

Multi-variate Probabilistic Time Series Forecasting via Conditioned Normalizing Flows

Kashif Rasul¹ Abdul-Saboor Sheikh¹ Ingmar Schuster¹ Urs Bergmann¹ Roland Vollgraf¹

Abstract

Time series forecasting is often fundamental to scientific and engineering problems and enables decision making. With ever increasing data set sizes, a trivial solution to scale up predictions is to assume independence between interacting time series. However, modeling statistical dependencies can improve accuracy and enable analysis of interaction effects. Deep learning methods are well suited for this problem, but multi-variate models often assume a simple parametric distribution and do not scale to high dimensions. In this work we model the multi-variate temporal dynamics of time series via an autoregressive deep learning model, where the data distribution is represented by a conditioned normalizing flow. This combination retains the power of autoregressive models, such as good performance in extrapolation into the future, with the flexibility of flows as a general purpose high-dimensional distribution model, while remaining computationally tractable. We show that it improves over the state-of-the-art for standard metrics on many real-world data sets with several thousand interacting time-series.

1. Introduction

Classical time series forecasting methods such as those in (Hyndman & Athanasopoulos, 2018) typically provide univariate forecasts and require hand tuned features to model seasonality and other parameters. Time series models based on recurrent neural networks (RNN), like LSTM (Hochreiter & Schmidhuber, 1997), have become popular methods due to their end-to-end training, the ease of incorporating exogenous covariates, and their automatic feature extraction abilities, which are the hallmarks of deep learning. Forecasting outputs can either be points or probability distributions, in which case the forecasts typically come with uncertainty bounds.

¹Zalando Research, Mühlenstraße 25, 10243 Berlin, Germany. Correspondence to: Kashif Rasul <kashif.rasul@zalando.de>.

The problem of modeling uncertainties in time series forecasting is of vital importance for assessing how much to trust the predictions for downstream tasks, such as anomaly detection or (business) decision making. Without probabilistic modeling, the importance of the forecast in regions of low noise (small variance around a mean value) versus a scenario with high noise cannot be distinguished. Hence, point estimation models ignore risk stemming from this noise, which would be of particular importance in some contexts such as making (business) decisions.

Finally, individual time series, in many cases, are statistically dependent on each other, and models need the capacity to adapt to this to improve forecast accuracy (Battaglia et al., 2018). For example, to model the demand for a retail article, it is important to not only model its sales dependent on its own past sales, but also to take into account the effect of interacting articles, which can lead to *cannibalization* effects in the case of article competition. As another example, consider traffic flow in a network of streets as measured by occupancy sensors. A disruption on one particular street will also ripple to occupancy sensors of nearby streets — a uni-variate model would arguable not be able to account for these effects.

In this work we propose an end-to-end trainable autoregressive deep learning model for probabilistic forecasting that explicitly models multi-variate time series and their temporal dynamics employing a normalizing flow architecture, like the Masked Autoregressive Flow (Papamakarios et al., 2017) or Real NVP (Dinh et al., 2017).

The main contributions of this paper are:

1. we propose a probabilistic method to model multi-variate time series, which is able to scale to thousands of time series and their interactions
2. we demonstrate that the model can uncover ground-truth dependency structure on toy data
3. the model establishes the new state-of-the-art on many real world data sets

The model further has the advantages that:

1. the underlying data distribution is modeled using a conditional normalizing flow, which enables adaptation to a broad class of underlying data distributions
2. it is highly efficient to train due to parallelization by using attention (Vaswani et al., 2017), unlike typical RNN-based time series models. Empirically, we observe an order of magnitude faster training times for Transformer-based models.

The paper first provides some background context in Section 2. We then cover related work in Section 3. Section 4 introduces our model and the experiments are detailed in Section 5. We conclude the paper with some discussion in Section 6.

2. Background

We briefly review the current time series forecasting landscape and present the building blocks of our method in this section.

2.1. Time Series Forecasting

Classical time series forecasting methods rely on the ARMA (see e.g. Box et al. 2015) method and its variants like ARIMA. Apart from the fact that these methods require manual feature engineering, they also suffer from the curse of dimensionality, require frequent re-training and are focused on model interpretability rather than test-set accuracy.

Deep learning models over the last years have shown impressive results over classical methods in many fields (Schmidhuber, 2015) like computer vision, speech recognition, natural language processing (NLP), and also time series forecasting, which is related to sequence modeling in NLP (Sutskever et al., 2014). Modern uni-variate point forecast methods like in (Oreshkin et al., 2020) are interpretable and fast to train on many target domains.

Uncertainty estimation for classical methods in the context of control theory have been worked on for decades, see e.g. (Dietz et al., 1997). The majority of the classic forecasting literature has focused on prediction of point estimates, such as the mean or the median of the distribution at a future time point. In the deep learning setting the two approaches have been to either model the data distribution explicitly or to use BNN as in (Zhu & Laptev, 2018). To estimate the underlying temporal distribution we can either learn the parameters of some target distribution as in the DeepAR method (Flunkert et al., 2017) or use mixture density models (McLachlan & Basford, 1988) operating on neural network outputs, called mixture density networks (MDN) (Bishop, 2006), as for example in the MD-RNN approach used to model handwriting (Graves, 2013). Recently (Rangapuram et al., 2018) combined a linear state

space model for each individual time series together with deep probabilistic models to additionally obtain interpretative time series predictions.

To model all time series jointly, i.e. capture interaction effects, one can use multi-variate Gaussian processes to capture the underlying structure of data (Vandenberg-Rodes & Shahbaba, 2015) or Low-rank Gaussian Copula processes via RNNs (Salinas et al., 2019). The temporal regularized matrix factorization framework (Yu et al., 2016) proposes learning the data dependencies, thus allowing the ability to forecast future values, via a matrix factorization approach. The LSTNet (Lai et al., 2018) approach uses Convolutional Neural Network (CNN) and RNN building blocks to model multi-variate time series for point forecasts. Bayesian models using hierarchical priors have also been proposed to share statistical strength between individual time series while keeping inference feasible (Chapados, 2014). The use of multi-head attention for time series forecasting has also recently been explored (Li et al., 2019) and it allows capturing long term dependencies where RNNs like the LSTM suffer.

2.2. Density Estimation via Normalizing Flows

Normalizing flows (Tabak & Turner, 2013) are mappings from \mathbb{R}^D to \mathbb{R}^D such that densities $p_{\mathcal{X}}$ on the input space $\mathcal{X} = \mathbb{R}^D$ are transformed into some simple distribution $p_{\mathcal{Z}}$ (e.g. an isotropic Gaussian) on the space $\mathcal{Z} = \mathbb{R}^D$. This mapping $f: \mathcal{X} \mapsto \mathcal{Z}$, is composed of a sequence of bijections or invertible functions. Due to the change of variables formula we can express $p_{\mathcal{X}}(\mathbf{x})$ by

$$p_{\mathcal{X}}(\mathbf{x}) = p_{\mathcal{Z}}(\mathbf{z}) \left| \det \left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \right) \right|,$$

where $\partial f(\mathbf{x})/\partial \mathbf{x}$ is the Jacobian of f at \mathbf{x} . Normalizing flows have the property that the inverse $\mathbf{x} = f^{-1}(\mathbf{z})$ is easy to evaluate and computing the Jacobian determinant takes $O(D)$ time.

The bijection introduced by Real NVP (Dinh et al., 2017) called the *coupling layer* satisfies the above two properties. It leaves part of its inputs unchanged and transforms the other part via functions of the un-transformed variables

$$\begin{cases} \mathbf{y}^{1:d} = \mathbf{x}^{1:d} \\ \mathbf{y}^{d+1:D} = \mathbf{x}^{d+1:D} \odot \exp(s(\mathbf{x}^{1:d})) + t(\mathbf{x}^{1:d}), \end{cases}$$

where \odot is an element wise product, s is a scaling and t a translation function from $\mathbb{R}^d \mapsto \mathbb{R}^{D-d}$, given by neural networks. To model a nonlinear density map $f(\mathbf{x})$, a number of coupling layers $\mathcal{X} \mapsto \mathcal{Y}_1 \mapsto \dots \mapsto \mathcal{Y}_{K-1} \mapsto \mathcal{Z}$ are composed together all the while alternating the dimensions which are unchanged and transformed. Via the change of variables formula the probability density function (PDF) of

the flow given a data point can be written as

$$\begin{aligned}\log p_{\mathcal{X}}(\mathbf{x}) &= \log p_{\mathcal{Z}}(\mathbf{z}) + \log |\det(\partial \mathbf{z} / \partial \mathbf{x})| \\ &= \log p_{\mathcal{Z}}(\mathbf{z}) + \sum_{i=1}^K \log |\det(\partial \mathbf{y}_i / \partial \mathbf{y}_{i-1})|. \end{aligned} \quad (1)$$

Note that the Jacobian for the Real NVP is a **block-triangular matrix** and thus the log-determinant simply becomes

$$\log |\det(\partial \mathbf{y}_i / \partial \mathbf{y}_{i-1})| = \text{sum}(\log |\text{diag}(\exp(s(\mathbf{y}_{i-1})))|), \quad (2)$$

where $\text{sum}()$ is the sum over all the vector elements, $\log()$ is the element-wise logarithm and $\text{diag}()$ is the diagonal of the Jacobian. This model, parameterized by the weights of the scaling and translation neural networks θ , is then trained via stochastic gradient descent (SGD) on training data points where for each batch \mathcal{D} we **maximize the average log likelihood** (1) given by

$$\mathcal{L} = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x} \in \mathcal{D}} \log p_{\mathcal{X}}(\mathbf{x}; \theta).$$

In practice Batch Normalization (Ioffe & Szegedy, 2015) is applied, as a bijection, to outputs of successive coupling layers to stabilize training of normalizing flows. This bijection implements the normalization procedure using a weighted moving average of the layer’s mean and standard deviation values, which has to be adapted to either training or inference regimes.

The Real NVP approach can be generalized, resulting in Masked Autoregressive Flows (MAF, Papamakarios et al. 2017) where the transformation layer is built as an autoregressive neural network in the sense that it takes in some input $\mathbf{x} \in \mathbb{R}^D$ and outputs $(\mathbf{c}^1, \dots, \mathbf{c}^D)$ with the requirement that this transformation is invertible and any output \mathbf{c}^i cannot depend on input dimensions $\mathbf{x}^{\geq i}$. **The Jacobian of this transformation is triangular and thus the Jacobian determinant is tractable.** Instead of using a RNN to share parameters across the D dimensions of \mathbf{x} one avoids this sequential computation by using masking, giving the method its name. The inverse however, needed for generating samples, is sequential.

By realizing that the scaling and translation function approximators don’t need to be invertible, it is straight-forward to implement conditioning of the PDF $p_{\mathcal{X}}(\mathbf{x}|\mathbf{h})$ on some additional information $\mathbf{h} \in \mathbb{R}^H$: we concatenate \mathbf{h} to the inputs of the scaling and translation function approximators of the coupling layers, i.e. $s(\text{concat}(\mathbf{x}^{1:d}, \mathbf{h}))$ and $t(\text{concat}(\mathbf{x}^{1:d}, \mathbf{h}))$ which are modified to map $\mathbb{R}^{d+H} \mapsto \mathbb{R}^{D-d}$. Another approach is to add a bias computed from \mathbf{h} to every layer inside the s and t networks as proposed

by (Korshunova et al., 2018). This does not change the log-determinant of the coupling layers given by (2). More importantly for us, for sequential data we can share parameters across the different conditioners by using RNNs in an autoregressive fashion.

For discrete data the distribution has differential entropy of negative infinity, which leads to arbitrary high likelihoods when training normalizing flow models, even on test data. To avoid this one can *dequantize* the data, often by adding Uniform[0, 1) noise. The log-likelihood of the continuous model is then lower-bounded by the log-likelihood of the discrete one as shown in Theis et al. (2016).

2.3. Self-attention

The self-attention based Transformer layer (Vaswani et al., 2017) has been used for sequence modeling with great success. The multi-head self-attention mechanism enables it to capture both long- and short-term dependencies in time series data. Essentially, the Transformer takes in a sequence $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_t]^T \in \mathbb{R}^{t \times D}$, and the multi-head self-attention transforms this into H distinct query matrices $\mathbf{Q}_h = \mathbf{X} \mathbf{W}_h^Q$, key matrices $\mathbf{K}_h = \mathbf{X} \mathbf{W}_h^K$ and value matrices $\mathbf{V}_h = \mathbf{X} \mathbf{W}_h^V$ where the \mathbf{W}_h^Q , \mathbf{W}_h^K , and \mathbf{W}_h^V are the learnable parameters. After these linear projections the scaled dot-product attention computes a sequence of vector outputs via:

$$\begin{aligned}\mathbf{O}_h &= \text{Attention}(\mathbf{Q}_h, \mathbf{K}_h, \mathbf{V}_h) \\ &= \text{softmax} \left(\frac{\mathbf{Q}_h \mathbf{K}_h^T}{\sqrt{d_K}} \cdot \mathbf{M} \right) \mathbf{V}_h, \end{aligned}$$

where a mask \mathbf{M} is applied to filter out right-ward attention (or future information leakage) by setting its upper-triangular elements to $-\infty$ and we normalize by d_K the dimension of the \mathbf{W}_h^K matrices. Afterwards all the H \mathbf{O}_h outputs are concatenated and linearly projected again.

One typically uses the Transformer in an encoder-decoder setup, where some warm up time series is passed through the encoder and the decoder can be used to learn and autoregressively generate outputs.

3. Related Work

Related to this work are models that combine normalizing flows for sequential modeling in some way. Transformation Autoregressive Networks (TANs, Oliva et al. 2018) which model the density of a multi-variate variable $\mathbf{x} \in \mathbb{R}^D$ as D conditional distributions $\prod_{i=1}^D p_{\mathcal{X}}(x^i | x^{i-1}, \dots, x^1)$, where the conditioning is given by a mixture model coming from the state of a RNN, and is then transformed via a bijection. The PixelSNAIL (Chen et al., 2018) method also models the joint as a product of conditional distributions, optionally with some global conditioning, via causal convolutions and

self-attention (Vaswani et al., 2017) to capture long-term temporal dependencies. These methods are well suited to modeling high dimensional data like images, however their use in modeling the temporal development of data has only recently been explored for example in VideoFlow (Kumar et al., 2019) which consists of a normalizing flow that autoregressively models the latent variable at time t as a Gaussian whose parameters are functions of the flow at previous time steps.

Using RNNs for modeling either multi-variate or temporal dynamics introduces sequential computational dependencies that are not amenable to parallelization. However, RNNs have been shown to be very effective in modeling sequential dynamics and we feel that it is important to nonetheless explore RNN-based temporal conditioning for multi-variate time series forecasting. A recent work in this direction (Hwang et al., 2019) employs bipartite flows with GRUs for temporal conditioning to develop a conditional generative model of multi-variate sequential data. The authors use a bidirectional training procedure to learn a generative model of observations that together with the temporal conditioning through a RNN, can also be conditioned on (observed) covariates that are modeled as additional conditioning variables in the latent space, which adds extra padding dimensions to the normalizing flow.

The Gaussian Copula Process method (Salinas et al., 2019) is a RNN-based time series model with a Gaussian copula process output model using a low-rank covariance structure to reduce the computational complexity and handle non-Gaussian marginal distributions. By using a low-rank approximation of the covariance matrix they obtain a computationally tractable method and are able to scale to multi-variate dimensions in the thousands with state-of-the-art results. We will compare our model to this method in what follows.

4. Temporal Conditioned Normalizing Flows

We denote the entities of a multi-variate time series by $x_t^i \in \mathbb{R}$ for $i \in \{1, \dots, D\}$ where t is the time index. Thus the multi-variate vector at time t is given by $\mathbf{x}_t \in \mathbb{R}^D$. In this paper, we investigate the effect of the time evolution and the output distribution module on both statistical accuracy and computation time. We find that an attention layer for time evolution together with an MAF output density model results in high forecasting accuracy. At the same time, attention results in orders of magnitudes faster training.

In the DeepAR model (Flunkert et al., 2017), the log-likelihood of each entity x_t^i at a time step $t \in [1, t_0]$ is maximized given an individual time series' history. This is done with respect to the parameters of the chosen distributional model (e.g. negative binomial for count data) via the

state of a RNN derived from its previous time step x_{t-1}^i and its corresponding covariates c_{t-1}^i . The emission distribution model, which is typically Gaussian for real-valued data or negative binomial for count data, is selected to best match the statistics of the time series and the network incorporates activation functions that satisfy the constraints of these distribution parameters, e.g. a `softplus()` for the scale parameter of the Gaussian.

A simple model for multi-variate real-valued data could use a factorizing distribution in the emissions. Shared parameters can then learn patterns across the individual time series through the temporal component — but the model falls short of capturing dependencies in the emissions of the model. For this, a full joint distribution at each time step must be modeled, for example by using a multivariate Gaussian model. However, modeling the full covariance matrix not only increases the number of parameters of the neural network by $O(D^2)$, making learning difficult, but computing the loss becomes expensive when D is large. Further, statistical dependencies in the emissions would be limited to second-order effects. These models are referred to as Vec-LSTM in Salinas et al. (2019).

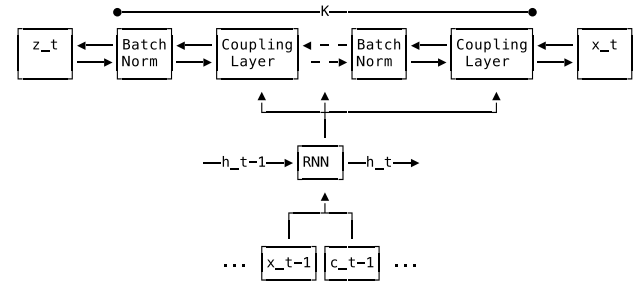


Figure 1. RNN Conditioned Real NVP model schematic at time t , consisting of K blocks of coupling layers and Batch Normalization, where in each coupling layer we condition \mathbf{x}_t and its transformations on the state of a shared RNN from the previous time step \mathbf{x}_{t-1} and its covariates \mathbf{c}_{t-1} which are typically time dependent and time independent features.

We wish to have a scalable model of D interacting time-series \mathbf{x}_t , and further to use a flexible distribution model on the emissions that allows to capture and represent higher order moments. To this end, we model the conditional joint distribution at time t of all time series $p_{\mathcal{X}}(\mathbf{x}_t | \mathbf{h}_t; \theta)$ with a flow, e.g. a Real NVP, conditioned on either the hidden state of a RNN at time t or an embedding of the time series up to t from an attention module. In the case of an autoregressive RNN (either a LSTM or a GRU; Chung et al. 2014), its hidden state \mathbf{h}_t is updated given the previous time step observation \mathbf{x}_{t-1} and its associated covariates \mathbf{c}_{t-1} (as in

Figure 1):

$$\mathbf{h}_t = \text{RNN}(\text{concat}(\mathbf{x}_{t-1}, \mathbf{c}_{t-1}), \mathbf{h}_{t-1}). \quad (3)$$

This model is autoregressive since it consumes the observation of the last time step \mathbf{x}_{t-1} as well as the recurrent state \mathbf{h}_{t-1} to produce the state \mathbf{h}_t on which we condition the current observation.

To get a powerful emission distribution model, we stack K layers of the flow model (Real NVP or MAF). Together with the RNN, we arrive at our model of the conditional distribution of the future of all time series, given its past $t \in [1, t_0 - 1]$ and all the covariates in $t \in [1, T]$. As the model is autoregressive it can be written as a product of factors

$$p_{\mathcal{X}}(\mathbf{x}_{t_0:T} | \mathbf{x}_{1:t_0-1}, \mathbf{c}_{1:T}) = \prod_{t=t_0}^T p_{\mathcal{X}}(\mathbf{x}_t | \mathbf{h}_t; \theta), \quad (4)$$

where θ denotes the set of all parameters of both the flow and the RNN.

For modeling the time evolution, we also investigate using an attention module (Vaswani et al., 2017). This is used to compute an embedding of the time series up to t , which we will also call \mathbf{h}_t for simplicity. In this case, the training time series is split into a warm up or encoding portion $\mathbf{x}_{1:t-1}$ and the output portion $\mathbf{x}_{t:T}$. See Figure 2 for a schematic of the overall model in this case. In training, care has to be taken to prevent using information from future time points and to preserve the autoregressive property by utilizing a mask that reflects the causal direction of progressing time, i.e. to mask out future time points.

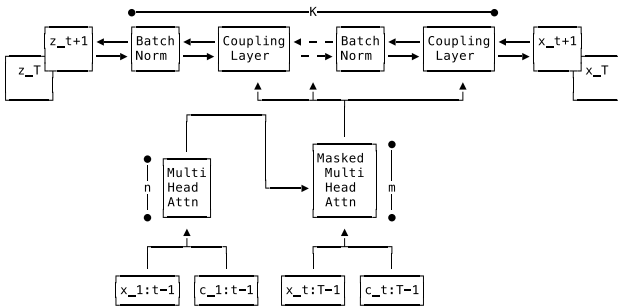


Figure 2. Transformer Conditioned Real NVP model schematic consisting of an encoder-decoder stack where the encoder takes in some context length of time series and then uses it to generate conditioning for the prediction length portion of the time series via a causally masked decoder stack. The output of the decoder is used as conditioning to train the flow. Note that the positional encodings are part of the covariates and unlike the RNN model, here all $\mathbf{x}_{t:T-1}$ time points are trained in parallel.

In real-world data the magnitudes of different time series can vary drastically. To normalize scales, we divide each

individual time series by its mean before feeding it into the model. The outputs are then correspondingly multiplied with the same mean values to match the original scale. This rescaling technique simplifies the problem for the model, which is reflected in significantly improved empirical performance.

4.1. Training

Given data of D time series, where at each time step we have $\mathbf{x}_t \in \mathbb{R}^D$ and their associated covariates \mathbf{c}_t , we maximize the log-likelihood given by (1) and (3)

$$\mathcal{L} = \frac{1}{|\mathcal{D}|T} \sum_{\mathbf{x}_{1:T} \in \mathcal{D}} \sum_{t=1}^T \log p_{\mathcal{X}}(\mathbf{x}_t | \mathbf{h}_t; \theta)$$

via SGD using Adam (Kingma & Ba, 2015) with respect to the parameters θ of the conditional flow and the RNN or Transformer. In practice, the time series $\mathbf{x}_{1:T}$ in a batch \mathcal{D} is selected from a random time window of size T within our training data, and the relative time steps are kept constant. This allows the model to learn to cold-start given only the covariates. This also increases the size of our training data when the training window size is not too large and allows us to trade-off computation time with memory consumption especially when D or T are large. Note that information about absolute time is only available to the RNN or Transformer via the covariates and not the relative position of \mathbf{x}_t in the training data.

The Transformer has computational complexity $O(T^2F)$ compared to a RNN which is $O(TF^2)$, where F is the dimension of the hidden state and T the time series length. This means for large multi-variate time series, i.e. $D > T$, and the assumption that the dimension of the hidden state grows proportional to the number of simultaneous time-series modeled, i.e. $F \propto D$, the Transformer flow model has smaller computational complexity. In addition, unlike the RNN, all the computation for training happens in parallel. The Transformer allows the model to access any part of the historic time series regardless of temporal distance and thus is able to generate better conditioning for the normalizing flow.

4.2. Covariates

We employ embeddings for categorical features (Charrington, 2018), which allows for relationships within a category, or its context, to be captured while training models. Combining these embeddings as features for time series forecasting yields powerful models like the first place winner of the Kaggle Taxi Trajectory Prediction¹ challenge (De Brébisson et al., 2015). The covariates \mathbf{c}_t we use are composed

¹<https://www.kaggle.com/c/pkdd-15-predict-taxi-service-trajectory-i>

of time-dependent (e.g. day of week, hour of day) and time-independent embeddings, if applicable, as well as lag features depending on the time frequency of the data set we are training on. All covariates must be known for the time periods we wish to forecast.

4.3. Inference

For inference we either obtain the hidden state $\tilde{\mathbf{h}}_{t_0+1}$ by passing a “warm up” time series $\mathbf{x}_{1:t_0}$ through the RNN or use the cold-start hidden state, i.e. we set $\mathbf{h}_{t_0+1} = \mathbf{h}_0$, and then by sampling a noise vector $\mathbf{z}_{t_0+1} \in \mathbb{R}^D$ from an isotropic Gaussian, go backward through the flow to obtain a sample of our time series for the next time step conditioned on this starting state $\tilde{\mathbf{x}}_{t_0+1} = f^{-1}(\mathbf{z}_{t_0+1} | \tilde{\mathbf{h}}_{t_0+1})$. We then use this sample and its covariate to obtain the next conditioning state $\tilde{\mathbf{h}}_{t_0+2}$ via the RNN and repeat till our inference horizon. This process of sampling trajectories from some initial state can be repeated many times to obtain empirical quantiles of the uncertainty of our prediction for arbitrary long forecast horizons.

The attention model similarly uses a warm up time series $\mathbf{x}_{1:t_0}$ and covariates and passes them through the encoder and then uses the decoder to output the conditioning for sampling from the flow. This sample is then used again in the decoder to iteratively sample the next conditioning state, similar to the inference procedure in seq-to-seq models.

5. Experiments

Here we discuss a toy experiment sanity-checking our model as well as results on six real world data sets.

5.1. Simulated Flow in a System of Pipes

In this experiment, we check for some basic properties of our model by simulating flow of a liquid in a system of pipes with valves. See Figure 3 for a depiction of the system.

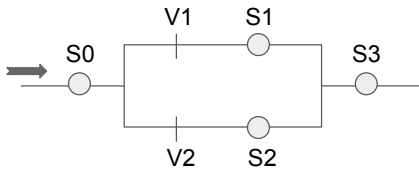


Figure 3. System of pipes with liquid flowing from left to right with sensors (S) and valves (V).

Liquid flows from left to right, where pressure at the first sensor in the system is given by $S_0 = X + 3$, $X \sim \text{Gamma}(1, 0.2)$ in the shape/scale parameterization of the Gamma distribution. The valves are given by $V_1, V_2 \sim \text{iid}$

$\text{Beta}(0.5, 0.5)$, and we have

$$S_i = \frac{V_i}{V_1 + V_2} S_0 + \epsilon_i$$

for $i \in \{1, 2\}$ and finally $S_3 = S_1 + S_2 + \epsilon_3$ with $\epsilon_* \sim \mathcal{N}(0, 0.1)$. With this simulation we check whether our model captures correlations in space and time. The correlation between S_1 and S_2 results from both having the same source, measured by S_0 . This is reflected by $\text{Cov}(S_1, S_2) > 0$, which is captured by our model.

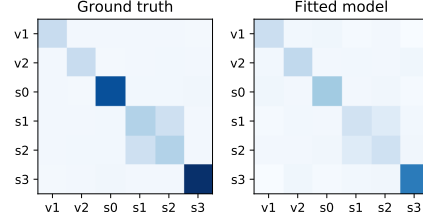


Figure 4. Covariance matrix for a fixed time point capturing the correlation between S_1 and S_2 . Darker means higher positive values.

The cross-covariance structure between consecutive time points in ground truth and as captured by our trained model is depicted in Figure 5. It reflects the true flow of liquid in the system from S_0 at time t to S_1 and S_2 at time $t + 1$, on to S_3 at time $t + 2$.

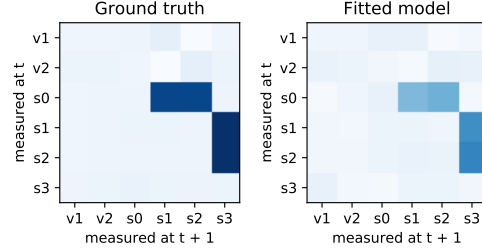


Figure 5. Cross-covariance matrix between consecutive time points capturing true flow of liquid in the pipe system. Darker means higher positive values.

5.2. Real World Data Sets

For evaluation we compute the *Continuous Ranked Probability Score* (CRPS) (Matheson & Winkler, 1976) on each individual time series, as well as on the sum of all time series (the latter denoted by CRPS_{sum}). CRPS measures the compatibility of a cumulative distribution function F with an observation x as

$$\text{CRPS}(F, x) = \int_{\mathbb{R}} (F(z) - \mathbb{I}\{x \leq z\})^2 dz \quad (5)$$

where $\mathbb{I}\{x \leq z\}$ is the indicator function which is one if $x \leq z$ and zero otherwise. CRPS is a proper scoring function, hence CRPS attains its minimum when the predictive

Table 1. Test set CRPS_{sum} comparison (lower is better) of models from (Salinas et al., 2019) and our models GRU-Real-NVP, GRU-MAF and Transformer-MAF. The *two* best methods are in bold where the mean and standard errors are obtained by re-running each method three times.

Data set	Vec-LSTM ind-scaling	Vec-LSTM lowrank-Copula	GP scaling	GP Copula	GRU Real-NVP	GRU MAF	Transformer MAF
Exchange	0.008±0.001	0.007±0.000	0.009±0.000	0.007±0.000	0.0064±0.001	0.005±0.001	0.005±0.001
Solar	0.391±0.017	0.319±0.011	0.368±0.012	0.337±0.024	0.331±0.02	0.315±0.023	0.301±0.014
Electricity	0.025±0.001	0.064±0.008	0.022±0.000	0.024±0.002	0.024±0.001	0.0208±0.000	0.0207±0.000
Traffic	0.087±0.041	0.103±0.006	0.079±0.000	0.078±0.002	0.078±0.001	0.069±0.002	0.056±0.001
Taxi	0.506±0.005	0.326±0.007	0.183±0.395	0.208±0.183	0.175±0.001	0.161±0.002	0.179±0.002
Wikipedia	0.133±0.002	0.241±0.033	1.483±1.034	0.086±0.004	0.078±0.001	0.067±0.001	0.063±0.003

distribution F and the data distribution equal. Employing the empirical CDF of F , i.e. $\hat{F}(z) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{X_i \leq z\}$ with samples $X_i \sim F$ as a natural approximation of the predictive CDF, CRPS can be directly computed from simulated samples of the conditional distribution (4) at each time point (Jordan et al., 2019). We take 100 samples to estimate the empirical CDF in practice. In comparison to log-likelihood, CRPS puts heavier penalty on the distribution tails (Gebetsberger et al., 2018). Finally, CRPS_{sum} is obtained by first summing across the D time-series — both for the ground-truth data, and sampled data (yielding $\hat{F}_{\text{sum}}(t)$ for each time-point). The results are then averaged over the prediction horizon, i.e. formally $\text{CRPS}_{\text{sum}} = \mathbb{E}_t \left[\text{CRPS} \left(\hat{F}_{\text{sum}}(t), \sum_i x_t^i \right) \right]$.

Our model is trained on the training split of each data set, and for testing we use a rolling windows prediction starting from the last point seen in the training data set and compare it to the test set.

We train on Exchange (Lai et al., 2018), Solar (Lai et al., 2018), Electricity², Traffic³, Taxi⁴ and Wikipedia⁵ open data sets, preprocessed exactly as in (Salinas et al., 2019), with their properties listed in Table 2. Both Taxi and Wikipedia consist of count data and are thus dequantized before being fed to the flow (and mean-scaled).

We use batch sizes of 32, with 100 batches per epoch and train for a maximum of 40 epochs with a learning rate of $1e-3$. The LSTM/GRU hyperparameters were the ones from (Salinas et al., 2019) and we used $K = 3$ or $K = 5$ stacks of normalizing flow bijections. We sample 100 times to report the metrics on the test set. The Transformer uses

²<https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014>

³<https://archive.ics.uci.edu/ml/datasets/PEMS-SF>

⁴<https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>

⁵https://github.com/mbohlkeschneider/gluon-ts/tree/mv_release/datasets

Table 2. Properties of the experiment data sets.

DATA SET	DIMENSION D	DOMAIN	FREQ.
EXCHANGE	8	\mathbb{R}^+	DAILY
SOLAR	137	\mathbb{R}^+	HOURLY
ELECTRICITY	370	\mathbb{R}^+	HOURLY
TRAFFIC	963	(0, 1)	HOURLY
TAXI	1,214	\mathbb{N}	30-MIN
WIKIPEDIA	2,000	\mathbb{N}	DAILY

$H = 8$ heads and $n = m = 3$ encoding and decoding layers and a dropout rate of 0.1. No extensive hyperparameter tuning was done. All the experiments run on a single Nvidia V-100 GPU.

We compare our method using GRU and two different normalizing flows (GRU-Real-NVP and GRU-MAF based on Real NVP and MAF, respectively) as well as a Transformer model with MAF (Transformer-MAF), with different RNN based methods and transformation schemes from (Salinas et al., 2019) and report the results in Table 1. Vec-LSTM-ind-scaling outputs the parameters of an independent normal distribution with mean-scaling, Vec-LSTM-lowrank-Copula parametrizes a low-rank plus diagonal covariance via Copula process. GP-scaling unrolls a LSTM with scaling on each individual time series before reconstructing the joint distribution. Similarly, GP-Copula unrolls a LSTM on each individual time series and then the joint emission distribution is given by a low-rank plus diagonal covariance Gaussian copula.

In Table 1 we observe that MAF with either RNN or self-attention mechanism for temporal conditioning achieves the state-of-the-art (to the best of our knowledge) CRPS_{sum} on all benchmarks. Moreover, bipartite flows with RNN either also outperform or are found to be competitive w.r.t. to the previous state-of-the-art results as listed in the first four columns of Table 1. Further analyses with other metrics (e.g. MSE) are reported in the Supplementary Material.

To assess how well our model captures dependencies in ex-

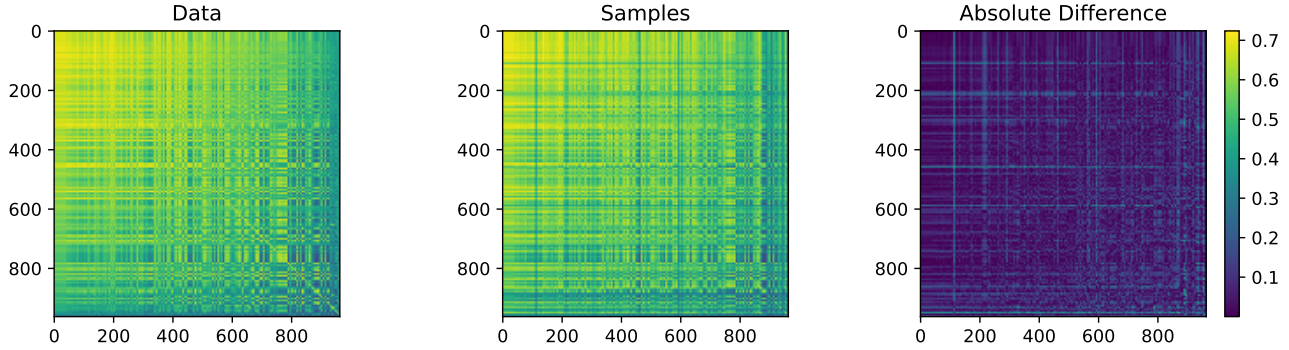


Figure 6. Analysis of the dependency structure extrapolation of the model. Left: Cross-covariance matrix computed from the test split of *Traffic* benchmark. Middle: Cross-covariance matrix computed from the mean of 100 sample trajectories drawn from the *Transformer-MAF* model’s extrapolation into the future (test split). Right: The absolute difference of the two matrices mostly shows small deviations between ground-truth and extrapolation.

trapolating the time series into the future versus real data, we plot in Figure 6 the cross-covariance matrix of observations (plotted left) as well as the mean of 100 sample trajectories (middle plot) drawn from *Transformer-MAF* model for the test split of *Traffic* data set. The right most plot in the figure illustrates the absolute difference between the two cross-covariance matrices. As can be seen, most of the covariance structure especially in the top-left region of highly correlated sensors is very well reflected in the samples drawn from the model.

6. Conclusion

We have presented a general method to model high dimensional probabilistic multivariate time series via combining conditional normalizing flows with an autoregressive model, such as a recurrent neural network or an attention module. Autoregressive models have a long-standing reputation for working very well for time series forecasting, as they show good performance in extrapolation into the future. The flow model, on the other hand, does not assume any simple fixed distribution class, but instead can adapt to a broad range of high-dimensional data distributions. The combination hence combines the extrapolation power of the autoregressive model class with the density estimation flexibility of flows. Further, it is computationally efficient, without the need of resorting to approximations (e.g. low-rank approximations of a covariance structure). Analysis on six commonly used time series benchmarks establish the new state-of-the-art performance, without much hyperparameter tuning.

A natural way to improve the model is to incorporate a better underlying flow model. For example, Table 1 showed that swapping a Real NVP flow in for a MAF improved

performance, which is a consequence of Real NVP lacking in density modeling performance compared to MAF. Likewise, we would expect other design choices of the flow model to improve performance, e.g. changes on the dequantization method, the specific affine coupling layer or more expressive conditioning, say via another Transformer. Recent improvements to flows, e.g. as proposed in the *Flow++* paper (Ho et al., 2019), to obtain expressive bipartite flow models, or better flow models to handle discrete categorical data (Tran et al., 2019), are left as future work to assess their usefulness. To our knowledge, it is however still an open problem how to model discrete ordinal data via flows — which would best capture the nature of some data sets (e.g. sales data).

Also, we would expect improvements in the time evolution module to improve the time series forecasts. For example, recent improvements to the Transformer like the *Reformer* (Kitaev et al., 2020) could improve memory efficiency. We leave these improvements to future investigations.

Finally, real-world applications might require training on very large number of interacting time series D , such as e.g. in sales modeling in e-Commerce, where D can be in the order of millions or more. Flows have been successfully applied to image modeling, which is comparable to the instantaneous dimensionality faced in this setting — but the memory requirement becomes infeasible for large time-series. In future work, we’ll investigate mechanisms for scalable training of time-series models, e.g. by subsampling time-series.

References

- Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V. F., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gülçehre, Ç., Song, F., Ballard, A. J., Gilmer, J., Dahl, G. E., Vaswani, A., Allen, K., Nash, C., Langston, V., Dyer, C., Heess, N., Wierstra, D., Kohli, P., Botvinick, M., Vinyals, O., Li, Y., and Pascanu, R. Relational inductive biases, deep learning, and graph networks. *CoRR*, abs/1806.01261, 2018.
- Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.
- Box, G., Jenkins, G., Reinsel, G., and Ljung, G. *Time Series Analysis: Forecasting and Control*. Wiley Series in Probability and Statistics. Wiley, 2015. ISBN 9781118674925.
- Chapados, N. Effective bayesian modeling of groups of related count time series. In *ICML*, volume 32 of *JMLR Workshop and Conference Proceedings*, pp. 1395–1403. JMLR.org, 2014.
- Charrington, S. TWiML & AI Podcast: Systems and software for machine learning at scale with Jeff Dean, 2018. URL <https://bit.ly/2G0LmGg>.
- Chen, X., Mishra, N., Rohaninejad, M., and Abbeel, P. PixelSNAIL: An improved autoregressive generative model. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 864–872, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL <http://proceedings.mlr.press/v80/chen18h.html>.
- Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.
- De Brébisson, A., Simon, E., Auvolet, A., Vincent, P., and Bengio, Y. Artificial neural networks applied to taxi destination prediction. In *Proceedings of the 2015th International Conference on ECML PKDD Discovery Challenge - Volume 1526, ECMLPKDDDC’15*, pp. 40–51, Aachen, Germany, Germany, 2015. CEUR-WS.org. URL <http://dl.acm.org/citation.cfm?id=3056172.3056178>.
- Dietz, R. D., Casavant, T. L., Scheetz, T. E., Braun, T. A., and Andersland, M. S. Modeling the impact of runtime uncertainty on optimal computation scheduling using feedback. In *1997 International Conference on Parallel Processing (ICPP ’97), August 11-15, 1997, Bloomington, IL, USA, Proceedings*, pp. 481–488, 1997. doi: 10.1109/ICPP.1997.622683. URL <https://doi.org/10.1109/ICPP.1997.622683>.
- Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using Real NVP, 2017. URL <https://arxiv.org/abs/1605.08803>.
- Flunkert, V., Salinas, D., and Gasthaus, J. DeepAR: Probabilistic forecasting with autoregressive recurrent networks, 2017. URL <http://arxiv.org/abs/1704.04110>.
- Gebetsberger, M., Messner, J. W., Mayr, G. J., and Zeileis, A. Estimation methods for nonhomogeneous regression models: Minimum continuous ranked probability score versus maximum likelihood. *Monthly Weather Review*, 146(12):4323–4338, 2018. doi: 10.1175/MWR-D-17-0364.1.
- Graves, A. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.
- Ho, J., Chen, X., Srinivas, A., Duan, Y., and Abbeel, P. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2722–2730, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL <http://proceedings.mlr.press/v97/ho19a.html>.
- Hochreiter, S. and Schmidhuber, J. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, Nov 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735.
- Hwang, S. J., Tao, Z., Kim, W. H., and Singh, V. Conditional recurrent flow: Conditional generation of longitudinal samples with applications to neuroimaging. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
- Hyndman, R. and Athanasopoulos, G. *Forecasting: principles and practice*. OTexts, 2018. ISBN 9780987507112.
- Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37, ICML’15*, pp. 448–456. JMLR.org, 2015. URL <http://dl.acm.org/citation.cfm?id=3045118.3045167>.
- Jordan, A., Krüger, F., and Lerch, S. Evaluating probabilistic forecasts with scoringrules. *Journal of Statistical Software, Articles*, 90(12):1–37, 2019. ISSN 1548-7660. doi: 10.18637/jss.v090.i12. URL <https://www.jstatsoft.org/v090/i12>.

- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.
- Kitaev, N., Kaiser, L., and Levskaya, A. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=rkgNKkHtvB>.
- Korshunova, I., Gal, Y., Gretton, A., and Dambre, J. Conditional BRUNO: A Deep Recurrent Process for Exchangeable Labelled Data. In *Bayesian Deep Learning workshop, NIPS*, 2018.
- Kumar, M., Babaeizadeh, M., Erhan, D., Finn, C., Levine, S., Dinh, L., and Kingma, D. VideoFlow: A Flow-Based Generative Model for Video. In *Workshop on Invertible Neural Nets and Normalizing Flows, ICML*, 2019.
- Lai, G., Chang, W.-C., Yang, Y., and Liu, H. Modeling long- and short-term temporal patterns with deep neural networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18*, pp. 95–104, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5657-2. doi: 10.1145/3209978.3210006. URL <http://doi.acm.org/10.1145/3209978.3210006>.
- Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y.-X., and Yan, X. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 5244–5254. Curran Associates, Inc., 2019.
- Matheson, J. E. and Winkler, R. L. Scoring rules for continuous probability distributions. *Management Science*, 22 (10):1087–1096, 1976.
- McLachlan, G. J. and Basford, K. E. *Mixture models: Inference and applications to clustering*. Marcel Dekker, New York, 1988.
- Oliva, J., Dubey, A., Zaheer, M., Poczos, B., Salakhutdinov, R., Xing, E., and Schneider, J. Transformation autoregressive networks. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3898–3907, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL <http://proceedings.mlr.press/v80/olival8a.html>.
- Oreshkin, B. N., Carпов, D., Chapados, N., and Bengio, Y. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=rlecqn4YwB>.
- Papamakarios, G., Pavlakou, T., and Murray, I. Masked autoregressive flow for density estimation. *Advances in Neural Information Processing Systems 30*, 2017.
- Rangapuram, S. S., Seeger, M. W., Gasthaus, J., Stella, L., Wang, Y., and Januschowski, T. Deep state space models for time series forecasting. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 7796–7805. Curran Associates, Inc., 2018.
- Salinas, D., Bohlke-Schneider, M., Callot, L., Medico, R., and Gasthaus, J. High-dimensional multivariate forecasting with low-rank gaussian copula processes. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 6824–6834. Curran Associates, Inc., 2019.
- Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.
- Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 3104–3112. Curran Associates, Inc., 2014.
- Tabak, E. G. and Turner, C. V. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.
- Theis, L., van den Oord, A., and Bethge, M. A note on the evaluation of generative models. In *International Conference on Learning Representations*, 2016. URL <http://arxiv.org/abs/1511.01844>. arXiv:1511.01844.
- Tran, D., Vafa, K., Agrawal, K., Dinh, L., and Poole, B. Discrete flows: Invertible generative models of discrete data. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 14692–14701. Curran Associates, Inc., 2019.
- Vandenberg-Rodes, A. and Shahbaba, B. Dependent Matérn processes for multivariate time series, 2015. URL <http://arxiv.org/abs/1502.03466>.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio,

S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc., 2017. URL <http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>.

Yu, H.-F., Rao, N., and Dhillon, I. S. Temporal regularized matrix factorization for high-dimensional time series prediction. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 847–855. Curran Associates, Inc., 2016.

Zhu, L. and Laptev, N. Deep and confident prediction for time series at Uber. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, volume 00, pp. 103–110, Nov. 2018. doi: 10.1109/ICDMW.2017.19. URL doi.ieeecomputersociety.org/10.1109/ICDMW.2017.19.