-linear operator ψ such that $\mathcal{K}\psi(f(t)) = \psi(f(t+1))$. The linear operator $\mathcal{K}$ is often called the Koopman operator. There are well-known examples of finite-dimensional operators with the same property, spanning what is known as a Koopman-invariant subspace (Brunton et al., 2021).

This has wide application in data-driven dynamics because it leads to more straightforward and general models. Typically, the difficulty is finding the non-linear observables operator ψ that describes the change of variables needed to obtain linear latent dynamics. Once the Koopman-invariant subspace has been found, methods that would usually involve time-stepping the non-linear dynamics can utilize linearity to find a closed-form for $f(t)$

* Corresponding author.
E-mail addresses: atmallen@uw.edu (A.T. Mallen), helange@uw.edu (H. Lange), kutz@uw.edu (J.N. Kutz).

https://doi.org/10.1016/j.ijforecast.2023.07.001
0169-2070/Published by Elsevier B.V. on behalf of International Institute of Forecasters.

by diagonalization. This puts many algorithms in a new complexity class.

Algorithmic approaches to Koopman theory give a global view of the dynamics, allowing for long-range forecasts. They usually try to estimate a low dimensional Koopman operator (or, to be more precise, a finite-dimensional linear operator that spans a Koopman-invariant subspace[1]) and a corresponding time-invariant non-linear function that projects into a subspace of the observables. However, the resulting optimization objective is difficult to solve because it is challenging to train Recurrent Neural Networks (RNNs) because of 'vanishing and exploding gradients'. When optimizing an RNN by truncated Taylor-series (i.e., backpropagation), the error flows back in time. It is repeatedly multiplied by the Jacobian associated with the recurrent connections of the network, which entails that the Jacobian Vector Product associated with the error at time $t + h$ propagated back to time t is $J^h e_{t+h}$. Depending on the eigenvalues of J, this quantity either collapses into the origin or explodes to $\pm\infty$ in the limit of h. The most common strategy to mitigate this effect is to choose a special RNN topology that results in a 'constant carousel of errors' (Hochreiter, 1991) most often by employing multiplicative gates as, for example, employed Long Short-Term Memories and Gated Recurrent Units (Chung, Gulcehre, Cho, & Bengio, 2014; Hochreiter & Schmidhuber, 1997). However, such a strategy cannot be used for Koopman-based approaches because it would violate the linearization constraint.

More recently, an approach called 'Koopman Forecast' (KF) was introduced that overcomes the problem of 'vanishing and exploding gradients' by employing the Fourier-series in conjunction with the Fast Fourier Transform (FFT) as opposed to Taylor-series as the underlying optimization engine to obtain the global optimum of the parameters that govern temporal dependencies (Lange, Brunton, & Kutz, 2021). This approach relies on the assumption that the associated Koopman operator has strictly imaginary eigenvalues, which assumes that the phenomenon we wish to model is stable and therefore exhibits quasi-periodic behaviour. As we will show later, KF leverages the periodicities of local losses to reconstruct the global loss.

In this paper, we extend this approach to the probabilistic setting. Any extension to the KF algorithm needs to ensure that the requirements for solving the global objective are kept intact, i.e., local losses exhibit periodicities that allow for the reconstruction of the global error. The techniques introduced in this paper ensure this by making a simple assumption: We assume that a quasi-periodically varying distribution can characterize the measurement of the dynamical system. This allows us to obtain individual KF models for the distribution parameters by maximizing the data likelihood and creating long-term probabilistic forecasts. The assumption of quasi-periodically varying probability distributions implies that quantities such as uncertainty exhibit quasi-periodic patterns. We empirically validate this assumption for signals in power systems and atmospheric chemistry.

[1] For the remainder of this paper, we will refer to any such operator as a Koopman operator.

2. Related work

The main contributions of our method are to robustly and probabilistically model seasonal time series using the capabilities of neural networks. The techniques introduced in this paper are general and therefore are, in some shape or form, related to time series models such as LSTMs (Hochreiter & Schmidhuber, 1997), Gated Recurrent Units (GRUs) (Chung et al., 2014), N-BEATS (Oreshkin, Carpo, Chapados, & Bengio, 2019), Neural and Gaussian processes (Garnelo, Rosenbaum et al., 2018; Garnelo, Schwarz et al., 2018; Williams & Rasmussen, 1996) and temporal convolutional networks (TCNs) (Bai, Kolter, & Koltun, 2018; Sen, Yu, & Dhillon, 2019). While these methods use deep learning to model the time series, they differ. Unlike DPK, which models global quasi-periodic patterns, LSTMs and GRUs are local recurrent models that lose accuracy for longer time horizons. N-BEATS is a neural method that includes frequency components to model seasonalities, which makes it among the most similar methods to DPK. However, it does not provide probabilistic models and involves a costly training and inference procedure. Because our technique assumes internal frequency and periodicity notions, Clockwork-RNNs are a closely related neural approach (Koutnik, Greff, Gomez, & Schmidhuber, 2014). Other works aim to model seasonality by decomposing time series into seasonal and trend components (Wen, Gao, Song, Sun, Xu, & Zhu, 2019; Yang, Wen, Yang, & Sun, 2021) before applying neural methods. In DPK, no such decomposition occurs before neural learning. Recent work on the deep reconstruction of strange attractors shares the goal of uncovering intrinsic quasi-periodic structures. Still, it assumes non-linear latent dynamics whilst our approach assumes linear latent dynamics (Gilpin, 2020). State space models (SSMs) are similar in character to our work, except that for SSMs, the latent space itself is stochastic, limiting their long-term forecasting ability. Recent work on Deep State Space models has extended SSMs to incorporate RNNs as a model of the latent dynamics (Rangapuram, Seeger, Gasthaus, Stella, Wang, & Januschowski, 2018). Conceptually, this work could also be seen as a special case of stochastic NeuralODEs (Chen, Rubanova, Bettencourt, & Duvenaud, 2018; Li, Wong, Chen, & Duvenaud, 2020) with linear latent dynamics. However, because of the linear latent dynamics, training is significantly easier and does not require the adjoint method or potentially expensive time steppers. The reader is referred to Hamilton (2020) for a recent review of different time series prediction techniques.

The finding that any non-linear dynamical system can be linearized globally was first introduced by Koopman in 1931 for Hamiltonian systems (Koopman, 1931) and was later generalized to continuous-spectrum systems (Koopman & Neumann, 1932). At the time, it was of considerable importance as a building block for advances in ergodic theory (Birkhoff, 1931; Birkhoff & Koopman, 1932; Moore, 2015; Neumann, 1932a, 1932b). Recently, interest in Koopman theory was renewed by work by Budišić, Mohr, and Mezić (2012), Mezić (2005) and Mezic (2013) and the development of *Dynamic Mode Decomposition* (Schmid, 2010) (DMD) which provided a cheap

and easy computational approximation to the Koopman operator (Rowley, Mezić, Bagheri, Schlatter, & Henningson, 2009). DMD was first introduced as a technique for modal decomposition to extract the dynamic behaviour in fluid flows but has since been applied in various fields, such as neuroscience (Brunton, Johnson, Ojemann & Kutz, 2016), epidemiology (Proctor & Eckhoff, 2015), acoustics (Song, Alizard, Robinet, & Gloerfelt, 2013), combustion modelling (Moeck, Bourgouin, Durox, Schuller, & Candel, 2013) and video processing (Erichson, Brunton, & Kutz, 2019). To be precise, DMD approximates the Koopman operator when the observables are restricted to direct measurements of the state variable (Brunton, Brunton, Proctor & Kut, 2016; Brunton et al., 2021; Kutz, Brunton, Brunton, & Proctor, 2016; Mezic, 2013; Rowley et al., 2009; Tu, Rowley, Luchtenburg, Brunton, & Kutz, 2014).

Many algorithms exist that attempt to estimate the Koopman operator from data. Most of these approaches rely on auto-encoder structures (Lusch, Kutz, & Brunton, 2018; Otto & Rowley, 2019; Takeishi, Kawahara, & Yairi, 2017; Wehmeyer & Noé, 2018; Yeung, Kundu, & Hodas, 2019) and solve an optimization objective that encourages linearity in 'Koopman space' and prediction accuracy. These approaches have been extended in various ways, e.g., the authors of Pan and Duraisamy (2019) extended Koopman theory to the probabilistic setting by employing Bayesian Neural Networks as encoders, whereas in Champion, Lusch, Kutz, and Brunton (2019), the linearity constraint was relaxed and substituted for sparsity in the latent dynamics. These approaches usually do not consider linearity in 'Koopman space' a constraint but rather the optimization objective and typically result in latent dynamics that are only approximately and locally linear.

In this paper, we build upon recent advancements that do away with the auto-encoder structure and enforce linearity and stability of the latent dynamics by construction (Lange et al., 2021; Mendible, Koch, Lange, Brunton, & Kutz, 2020).

3. Probabilistic koopman forecast

As discussed earlier, the class of models introduced in this work build upon the KF algorithm (Lange et al., 2021). The KF algorithm exploits the fact that the linear latent dynamics of any stable, non-linear dynamical system with finite-dimensional discrete Koopman operator with unit magnitude complex eigenvalues can be expressed as a vector of sines and cosines. In Koopman-theoretic terms, the non-linear operator ψ is the neural network, and the linear Koopman operator $\mathcal{K}$ is the conceptual operator that takes the vector of sinusoids at time t to time $t + \mathrm{d}t$. We never need to express this linear operator because the frequencies of the sinusoids imply it. Fitting the KF algorithm to data points $x_t \in \mathbb{R}^n$ distributed uniformly over time, therefore, requires solving the following optimization objective to search for $\vec{\omega}$:

$$\mathcal{L}(\Theta, \vec{\omega}) = \sum_{t=1}^{T} \left(g_\Theta \left(\begin{bmatrix} \cos(\vec{\omega} t) \\ \sin(\vec{\omega} t) \end{bmatrix} \right) - x_t \right)^2 = \sum_{t=1}^{T} L(\Theta, \vec{\omega}, t) \tag{1}$$

with g usually being a NN describing the inverse of the observables ψ with parameters Θ. The dimensionality of $\vec{\omega}$ tends to be surprisingly small, i.e. less than five or six. Even though this objective is non-convex and non-linear in $\vec{\omega}$, and therefore not amenable to optimization with stochastic gradient descent (SGD), it nevertheless can *and needs to* be solved globally. This is achieved by exploiting periodicities in $L(\Theta, \vec{\omega}, t)$, specifically that L is periodic with period $2\pi/t$ with respect to each frequency ω_i since $L(\Theta, \vec{\omega}, t) = L(\Theta, \vec{\omega} + 2\pi/t, t)$. Note that the period of $2\pi/t$ is with respect to frequency (ω_i), rather than time (t). In conjunction with the FFT, this realization allows for the global reconstruction of $\mathcal{L}$ (Lange et al., 2021). Thus, the KF can be described as a technique to fit a non-linear combination of sinusoids to data, producing a model that makes a stable forecast of future data. Note that unlike $\vec{\omega}$, Θ is optimized using SGD. Searching for Θ can be understood as searching for the observables such that the dynamics are linear. We direct the readers to Lange et al. (2021) for a more comprehensive discussion of optimizing to search for $\vec{\omega}$.

This work extends this technique to allow for probabilistic forecasting for systems with periodically varying uncertainty. Specifically, we assume that the data is drawn from a time-varying but, if conditioned on model parameters, the independent random variable X_t is distributed as P with parameters exhibiting quasi-periodic temporal patterns. We then task a DPK model to predict these parameters forward in time. Let θ_t denote the parameters of P, and $p(x_t|\theta_t) = P(X_t = x_t|\theta_t)$ denote X_t's probability density or mass under P. For example, if we take P to be a Gaussian, then $X_t \sim \mathcal{N}(\theta_t = \{\mu(t), \sigma(t)\})$. The assumptions of the respective algorithms can be summarized as follows:

$$\underbrace{x_t = g_\Theta \left(\begin{bmatrix} \cos(\vec{\omega} t) \\ \sin(\vec{\omega} t) \end{bmatrix} \right)}_{\text{Koopman Forecast}} \tag{2}$$

$$\underbrace{x_t \sim P \left(X_t \mid \theta_t = g_\Theta \left(\begin{bmatrix} \cos(\vec{\omega} t) \\ \sin(\vec{\omega} t) \end{bmatrix} \right) \right)}_{\text{DPK}} \tag{3}$$

More concretely, we train the parameters Θ of the NN g_Θ to minimize the negative log-likelihood of observing the data x_t given our model. In practice, a separate NN is used to fit each distribution parameter.

Unlike KF, instead of minimizing the squared error, we minimize the negative log-likelihood:

$$\mathcal{L}_P(\Theta, \vec{\omega}) = \sum_{t=1}^{T} -\log p \left(x_t | g_\Theta \left(\begin{bmatrix} \cos(\vec{\omega} t) \\ \sin(\vec{\omega} t) \end{bmatrix} \right) \right) \tag{4}$$

$$= \sum_{t=1}^{T} L_P(\Theta, \vec{\omega}, t) \tag{5}$$

Note that independent of the choice of P, $L_P(\Theta, \vec{\omega}, t) = L_P(\Theta, \vec{\omega} + 2\pi/t, t)$ which in turn means that the 'trick' of the KF algorithm to reconstruct $\mathcal{L}$ is applicable independent of the assumption of the underlying distribution. In general, our framework can be used as long as the density or mass function of the distribution P is differentiable with respect to the distribution parameters.

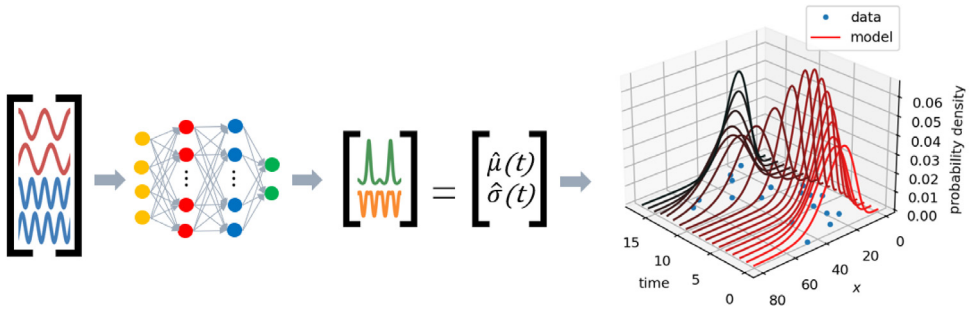

Fig. 1. Left: an overview of our model when the data is assumed to be normally distributed. Right: our model recovers a probability distribution for every point in time, given only a single realization from each distribution, and then projects into the future.

By assuming that the parameters evolve according to a non-linear function of sinusoids, we extract correlations between measurements that are a) adjacent in time because we assume g to be smooth and b) multiples of one period apart. This constraint allows us to faithfully estimate a probability distribution for every point in time given only a single realization from each distribution and then extrapolate that into the future as shown in Fig. 1.

In the simplest case, if we assume P to be a time-varying Gaussian—i.e. $X_t \sim \mathcal{N}(\mu(t), \sigma(t))$, the loss becomes

$$\mathcal{L}_{\mathcal{N}} = \sum_{t=1}^{T} \left(\frac{(x_t - \hat{\boldsymbol{\mu}}(t))^2}{2\hat{\boldsymbol{\sigma}}^2(t)} + \ln(\hat{\boldsymbol{\sigma}}(t)) \right). \tag{6}$$

This bears a strong resemblance to the ordinary mean-square error, except that squared errors are weighted inversely by $\hat{\sigma}^2(t)$, and there is an added penalty for increasing $\hat{\boldsymbol{\sigma}}(t)$. Because $\hat{\mu}(t)$ and $\hat{\sigma}(t)$ are constrained to be quasi-periodic, the model is only suitable when $\mu(t)$ and $\sigma(t)$ are approximately quasi-periodic.

4. Experiments

To optimize $\mathcal{L}$ with respect to the NN parameters, we use SGD with a learning rate tuned to 10^{-4} and a weight decay regularization of 10^{-3}. The data is normalized to zero mean and unit range for most experiments before passing through DPK, except for discrete distributions and the synthetic experiments below.

The technique to extract angular frequencies ω_i described in Lange et al. (2021), because of the *Unknown Phase Problem*, can often lead to numerical instabilities and, therefore, significant variance during training. To avoid this, for the natural data experiments, we hand-pick and fix ω_i. Thus, the natural data experiments probe the effectiveness of modelling complex phenomena by NNs driven by sinusoids rather than the KF's ability to extract these frequencies from data. The seasonalities are known in many time series applications, including those investigated here. In what follows, we show the effectiveness of DPK applied to various synthetic and natural data, including electricity demand and atmospheric pollution. Our experiments do not focus on short-term, non-quasi-periodic time series because we do not expect to contribute here.

4.1. Synthetic experiments

4.1.1. Recovering time-varying distribution parameters: Gaussian and gamma

We generate two sets of synthetic data to test our model's ability to recover probability distributions whose parameters evolve according to arbitrary non-linear but quasi-periodic functions of time. We conduct experiments on Gaussian- and gamma-distributed data. The gamma distribution is used to model data imbued with a non-negativity constraint and is parameterized by a shape α and scale β variable. Both the Gaussian and the gamma model use a fully-connected NN with a 256-node first hidden layer and a 64-node second hidden layer with *tanh* activation to model each distribution parameter independently. We recover time-varying distribution parameters obeying the following quasi-periodic functions of time:

$$\text{Gaussian: } \mu(t) = 2\sin(1 + \sin(2\pi t/48)) \tag{7}$$

$$\sigma(t) = \exp[\sin(2\pi t/31) - 1] + 1/2 \tag{8}$$

$$\text{Gamma: } \alpha(t) = (\exp[\sin(2\pi t/96)] + \cos(2\pi t/12))^2 + 4 \tag{9}$$

$$\beta(t) = \sin(2\pi t/12)/2 + \cos(2\pi t/96) + 2 \tag{10}$$

We also compare three related neural time-series models: an LSTM, a TCN, and generic N-BEATS. All four models are given 4000 training data points, and we compare the deterministic forecasts to the DPK mean forecast for various forecast horizons using mean absolute error (MAE). More experimental details can be found in the appendix and provided code. Qualitatively, N-BEATS and the TCN can make reasonably accurate forecasts but cannot discern the smoothness of the predictable component of the time series. The LSTM falls out of phase for longer forecast horizons. Plots of the forecasts and DPK's recovered distribution parameters are shown in Fig. 2.

Because LSTMs and TCNs have a probabilistic variant, we provide additional results that compare their probabilistic forecasts to those of DPK. This experiment is again split into 4000 training and 400 test data points, employing the same DPK forecast. The configuration of the LSTM and TCN models is the same as in the point forecasting case, except that we provide them with the appropriate gamma or Gaussian likelihood function. Forecasts are scored using pinball loss, as described in Eq. (11), to be

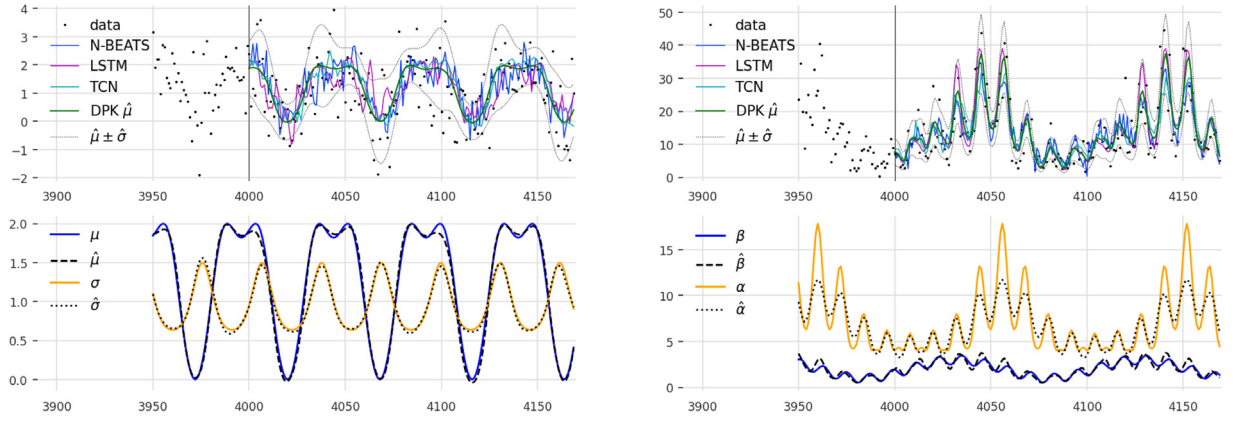

Fig. 2. Top: Forecasts of time-varying Gaussian (left) and gamma (right) distributed data. Bottom: DPK-learned parameters versus the parameters used to generate the data.

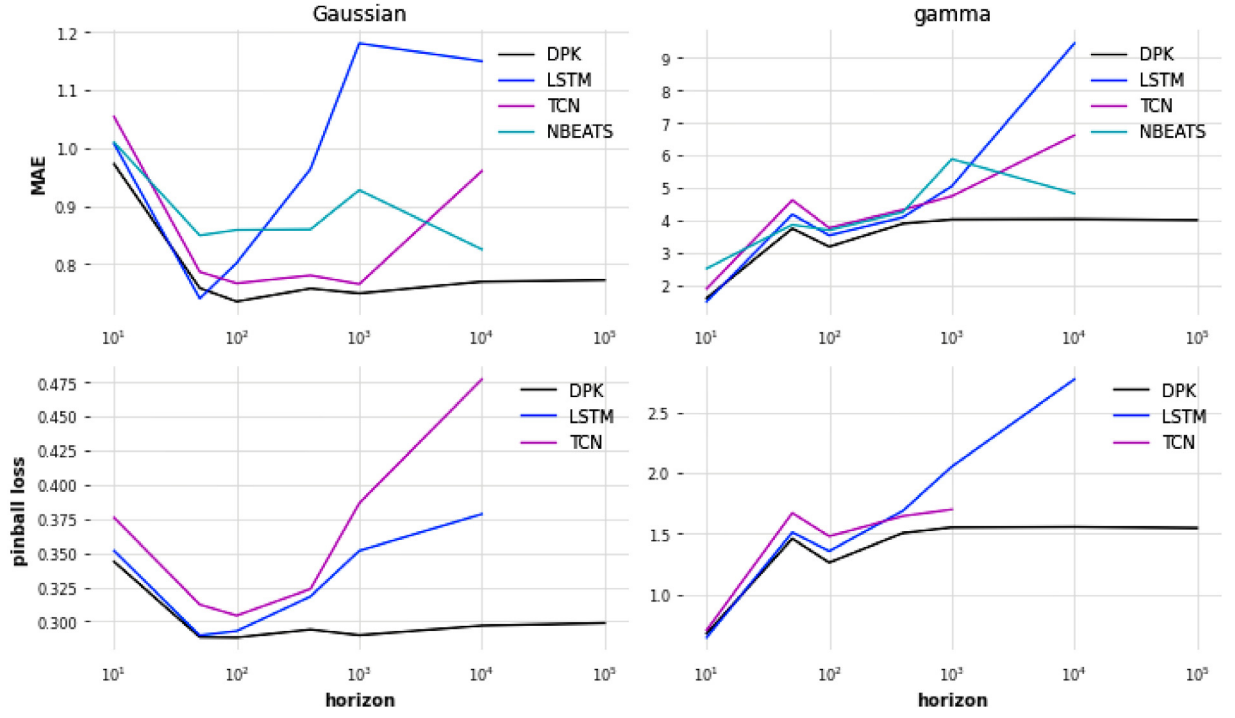

Fig. 3. Performance of various methods in forecasting Gaussian (left) and gamma (right) distributed data. DPK outperforms traditional ML forecasting approaches for long forecast horizons in deterministic and probabilistic cases. DPK can also quickly make long forecasts where autoregressive techniques cannot because DPK does not require time-stepping.

consistent with later experiments. Therefore, this experiment tests the ability of each forecasting technique to recover the time-varying shape of the probability distribution given the distribution family (gamma or Gaussian). We find that DPK outperforms LSTM and TCN for the Gaussian and gamma-distributed data, as shown in Fig. 3.

4.2. Natural data experiments

4.2.1. Global energy forecasting competition

Additionally, we evaluate our model on data from the most recent Global Energy Forecasting Competition (GEFCom) 2017. For each combination of month in 2017 and zone in New England, contestants are tasked with forecasting nine quantiles of the electricity demand for the respective month and zone at an hourly rate. The data is publicly available from ISO New England and can be found in Hong, Xie, and Black (2019).

Competing teams are evaluated relative to a competitive baseline model called the vanilla benchmark (VB). VB is an ensemble of multiple linear regression (MLR) forecasts that account for temporal patterns and the effects of temperature (Hong et al., 2019). Various temperature scenarios are drawn from the historical temperature data to create an ensemble of MLR forecasts, whose empirical distribution induces the uncertainty patterns in the model.

Table 1
Comparison against top 7 of 177 GEFCom 2017 teams.

Team	Description	$\bar{R}_{\text{VB}}$
Tangent Works	*unknown/proprietary*	12.9%
It Can Be Done	Quantile Regression, GBM	11.9%
Orbuculum	Quantile Random Forest, NN, GBM (Smyl & Hua, 2019)	11.5%
Black Analytics	MLR (Hong et al., 2019)	9.1%
Dmlab	XGBoost, QGBRT, Generalized Additive Models	8.4%
Simple But Good	Quantile Regression, MLR (Ziel, 2019)	7.1%
Rain Benchmark	MLR (Hong et al., 2019)	5.7%
	DPK	**15.4%**

Models are compared to VB using pinball loss (Koenker & Hallock, 2001):

$$\text{Pinball}(\hat{x}, q, t) = \begin{cases} (1 - \frac{q}{100})(\hat{x}_{q,t} - x_t) & \text{if } \hat{x}_{q,t} > x_t \\ \frac{q}{100}(x_t - \hat{x}_{q,t}) & \text{otherwise,} \end{cases} \tag{11}$$

with $\hat{x}_{q,t}$ denoting the qth quantile of the respective probabilistic forecast $\hat{x}$ at time t and x_t the observed value, where $q \in \mathcal{Q} = \{10, 20, \ldots, 90\}$. Furthermore, let E be the average of $\text{Pinball}(\hat{x}, q, t)$, i.e.: $E(\hat{x}) = \frac{1}{9T}\sum_{t=1}^{T}\sum_{q\in\mathcal{Q}} \text{Pinball}(\hat{x}, q, t)$. The final score of each competing algorithm is the relative improvement over VB, i.e. $R_{\text{VB}} = (1 - \frac{E}{E_{\text{VB}}})100\%$, with E_{VB} being the error of VB.

We choose and fix $\vec{\omega} = [\frac{2\pi}{24}, \frac{2\pi}{7\times 24}, \frac{2\pi}{365.24\times 24}]$. However, demand also has a non-periodic trend, which can be thought of as a limiting case as $\omega_i \to 0^+$, so we also pass the time t since the start of training into the NN alongside the vector of sinusoids. It should be noted that adding this term empirically does not result in unstable forecasts for long-time horizons (typically approaching a horizontal asymptote instead), as can be seen by our results. We assume energy consumption to be skew-normally distributed. The skew-normal distribution, denoted $\mathcal{S}(\xi, k, \alpha)$, is a generalization of the normal distribution that allows for nonzero skewness. Its PDF is given by

$$f_{X\sim\mathcal{S}}(x; \xi, k, \alpha) = \frac{1}{k\sqrt{2\pi}} e^{-\frac{(x-\xi)^2}{2k^2}} \int_{-\infty}^{\alpha\left(\frac{x-\xi}{k}\right)} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \tag{12}$$

The maximum likelihood loss function is

$$\mathcal{L}_{\mathcal{S}} = \sum_{t=1}^{T} \left(\frac{(x_t - \hat{\xi}(t))^2}{2\hat{k}^2(t)} + \ln(\hat{k}(t)) - \ln\left(\int_{-\infty}^{\hat{\alpha}(t)\left(\frac{x_t - \hat{\xi}(t)}{\hat{k}(t)}\right)} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \right) \right). \tag{13}$$

Strategies to avoid numerical instability in the loss are described in the appendix. The skew-normal model is comprised of 3 independent fully-connected NNs, each with a 256-node first hidden layer and a 64-node second hidden layer (except for the α parameter's NN, which only has 32-s hidden layer nodes) with *tanh* activation.

Competing teams were allowed to exploit historical demand, temperature, and humidity data since 2003 to forecast demand and usually incorporate information such as indicators of whether a day is a holiday. Our model does not use any covariates. We train our model on the years of historical data and produce a 1-month forecast, with a 52-day gap between training and testing as mentioned in Smyl and Hua (2019). We then compute the required nine quantile forecasts for each zone and month. As detailed in the appendix, our training loss function is modified to increase the weight of accurately modelling recent data and data from the time of year being tested.

Table 1 shows the average relative performance $\bar{R}_{\text{VB}}$ of the best seven models globally from the 177 teams competing in the qualifying round versus that of our DPK approach. With a mean relative improvement $\bar{R}_{\text{VB}}$ of 15.4%, DPK outperforms all competing teams. See the appendix for R_{VB} values at each zone and month. The standard deviation across the 108 zone-month pairs is 13.4%. We also computed the standard deviation over ten randomly initialized trials for three randomly selected forecasts, which was 4.33% for Connecticut in December, 1.79% for Northeast Massachusetts in July, and 1.33% for Southeast Massachusetts in October. As described in Hong et al. (2019), competitors used quantile regression, MLR, gradient boosting machines (GBMs), NNs, quantile gradient boosted regression trees (QGBRTs), and others (Smyl & Hua, 2019; Ziel, 2019).

While DPK, as demonstrated, provides state-of-the-art forecasts, it also provides insightful descriptions of uncertainty. As can be seen in Fig. 4a and b, uncertainty is higher on weekdays compared to weekends, and in Summer and Winter than Spring and Fall, while the diurnal pattern depends on the time of year.

Due to time-varying uncertainty, errors can be evaluated by standardizing them using model parameters (i.e. model-standardized residuals $r_t = (x_t - \hat{\mu}(t))/\hat{\sigma}(t)$). We measure the mean and root-mean-square (RMS) of the model-standardized residuals for each of the 108 forecasts as indicators of bias and overconfidence, respectively, with 0 being perfect bias and 1 being perfect RMS. On the training interval, our model-standardized residuals had a mean bias of -0.003 and an RMS of 1.01, while they had a mean test bias of -0.184 and a test RMS of 1.19. This means that the model is 19% overconfident, a sign of slight overfitting to the training data. Month-by-zone mean bias and RMS can be found in the appendix. The distribution of de-skewed model-standardized residuals can be seen in Fig. 4c. De-skewing involves mapping the quantiles of a skew-normal distribution to those of a normal distribution. Mathematically, $z_{\text{de-skewed}} = \text{ICDF}_{\mathcal{N}_{0,1}}(\text{CDF}_{\mathcal{S}_{\hat{\xi},\hat{k},\hat{\alpha}}}(x))$.

4.2.2. Atmospheric pollution

Nitrogen dioxide (NO_2), ozone (O_3), and small particulate matter ($PM_{2.5}$) are three critical atmospheric pollutants that the National Aeronautics and Space Administration (NASA) regularly forecasts in more than 1000 global locations using the Goddard Earth Observing System Composition Forecast (GEOS-CF). The data is publicly

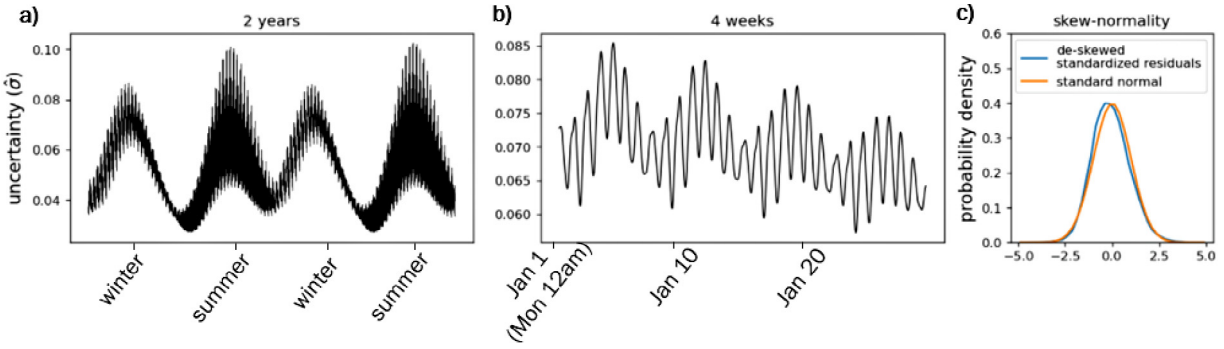

Fig. 4. Periodic trends in uncertainty over time. (a) and (b) demonstrate annual, weekly, and daily periodicities in uncertainty according to our model. (c) demonstrates that our model accurately captures its own uncertainty. Table 1 demonstrates that our model's uncertainty is representative of the uncertainty of the energy forecasting community as a whole.

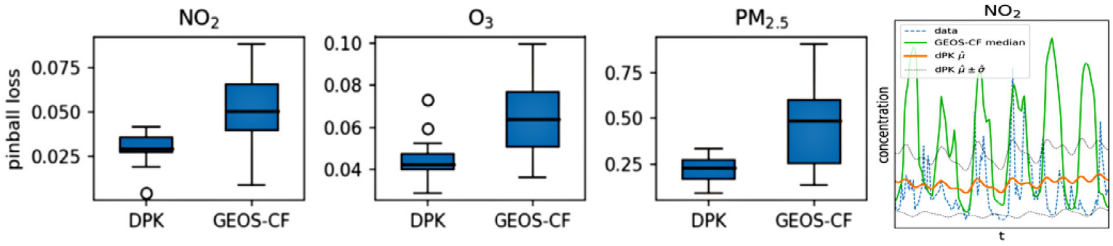

Fig. 5. Comparison against GEOS-CF atmospheric chemistry forecasts for three pollutants and 50 stations.

available at Keller et al. (2021). GEOS-CF uses a model physics package as described in Keller et al. (2021) to simulate the next five days of these pollutants daily. Thus, there are 5 model forecasts for every hour of the day. Although performed at different times, we interpret the ensemble of these predictions as an empirical distribution to allow for probabilistic comparisons.

For this comparison, we take 50 random locations, train our model on the available data from just 2018, and make a forecast for 2019. The observations are recorded hourly, but there are many gaps of up to a few months, so we pass a vector of times t into DPK because x is not uniform over time.

We assume that concentrations of these pollutants follow a gamma distribution. The same gamma model is used as in the synthetic experiments. Note that our forecast horizon is one year, while the GEOS-CF forecast horizon is five days. See Fig. 5 for a comparison of pinball loss scores, showing DPK performs better and more consistently for all three pollutants. In contrast to the GEOS-CF predictions, DPK predictions show a bimodal distribution of NO_2 over the day, once in the morning and again in the evening, suggesting that NO_2 is affected by rush-hour traffic. Thus, DPK could find applications in validating physics-based modelling approaches that often rely on qualitative assumptions of, for example, pollutants emitted by commuter traffic. The appendix contains a more in-depth comparison that shows DPK outperforms GEOS-CF in terms of bias and root-mean-square error and equals it in terms of Pearson correlation when evaluated as a point forecast.

4.3. Exploratory demonstrations

Here, we further demonstrate how our interpretable probabilistic extension of the KF can be used for exploratory data analysis to produce insightful descriptions of time series, such as the chaotic Duffing oscillator and mouse neural data.

4.3.1. A probabilistic description of chaos: Duffing oscillator

As discussed earlier, linearizing chaotic dynamical systems usually requires an infinite dimensional Koopman operator rendering algorithmic applications of Koopman theory difficult. Taking inspiration from ergodic theory and Liouville's theorem (Lutzen, 1985), we demonstrate that the probabilistic description of dynamics introduced in this paper allows for a low-dimensional (in this case 1 dimensional) description of chaos that allows for the prediction of occupancy measures. Because it is driven by a sinusoid, we choose the Duffing oscillator (Wiggins, 1987) as an example. The Duffing equation is a second-order differential equation that describes a family of oscillators varying from periodic to chaotic. It finds applications in modelling oscillators that do not obey Hooke's law (Rychlewski, 1984) and is defined as:

$$\ddot{x} + \delta\dot{x} + \alpha x + \beta x^3 = \gamma \cos(\omega t) \tag{14}$$

We chose parameters in the chaotic regime, ($\omega = 1.4$, $\gamma = 0.375, \delta = 0.1, \alpha = -1, \beta = 1$), and discretize the resulting system spatially by choosing 20 equally sized bins between the minimum and maximum value x_t takes. As the underlying probability distribution, we choose a

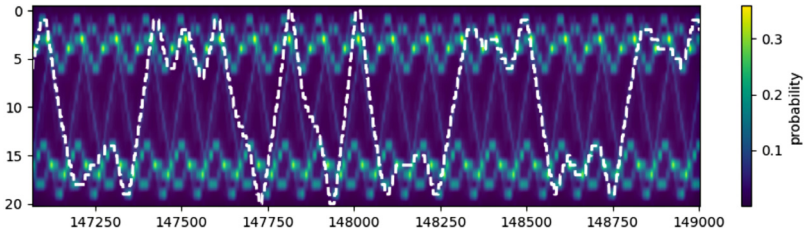

Fig. 6. Probabilistic forecast of the chaotic Duffing oscillator. A brighter background indicates DPK believes the Duffing oscillator is more likely to occupy that region at that time.

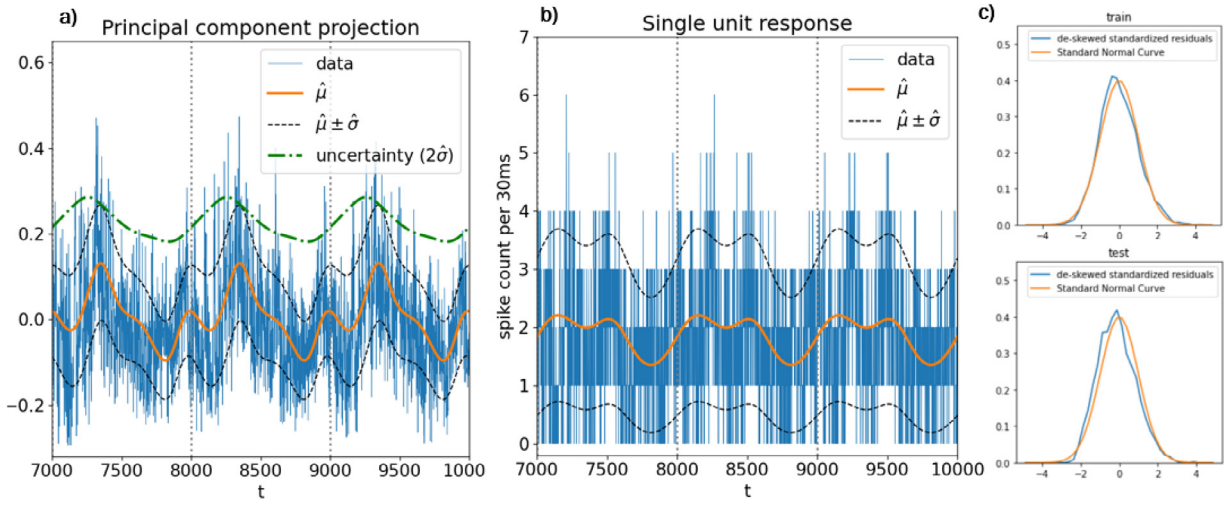

Fig. 7. (a) A forecast of the principal component projection of a mouse's primary visual cortical response to 3 consecutive repetitions of a 30-s video. (b) A forecast of the response of an individual unit (neuron) to the video. (c) Distribution of de-skewed model-standardized residuals on train and test datasets for PCA projected data.

categorical distribution with 20 categories representing the respective bin. The model is trained with a single driving frequency and a *softmax* output with negative log-likelihood loss and a learning rate of 3×10^{-4} on the first 100,000 time steps without regularization. The fully-connected NN uses a 128-node first hidden layer and a 256-node second hidden layer with *tanh* activation. Fig. 6 shows the predictions and a single realization of the system outside the training regime. The plot shows the probability that the system is in state i (or bin i) at time t. DPK successfully extracts non-trivial occupancy patterns from the training data. For example, it was discovered that the occupancy distribution is surprisingly sharp when transitioning between lobes. Even though a point forecast of the system's behaviour cannot be accurate past bifurcation points because of the system's chaotic nature, predicting the behaviour when transitioning between lobes is possible.

4.3.2. Mouse cortical function

To characterize the strength and variability of a mouse's neural response to visual stimuli, we employed our model on recordings from the mouse's primary visual cortex in response to visual stimuli. The data was obtained from the public Allen Brain Observatory visual coding dataset (de Vries et al., 2020). During the experiment, the mouse was presented with a 30-s natural movie looped ten times.[2] The neural data is therefore driven by a 30-s periodicity, which is input to DPK. As a preprocessing step, we perform principal component analysis (PCA) on the firing data of the 60 measured units (one unit approximately corresponds to one neuron, barring measurement uncertainties) to extract correlated patterns of firing and project the 60-dimensional neural firing data onto the first PCA mode. DPK is trained on the first seven presentations of the video with the skew-normal model architecture described in the GEFCom experiment and makes a forecast for the remaining three loops.

Neural firing is often described as a Poisson process for small time windows. We, therefore, model the spike counts of a particular unit using a time-varying Poisson random variable with rate λ in a subsequent experiment. The model is trained on the first seven loops of the video and makes a forecast for the remaining three loops. The model uses a 256-node first hidden layer and a 64-node second hidden layer with *tanh* activation. In Fig. 7a, there are, in fact, strong periodic trends in uncertainty in the principal component projection of the response, which corresponds qualitatively with more dynamic portions of

[2] http://observatory.brain-map.org/visualcoding/stimulus/natural_movies.

the video. The interpretable architecture of DPK, combined with its assumption of quasi-periodically varying distribution parameters, allows for such insights.

5. Conclusions

Many time series, while governed by complex dynamics, are ultimately driven by reliable periodic phenomena such as Earth's rotation and orbit about the sun. This paper introduced a class of models incorporating this simple but powerful assumption: many phenomena can be characterized by a probability distribution whose parameters vary quasi-periodically with time. The resulting technique has a low computational and memory footprint and is easy to implement and understand. Despite the technique's simplicity, it often outperforms domain-specific competitors on important tasks. We empirically show its effectiveness in energy demand forecasting, atmospheric chemistry modelling, and synthetic datasets.

The methods introduced in this paper have a strong inductive bias. While this means they do not work well for data that is not quasi-periodic, the class of quasi-periodic phenomena is vast. Suppose the data is known to be produced by a quasi-periodic system. In that case, a DPK model might be the most appropriate modelling choice because, unlike other neural time series models, they will always produce quasi-periodic predictions. Thus, DPK constitutes a class of 'stiff' models that can fit quasi-periodic phenomena tightly. While it is often hard to reason what a neuronal time series model has learned, DPK models have an interpretable architecture. We demonstrated several examples where DPK's probabilistic predictions could be meaningfully interpreted, producing insights into the quasi-periodic uncertainty of energy demand, rush-hour-related pollution, and occupancy measures of a chaotic dynamical system.

5.1. Future work

Because of the *Unknown Phase Problem*, optimal frequencies can be challenging to learn if they are not already known. Therefore mitigating its effects could improve the applicability of DPK. Another route of potential future work is generalizing DPK to arbitrary distributions by the application of Bayesian approaches (Blei, Kucukelbir, & McAuliffe, 2017; Kingma & Welling, 2013) in conjunction with flexible and scalable approximate posterior distributions parameterized by NNs (Louizos & Welling, 2017; Rezende & Mohamed, 2015). For such an approach, the Bayesian encoder would be driven by sinusoids and map to a probabilistic latent space from which samples would be drawn and fed into the decoder. Such an approach could, in principle, learn arbitrary probability distributions and not require a priori knowledge of the shape of the distributions to model.

Acknowledgments

The authors are grateful to Christoph Keller for providing the GEOS-CF model and observed data. This work was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research (ASCR) under Contract DE-AC02-06CH11347. Pacific Northwest National Laboratory is operated by Battelle for the DOE under Contract DE-AC05-76RL01830.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2023.07.001.

References

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271.

Birkhoff, G. D. (1931). Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences*, *17*(12), 656–660.

Birkhoff, G. D., & Koopman, B. (1932). Recent contributions to the ergodic theory. *Proceedings of the National Academy of Sciences of the United States of America*, *18*(3), 279.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, *112*(518), 859–877.

Brunton, S. L., Brunton, B. W., Proctor, J. L., & Kutz, J. N. (2016). Koopman invariant subspaces and finite linear representations of nonlinear dynamical systems for control. *PLoS One*, *11*(2), Article e0150171.

Brunton, S. L., Budišić, M., Kaiser, E., & Kutz, J. N. (2021). Modern koopman theory for dynamical systems. arXiv preprint arXiv:2102.12086.

Brunton, B. W., Johnson, L. A., Ojemann, J. G., & Kutz, J. N. (2016). Extracting spatial–temporal coherent patterns in large-scale neural recordings using dynamic mode decomposition. *Journal of neuroscience methods*, *258*, 1–15.

Budišić, M., Mohr, R., & Mezić, I. (2012). Applied Koopmanism. *Chaos. An Interdisciplinary Journal of Nonlinear Science*, *22*(4), Article 047510.

Champion, K., Lusch, B., Kutz, J. N., & Brunton, S. L. (2019). Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences*, *116*(45), 22445–22451.

Chen, R. T., Rubanova, Y., Bettencourt, J., & Duvenaud, D. (2018). Neural ordinary differential equations. arXiv preprint arXiv:1806.07366.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, (3), 326–334.

de Vries, S., Lecoq, J., Buice, M., et al. (2020). A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nature Neuroscience*, *23*, 138–151. http://dx.doi.org/10.1038/s41593-019-0550-9.

Erichson, N. B., Brunton, S. L., & Kutz, J. N. (2019). Compressed dynamic mode decomposition for real-time object detection. *Journal of Real-Time Image Processing*, *16*(5), 1479–1492.

Garnelo, M., Rosenbaum, D., Maddison, C., Ramalho, T., Saxton, D., Shanahan, M., et al. (2018). Conditional neural processes. In *International conference on machine learning* (pp. 1704–1713). PMLR.

Garnelo, M., Schwarz, J., Rosenbaum, D., Viola, F., Rezende, D. J., Eslami, S., et al. (2018). Neural processes. arXiv preprint arXiv:1807.01622.

Gilpin, W. (2020). Deep reconstruction of strange attractors from time series. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems, vol. 33* (pp. 204–216). Curran Associates, Inc..

Hamilton, J. D. (2020). *Time series analysis*. Princeton University Press.

Hochreiter, S. (1991). Untersuchungen zu dynamischen neuronalen Netzen [in German] Diploma thesis. *TU Münich*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hong, T., Xie, J., & Black, J. (2019). Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting. *International Journal of Forecasting*, [ISSN: 0169-2070] *35*(4), 1389–1399. http://dx.doi.org/10.1016/j.ijforecast.2019.02.006.

Keller, C. A., Knowland, K. E., Duncan, B. N., Liu, J., anderson, D. C., Das, S., et al. (2021). Description of the NASA GEOS composition forecast modeling system GEOS-CF v1.0. *Journal of Advances in Modeling Earth SystemsS*, *13*(4), http://dx.doi.org/10.1029/2020MS002413.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Koenker, R., & Hallock, K. F. (2001). Quantile regression. *Journal of economic perspectives*, *15*(4), 143–156.

Koopman, B. O. (1931). Hamiltonian systems and transformation in Hilbert space. *Proceedings of the National Academy of Sciences of the United States of America*, *17*(5), 315.

Koopman, B. O., & Neumann, J. v. (1932). Dynamical systems of continuous spectra. *Proceedings of the National Academy of Sciences of the United States of America*, *18*(3), 255.

Koutnik, J., Greff, K., Gomez, F., & Schmidhuber, J. (2014). A clockwork rnn. In *International conference on machine learning* (pp. 1863–1871). PMLR.

Kutz, J. N., Brunton, S. L., Brunton, B. W., & Proctor, J. L. (2016). Dynamic mode decomposition: data-driven modeling of complex systems. SIAM.

Lange, H., Brunton, S. L., & Kutz, J. N. (2021). From fourier to koopman: Spectral methods for long-term time series prediction. *Journal of Machine Learning Research*, *22*(41), 1–38.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436.

Li, X., Wong, T.-K. L., Chen, R. T., & Duvenaud, D. (2020). Scalable gradients for stochastic differential equations. In *International conference on artificial intelligence and statistics* (pp. 3870–3882). PMLR.

Louizos, C., & Welling, M. (2017). Multiplicative normalizing flows for variational bayesian neural networks. In *International conference on machine learning* (pp. 2218–2227). PMLR.

Lusch, B., Kutz, J. N., & Brunton, S. L. (2018). Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, *9*(1), 4950.

Lutzen, J. (1985). Liouville's differential calculus of arbitrary order and its electrodynamical origin. In *Proc. 19th Nordic congress mathenzaticians* (pp. 149–160). Icelandic Mathematical Soc. Reykjavik.

Mendible, A., Koch, J., Lange, H., Brunton, S., & Kutz, N. (2020). Data-driven modeling of detonation wave interactions in rotating detonation engines. *Bulletin of the American Physical Society*.

Mezić, I. (2005). Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dynamics*, *41*(1–3), 309–325.

Mezic, I. (2013). Analysis of fluid flows via spectral properties of the Koopman operator. *Annual Review of Fluid Mechanics*, *45*, 357–378.

Moeck, J. P., Bourgouin, J.-F., Durox, D., Schuller, T., & Candel, S. (2013). Tomographic reconstruction of heat release rate perturbations induced by helical modes in turbulent swirl flames. *Experiments in fluids*, *54*(4), 1498.

Moore, C. C. (2015). Ergodic theorem, ergodic theory, and statistical mechanics. *Proceedings of the National Academy of Sciences*, *112*(7), 1907–1911.

Nathan Kutz, J., Proctor, J. L., & Brunton, S. L. (2018). Applied Koopman theory for partial differential equations and data-driven modeling of spatio-temporal systems. *Complexity*, *2018*.

Neumann, J. v. (1932a). Physical applications of the ergodic hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, *18*(3), 263.

Neumann, J. v. (1932b). Proof of the quasi-ergodic hypothesis. *Proceedings of the National Academy of Sciences*, *18*(1), 70–82.

Oreshkin, B., Carpo, D., Chapados, N., & Bengio, Y. (2019). N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *ICLR*.

Otto, S. E., & Rowley, C. W. (2019). Linearly recurrent autoencoder networks for learning dynamics. *SIAM Journal on Applied Dynamical Systems*, *18*(1), 558–593.

Pan, S., & Duraisamy, K. (2019). Physics-informed probabilistic learning of linear embeddings of non-linear dynamics with guaranteed stability. arXiv preprint arXiv:1906.03663.

Proctor, J. L., & Eckhoff, P. A. (2015). Discovering dynamic patterns from infectious disease data using dynamic mode decomposition. *International health*, *7*(2), 139–145.

Rangapuram, S. S., Seeger, M. W., Gasthaus, J., Stella, L., Wang, Y., & Januschowski, T. (2018). Deep state space models for time series forecasting. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems, vol. 31*. Curran Associates, Inc., URL https://proceedings.neurips.cc/paper/2018/file/5cf68969fb67aa6082363a6d4e6468e2-Paper.pdf.

Rezende, D., & Mohamed, S. (2015). Variational inference with normalizing flows. In *International conference on machine learning* (pp. 1530–1538). PMLR.

Rowley, C. W., Mezić, I., Bagheri, S., Schlatter, P., & Henningson, D. S. (2009). Spectral analysis of nonlinear flows. *Journal of fluid mechanics*, *641*, 115–127.

Rychlewski, J. (1984). On Hooke's law. *Journal of Applied Mathematics and Mechanics*, *48*(3), 303–314.

Schmid, P. J. (2010). Dynamic mode decomposition of numerical and experimental data. *Journal of fluid mechanics*, *656*, 5–28.

Scholkopf, B., & Smola, A. J. (2001). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press.

Sen, R., Yu, H.-F., & Dhillon, I. (2019). Think globally, act locally: a deep neural network approach to high-dimensional time series forecasting. arXiv preprint arXiv:1905.03806.

Smyl, S., & Hua, N. G. (2019). Machine learning methods for GEFCom2017 probabilistic load forecasting. *International Journal of Forecasting*, [ISSN: 0169-2070] *35*(4), 1424–1431. http://dx.doi.org/10.1016/j.ijforecast.2019.02.002.

Song, G., Alizard, F., Robinet, J.-C., & Gloerfelt, X. (2013). Global and Koopman modes analysis of sound generation in mixing layers. *Physics of Fluids*, *25*(12), Article 124101.

Takeishi, N., Kawahara, Y., & Yairi, T. (2017). Learning Koopman invariant subspaces for dynamic mode decomposition. In *Advances in neural information processing systems* (pp. 1130–1140).

Tu, J. H., Rowley, C. W., Luchtenburg, D. M., Brunton, S. L., & Kutz, J. N. (2014). On dynamic mode decomposition: theory and applications. *Journal of Computational Dynamics*, *1*(2), 391–421.

Wehmeyer, C., & Noé, F. (2018). Time-lagged autoencoders: deep learning of slow collective variables for molecular kinetics. *Journal of Chemical Physics*, *148*(24), Article 241703.

Wen, Q., Gao, J., Song, X., Sun, L., Xu, H., & Zhu, S. (2019). RobustSTL: A Robust seasonal-trend decomposition algorithm for long time series. In *Proceedings of the AAAI Conference on artificial intelligence, vol. 33, no. 01* (pp. 5409–5416). http://dx.doi.org/10.1609/aaai.v33i01.33015409, URL https://ojs.aaai.org/index.php/AAAI/article/view/4480.

Wiggins, S. (1987). Chaos in the quasiperiodically forced duffing oscillator. *Physics Letters. A*, *124*(3), 138–142.

Williams, M. O., Kevrekidis, I. G., & Rowley, C. W. (2015). A data–driven approximation of the koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, *25*(6), 1307–1346.

Williams, C. K., & Rasmussen, C. E. (1996). *Gaussian processes for regression*. MIT.

Yang, L., Wen, Q., Yang, B., & Sun, L. (2021). A robust and efficient multi-scale seasonal-trend decomposition. In *ICASSP 2021 - 2021 IEEE International conference on acoustics, speech and signal processing (ICASSP)* (pp. 5085–5089).

Yeung, E., Kundu, S., & Hodas, N. (2019). Learning deep neural network representations for Koopman operators of nonlinear dynamical systems. In *2019 American control conference (ACC)* (pp. 4832–4839). IEEE.

Ziel, F. (2019). Quantile regression for the qualifying match of GEFCom2017 probabilistic load forecasting. *International Journal of Forecasting*, [ISSN: 0169-2070] *35*(4), 1400–1408. http://dx.doi.org/10.1016/j.ijforecast.2018.07.004.