Probabilistic Time Series Forecasting with Structured Shape and Temporal Diversity

Vincent Le Guen [1,2]
vincent.le-guen@edf.fr

Nicolas Thome [2]
nicolas.thome@cnam.fr

[1] EDF R&D, Chatou, France
[2] Conservatoire National des Arts et Métiers, CEDRIC, Paris, France

Abstract

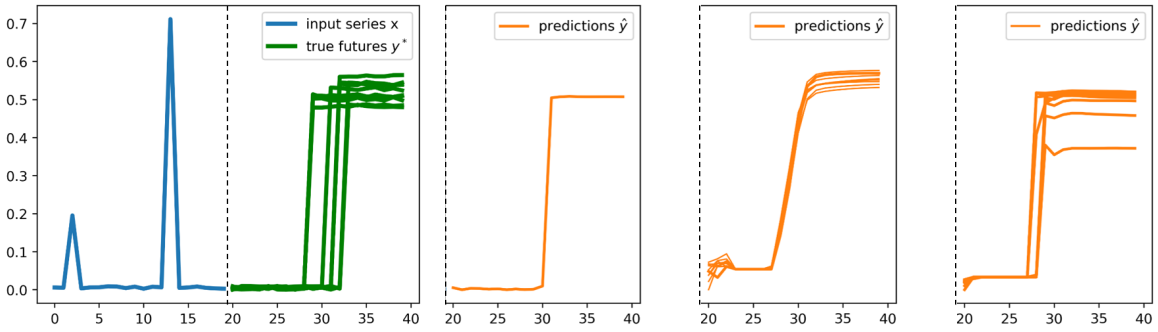
Probabilistic forecasting consists in predicting a distribution of possible future outcomes. In this paper, we address this problem for non-stationary time series, which is very challenging yet crucially important. We introduce the STRIPE model for representing structured diversity based on shape and time features, ensuring both probable predictions while being sharp and accurate. STRIPE is agnostic to the forecasting model, and we equip it with a diversification mechanism relying on determinantal point processes (DPP). We introduce two DPP kernels for modeling diverse trajectories in terms of shape and time, which are both differentiable and proved to be positive semi-definite. To have an explicit control on the diversity structure, we also design an iterative sampling mechanism to disentangle shape and time representations in the latent space. Experiments carried out on synthetic datasets show that STRIPE significantly outperforms baseline methods for representing diversity, while maintaining accuracy of the forecasting model. We also highlight the relevance of the iterative sampling scheme and the importance to use different criteria for measuring quality and diversity. Finally, experiments on real datasets illustrate that STRIPE is able to outperform state-of-the-art probabilistic forecasting approaches in the best sample prediction.

1 Introduction

Time series forecasting consists in analysing historical signal correlations to anticipate future outcomes. In this work, we focus on probabilistic forecasting in non-stationary contexts, i.e. we aim at producing plausible and diverse predictions where future trajectories can present sharp variations. This forecasting context is of crucial importance in many applicative fields, e.g. climate [62, 34, 15], optimal control or regulation [66, 41], traffic flow [39, 38], healthcare [8, 1], stock markets [14, 7], *etc*. Our motivation is illustrated in the example of the blue input in Figure 1(a): we aim at performing predictions covering the full distribution of future trajectories, whose samples are shown in green.

State-of-the-art methods for time series forecasting currently rely on deep neural networks, which exhibit strong abilities in modeling complex nonlinear dependencies between variables and time. Recently, increasing attempts have been made for improving architectures for accurate predictions [31, 53, 37, 42, 35] or for making predictions sharper, e.g. by explicitly modeling dynamics [9, 16, 50], or by designing specific loss functions addressing the drawbacks of blurred prediction with mean squared error (MSE) training [12, 47, 33, 58]. Although Figure 1(b) shows that such approaches produce sharp and realistic forecasts, their deterministic nature limits them to a single trajectory prediction without uncertainty quantification.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.



(a) True predictive distribution (b) Sharp loss [33] (c) deep stoch model [65] (d) STRIPE (ours)

Figure 1: We address the probabilistic time series forecasting problem. (a) Recent deep learning models include a specific loss enabling sharp predictions [12, 47, 33, 58] (b), but are inadequate for producing diverse forecasts. On the other hand, probabilistic forecasting approaches based on generative models [65, 46] loose the ability to generate sharp forecasts (c). The proposed STRIPE model (d) produces both sharp and diverse future forecasts.

Methods targeting probabilistic forecasting enable to sample diverse predictions from a given input. This includes deterministic methods that predict the quantiles of the predictive distribution or probabilistic methods that sample future values from a learned approximate distribution, parameterized explicitly (e.g. Gaussian [52, 45, 51]), or implicitly with latent generative models [65, 29, 46]. These approaches are commonly trained using MSE or variants for probabilisting forecasts, e.g. quantile loss [28], and consequently often loose the ability to represent sharp predictions, as shown in Figure 1(c) for [65]. These generative models also lack an explicit structure to control the type of diversity in the latent space.

In this work, we introduce a model for including Shape and Time diverRsIty in Probabilistic forEcasting (STRIPE). As shown in Figure 1(d), this enables to produce sharp and diverse forecasts, which fit well the ground truth distribution of trajectories in Figure 1(a).

STRIPE presented in section 3 is agnostic to the predictive model, and we use both deterministic or generative models in our experiments. STRIPE encompasses the following contributions. Firstly, we introduce a structured shape and temporal diversity mechanism based on determinantal point processes (DPP). We introduce two DPP kernels for modeling diverse trajectories in terms of shape and time, which are both differentiable and proved to be positive semi-definite (section 3.1). To have an explicit control on the diversity structure, we also design an iterative sampling mechanism to disentangle shape and time representations in the latent space (section 3.2).

Experiments are conducted in section 4 on synthetic datasets to evaluate the ability of STRIPE to match the ground truth trajectory distribution. We show that STRIPE significantly outperforms baseline methods for representing diversity, while maintaining the accuracy of the forecasting model. Experiments on real datasets further show that STRIPE is able to outperform state-of-the-art probabilistic forecasting approaches when evaluating the best sample (i.e. diversity), while being equivalent based on its mean prediction (i.e. quality).

2 Related work

Deterministic time series forecasting Traditional time series forecasting methods, including linear autoregressive models such as ARIMA [6] or exponential smoothing [27], handle linear dynamics and stationary time series (or made stationary by modeling trends and seasonality). Deep learning has become the state-of-the-art for automatically modeling complex long-term dependencies, with many works focusing on architecture design based on temporal convolution networks [5, 53], recurrent neural networks (RNNs) [31, 64, 44], or Transformer [57, 37]. Another crucial topic more recently studied in the non-stationary context is the choice of a suitable loss function. As an alternative to the mean squared error (MSE) largely used as a proxy, new differentiable loss functions were proposed to enforce more meaningful criteria such as shape and time [47, 12, 33, 58], e.g. soft-DTW based on

dynamic time warping [12, 4] or the DILATE loss with a soft-DTW term for shape and a smooth temporal distortion index (TDI) [20, 56] for accurate temporal localization. These works toward sharper predictions were however only studied in the context of deterministic predictions and not for multiple outcomes.

Probabilistic forecasting For describing the conditional distribution of future values given an input sequence, a first class of deterministic methods add variance estimation with Monte Carlo dropout [67, 32] or predict the quantiles of this distribution [61, 21, 60] by minimizing the pinball loss [28, 49] or the continuous ranked probability score (CRPS) [23]. Other probabilistic methods try to approximate the predictive distribution, *explicitly* with a parametric distribution (e.g. Gaussian for DeepAR [52] and variants [45, 51]), or *implicitly* with a generative model with latent variables (e.g. with conditional variational autoencoders (cVAEs) [65], conditional generative adversarial networks (cGANs) [29], normalizing flows [46]). However, these methods lack the ability to produce sharp forecasts by minimizing variants of the MSE (pinball loss, gaussian maximum likelihood), at the exception of cGANs - but which suffer from mode collapse that limits predictive diversity. Moreover, these generative models are generally represented by unstructured distributions in the latent space (e.g. Gaussian), which do not allow to have an explicit control on the targeted diversity.

Diverse predictions For improving the diversity of predictions, several repulsive schemes were studied such as the variety loss [26, 55] that consists in optimizing the best sample, or entropy regularization terms [13, 59] that encourage a uniform distribution and thus more diverse samples. Submodular distribution functions such as determinantal point processes (DPP) [30, 48, 40] are an appealing probabilistic tool to enforce structured diversity via the choice of a positive semi-definite kernel. DPPs has been successfully applied in various contexts, e.g. document summarization [24], recommendation systems [22], object detection [2], and very recently to image generation [17] and diverse trajectory forecasting [65]. GDPP [17] is based on matching generated and true sample diversity by aligning the corresponding DPP kernels, and thus limits their use in datasets where the full distribution of possible outcomes is accessible. In contrast, our approach is applicable in realistic scenarii where only a single label is available for each training sample. Although we share with [65] the goal to use DPP as diversification mechanism, the main limitation in [65] is to use the MSE loss for training the prediction and diversification models, leading to blurred prediction, as illustrated in Figure 1(c). Our approach is able to generate sharp and diverse predictions ; we also highlight the importance in STRIPE to use different criteria for training the prediction model (quality) and the diversification mechanism in order to make them cooperate.

3 Shape and time diversity for probabilistic time series forecasting

We introduce the STRIPE model for including shape and time diversity for probabilistic time series forecasting, which is depicted in Figure 2. Given an input sequence $\mathbf{x}_{1:T} = (\mathbf{x}_1, ..., \mathbf{x}_T) \in \mathbb{R}^{p \times T}$, our goal is to sample a set of N diverse and plausible future trajectories $\hat{\mathbf{y}}^{(i)} = (\hat{\mathbf{y}}_{T+1}, ..., \hat{\mathbf{y}}_{T+\tau}) \in \mathbb{R}^{d \times \tau}$ from the data future distribution $\hat{\mathbf{y}}^{(i)} \sim p(.|\mathbf{x}_{1:T})$.

STRIPE builds upon a general Sequence To Sequence (Seq2Seq) architecture dedicated to multi-step time series forecasting: the input time series $\mathbf{x}_{1:T}$ is fed into an encoder that summarizes the input into a latent vector h. Note that our method is agnostic to the specific choice of the forecasting model: it can be a deterministic RNN, or a probabilistic conditional generative model (e.g. cVAE [65], cGAN [29], normalizing flow [46]).

For training the predictor (upper part in Figure 2), we concatenate h with a vector $\mathbf{0}_k \in \mathbb{R}^k$ (free space left for the diversifying variables) and a decoder produces a forecasted trajectory $\hat{\mathbf{y}}^{(0)} = (\hat{\mathbf{y}}^{(0)}_{T+1}, ..., \hat{\mathbf{y}}^{(0)}_{T+\tau})$. The predictor minimizes a quality loss $\mathcal{L}_{quality}(\hat{\mathbf{y}}^{(0)}, \mathbf{y}^{(0)})$ between the predicted $\hat{\mathbf{y}}^{(0)}$ and ground truth future trajectory $\mathbf{y}^{(0)}$. In our non-stationary context, we train the STRIPE predictor with $\mathcal{L}_{quality}$ based on the recently proposed DILATE loss [33], that has proven successful for enforcing sharp predictions with accurate temporal localization.

For introducing structured diversity (lower part in Figure 2), we concatenate h with diversifying latent variables $z \in \mathbb{R}^k$ and produce N future trajectories $\left\{\hat{\mathbf{y}}^{(i)}\right\}_{i=1,..,N}$. Our key idea is to augment $\mathcal{L}_{quality}(\cdot)$ with a diversification loss $\mathcal{L}_{diversity}(\cdot; \mathcal{K})$ parameterized by diversity kernel $\mathcal{K}$ and

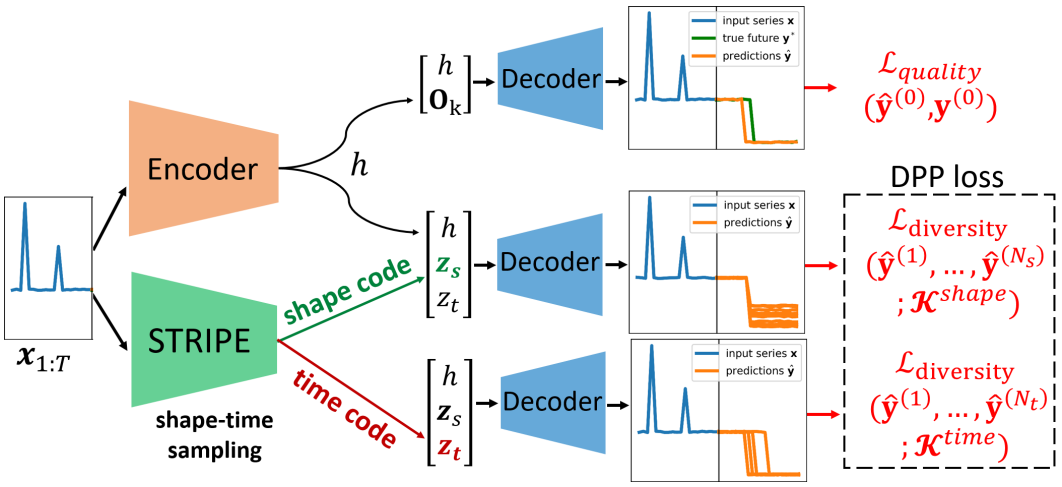


Figure 2: Our STRIPE model builds upon a Seq2Seq architecture trained with a quality loss $\mathcal{L}_{quality}$ enforcing sharp predictions. Our contributions rely on the design of a diversity loss $\mathcal{L}_{diversity}$ based on a specific Determinantal Point Processes (DPP). We design admissible shape and time DPP kernels, i.e. positive semidefinite, and differentiable for end-to-end training with deep models (section 3.1). We also introduce an iterative DDP sampling mechanism to generate disentangled latent codes between shape and time, supporting the use of different criteria for diversity and quality (section 3.2).

balanced by hyperparameter $\lambda \in \mathbb{R}$, leading to the overall objective training function:

$$\mathcal{L}_{STRIPE}(\hat{\mathbf{y}}^{(0)},...,\hat{\mathbf{y}}^{(N)},\mathbf{y}^{(0)};\mathcal{K}) = \mathcal{L}_{quality}(\hat{\mathbf{y}}^{(0)},\mathbf{y}^{(0)}) + \lambda\, \mathcal{L}_{diversity}(\hat{\mathbf{y}}^{(1)},...,\hat{\mathbf{y}}^{(N)};\mathcal{K}) \quad (1)$$

We highlight that STRIPE is applicable with a single target trajectory $\mathbf{y}^{(0)}$, i.e. we do not require the full trajectory distribution. We now detail how the $\mathcal{L}_{diversity}(\cdot;\mathcal{K})$ loss is designed to ensure diverse shape and time predictions.

3.1 STRIPE diversity module based on determinantal point processes

Our $\mathcal{L}_{diversity}$ loss relies on determinantal point processes (DPP) that are a convenient probabilistic tool for enforcing structured diversity via adequately chosen positive semi-definite kernels. For comparing two time series $\mathbf{y}_1$ and $\mathbf{y}_2$, we introduce the two following kernels $\mathcal{K}^{shape}$ and $\mathcal{K}^{time}$, for finely controlling the shape and temporal diversity:

$$\mathcal{K}^{shape}(\mathbf{y_1},\mathbf{y_2}) = e^{-DTW_\gamma(\mathbf{y_1},\mathbf{y_2})} \quad (2)$$

$$\mathcal{K}^{time}(\mathbf{y_1},\mathbf{y_2}) = TDI_\gamma(\mathbf{y_1},\mathbf{y_2}) = \frac{1}{Z}\sum_{\mathbf{A}\in\mathcal{A}_{\tau,\tau}} \langle \mathbf{A},\mathbf{\Omega}\rangle \exp^{-\frac{\langle \mathbf{A},\mathbf{\Delta}(\mathbf{y_1},\mathbf{y_2})\rangle}{\gamma}} \quad (3)$$

where $DTW_\gamma(\mathbf{y_1},\mathbf{y_2}) := -\gamma \log\left(\sum_{\mathbf{A}\in\mathcal{A}_{\tau,\tau}} \exp^{-\frac{\langle \mathbf{A},\mathbf{\Delta}(\mathbf{y_1},\mathbf{y_2})\rangle}{\gamma}}\right)$ is a smooth relaxation of Dynamic Time Warping (DTW) [12], and $\mathcal{K}^{time}$ corresponds to a smooth Temporal Distortion Index (TDI) [20, 33]: $\gamma > 0$ denotes the smoothing coefficient, $\mathbf{A} \subset \{0,1\}^{\tau\times\tau}$ is a warping path between two time series of length τ, $\mathcal{A}_{\tau,\tau}$ the set of all feasible warping paths and $\Delta(\mathbf{y_1},\mathbf{y_2}) = [\delta((\mathbf{y_1})_i,(\mathbf{y_2})_j)]_{1\leq i,j\leq\tau}$ is a pairwise cost matrix between time steps of both series with similarity measure $\delta : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, $\mathbf{\Omega}$ is a $\tau\times\tau$ matrix penalizing the deviation of warping paths from the main diagonal and Z is the partition function. These kernels are derived from the two components of the DILATE loss [33] ; however in contrast to the deterministic nature of DILATE, they are used in a probabilistic context for producing sharp and diverse forecasts.

$\mathcal{K}^{shape}$ and $\mathcal{K}^{time}$ are differentiable by design[1], making them suitable for end-to-end training with back-propagation. We also derive the key following result for ensuring the submodularity properties of DPPs, that we prove in supplementary 1:

[1] In the limit case $\gamma \to 0$, DTW_γ (resp. TDI_γ) recovers the standard DTW (resp. TDI).

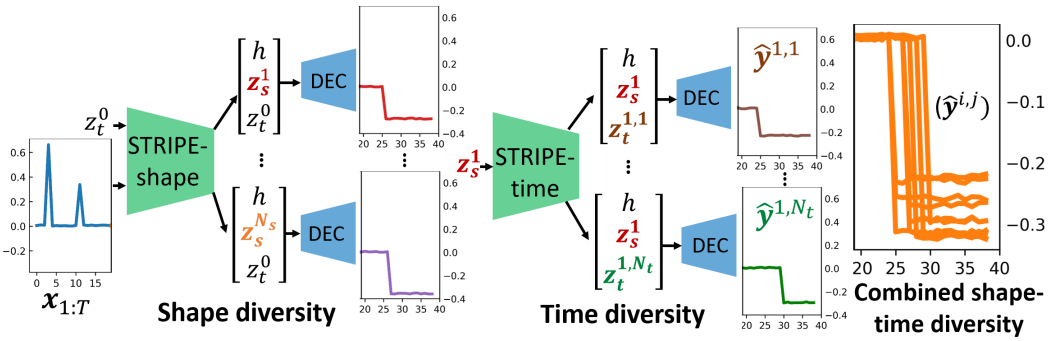


Figure 3: At test time, STRIPE sequential shape and time sampling scheme that leverages the disentangled latent space. STRIPE-shape first proposes diverse shape latent variables. For each generated shape, STRIPE-time further enhances its temporal variability, leading to a final set of accurate predictions with shape and time diversity.

Proposition 1. *Providing that κ is a positive semi-definite (PSD) kernel κ such that $\frac{\kappa}{1+\kappa}$ is also PSD, if we define the cost matrix Δ with general term $\delta(y_i, y_j) = -\gamma \log \kappa(y_i, y_j)$, then $\mathcal{K}^{shape}$ and $\mathcal{K}^{time}$ defined respectively in Equations (2) and (3) are PSD kernels.*

In practice, we choose $\kappa(u, v) = \frac{1}{2} e^{-\frac{(u-v)^2}{\sigma^2}} (1 - \frac{1}{2} e^{-\frac{(u-v)^2}{\sigma^2}})^{-1}$ that fullfills Prop 1 requirements.

DPP diversity loss We combine the two differentiable PSD kernels $\mathcal{K}^{shape}$ and $\mathcal{K}^{time}$ with the DPP diversity loss from [65] defined as the negative expected cardinality of a random subset Y (of a ground set $\mathcal{Y}$ of N items) sampled from the DPP of kernel $\mathcal{K}$ (denoted as $\mathbf{K}$ in matrix form of shape $N \times N$). This loss is differentiable and can be efficiently computed in closed-form:

$$\mathcal{L}_{diversity}(\mathcal{Y}; \mathbf{K}) = -\mathbb{E}_{Y \sim DPP(\mathbf{K})} |Y| = -Trace(\mathbf{I} - (\mathbf{K} + \mathbf{I})^{-1}) \tag{4}$$

Intuitively, a larger expected cardinality means a more diverse sampled set according to kernel $\mathcal{K}$. We provide more details on DPPs and the derivation of $\mathcal{L}_{diversity}$ in supplementary 2.

3.2 STRIPE learning and sequential shape and temporal diversity sampling

To maximize shape and time diversity with Eq (1) and (4), a naive way is to consider the combined kernel $\mathcal{K}^{shape} + \mathcal{K}^{time}$ which is also PSD. However, this reduces to using the same criterion for quality and diversity, i.e. DILATE [33]. This directly makes $\mathcal{L}_{diversity}$ conflicts with $\mathcal{L}_{quality}$ and harms prediction performances, as shown in ablation studies (section 4.2). Another simple solution is to diversify using $\mathcal{K}^{shape}$ and $\mathcal{K}^{time}$ independently, which prevents from modeling joint shape and time variations, and intrinsically limits the expressiveness of the diversification scheme. In contrast, we propose a sequential shape and temporal diversity sampling scheme, which enables to jointly model variations in shape and time without altering prediction quality. We now detail how the STRIPE models are trained and then used at test time.

STRIPE-shape and STRIPE-time learning We start by independently training two proposal modules STRIPE-shape and STRIPE-time (and their respective encoders and decoders) by optimizing Eq (1) with $\mathcal{L}_{STRIPE}(\cdot; \mathcal{K}^{shape})$ (resp. $\mathcal{L}_{STRIPE}(\cdot; \mathcal{K}^{time})$). To this end, we complement the latent state h of the forecaster with a diversifying latent variable $z \in \mathbb{R}^k$ decomposed into shape $z_s \in \mathbb{R}^{k/2}$ and temporal $z_t \in \mathbb{R}^{k/2}$ components: $z = (z_s, z_t) \in \mathbb{R}^k$. As illustrated in Figure 3, STRIPE-shape (the description of STRIPE-time is symmetric) is a proposal neural network that produces N_s different shape latent codes $z_s^{(i)}$ (the output of the STRIPE-shape neural network is of shape $N_s \times k$). The decoder takes the concatenated state $(h, z_s^{(i)}, z_t)$ for a fixed z_t and produces N_s future trajectories $\hat{\mathbf{y}}^{(i)}$, whose diversity is maximized with $\mathcal{L}_{diversity}(\hat{\mathbf{y}}^{(1)}, ..., \hat{\mathbf{y}}^{(N_s)}; \mathbf{K}^{shape})$ in Eq (4).

Sequential sampling at test time Once the STRIPE-shape and STRIPE-time models (and their corresponding encoders and decoders) are learned, test-time sampling (illustrated in Figure 3 and detailed in Algorithm 1) consists in sequentially maximizing the shape diversity with STRIPE-shape (different guesses about the step amplitude in Figure 3) and the temporal diversity of each shape with STRIPE-time (the temporal localization of the step).
Notice that the ordering shape+time is actually important since the notion of time diversity between two time series is only meaningful if they have a similar shape (so that computing the DTW optimal path has a sense).
As shown in our experiments, this two-steps scheme (denoted STRIPE S+T) leads to more diverse predictions with both shape and time criteria compared to using the shape or time kernels alone.

Algorithm 1: STRIPE S+T sampling at test time

Sample an initial $z_t^{(0)} \sim \mathcal{N}(0, \mathbf{I})$
$z_s^{(1)}, ..., z_s^{(N_s)} =$
STRIPE-shape$(\mathbf{x}_{1:T}, z_t^{(0)})$
for *i=1..*N_s **do**
  $z_t^{(i,1)}, ..., z_t^{(i,N_t)} =$
  STRIPE-time$(\mathbf{x}_{1:T}, z_s^{(i)})$
  for *j=1..*N_t **do**
    $\hat{\mathbf{y}}_{T+1:t+\tau}^{(i,j)} =$
    $Decoder(\mathbf{x}_{1:T}, (z_s^{(i)}, z_t^{(i,j)}))$
  end
end

4 Experiments

To illustrate the relevance of STRIPE, we carry out experiments in two different settings: in the first one, we compare the ability of forecasting methods to capture the full predictive distribution of future trajectories on a synthetic dataset with multiple possible futures for each input. To validate our approach in realistic settings, we evaluate STRIPE on 2 standard real datasets (traffic & electricity) where we evaluate the best (resp. the mean) sample metrics as a proxy for diversity (resp. quality).

Implementation details: To handle the inherent ambiguity of the synthetic dataset (multiple targets for one input), our STRIPE model is based on a natively stochastic model (cVAE). Since this situation does not arise exactly for real-world datasets, we choose in this case a deterministic Seq2Seq predictor with 1 layer of 128 Gated Recurrent Units (GRU) [10]. In our experiments, all methods produce N=10 future trajectories that are compared to the unique (or multiple) ground truth(s). For a fair comparison, STRIPE S+T generates $N_s \times N_t = 10 \times 10 = 100$ predictions and we randomly sample N=10 predictions for evaluation. Further neural network architectures and implementation details are described in supplementary 3.1. Our PyTorch code implementing STRIPE is available at https://github.com/vincent-leguen/STRIPE.

4.1 Synthetic dataset with multiple futures

We use a synthetic dataset similar to [33] that consists in predicting step functions based on a two-peaks input signal (see Figure 1). For each input series of 20 timesteps, we generate 10 different future series of length 20 by adding noise on the step amplitude and localisation. The dataset is composed of $100 \times 10 = 1000$ time series for each train/valid/test split (further dataset description in supplementary 3.1).

Metrics: In this multiple futures context, we define two specific discrepancy measures $H_{quality}(\ell)$ and $H_{diversity}(\ell)$ for assessing the divergence between the predicted and true distributions of futures trajectories for a given loss ℓ ($\ell =$ MSE or DILATE in our experiments):

$$H_{quality}(\ell) = \mathbb{E}_{\mathbf{x}\in\mathcal{D}_{test}} \mathbb{E}_{\hat{\mathbf{y}}} \left[\inf_{\mathbf{y}\in F(\mathbf{x})} \ell(\hat{\mathbf{y}}, \mathbf{y}) \right] \quad H_{diversity}(\ell) = \mathbb{E}_{\mathbf{x}\in\mathcal{D}_{test}} \mathbb{E}_{\mathbf{y}\in F(\mathbf{x})} \left[\inf_{\hat{\mathbf{y}}} \ell(\hat{\mathbf{y}}, \mathbf{y}) \right]$$

$\mathrm{H}_{quality}$ penalizes forecasts $\hat{\mathbf{y}}$ that are far away from a ground truth future of $\mathbf{x}$ denoted $\mathbf{y} \in F(\mathbf{x})$ (similarly to the precision concept in pattern recognition) whereas $\mathrm{H}_{diversity}$ penalizes when a true future is not covered by a forecast (similarly to recall). We also use the continuous ranked probability score (CRPS)[2] which is a standard *proper scoring rule* [23] for assessing probabilistic forecasts [21].

[2] An intuitive definition of the CRPS is the pinball loss integrated over all quantile levels. The CRPS is minimized when the predicted future distribution is identical to the true future distribution.

Table 1: Forecasting results on the synthetic dataset with multiple futures for each input, averaged over 5 runs (mean ± standard deviation). Best equivalent method(s) (Student t-test) shown in bold. Metrics are scaled (MSE × 1000, DILATE ×100, CRPS × 1000) for readability.

Methods	$H_{quality}$ (.)(↓)		$H_{diversity}$(.) (↓)		CRPS (↓)
	MSE	DILATE	MSE	DILATE	
Deep AR [52]	26.6 ± 6.4	67.0 ± 12.0	**15.2 ± 3.4**	45.4 ± 4.3	62.4 ± 9.9
cVAE MSE	11.8 ± 0.5	48.8 ± 3.2	20.0 ± 0.6	85.4 ± 7.0	76.4 ± 3.0
variety loss [55] MSE	13.1 ± 2.7	50.9 ± 4.7	19.6 ± 1.1	84.7 ± 2.2	80.1 ± 3.3
Entropy regul. [13] MSE	12.0 ± 0.7	51.5 ± 2.9	19.7 ± 0.7	89.5 ± 7.4	78.9 ± 2.9
Diverse DPP [65] MSE	15.9 ± 2.6	56.6 ± 2.8	16.5 ± 1.5	59.6 ± 5.6	80.5 ± 6.1
GDPP [17] kernel MSE	11.7 ± 1.3	47.5 ± 3.1	19.5 ± 0.4	82.3 ± 5.2	74.0 ± 4.5
STRIPE S+T (ours)	12.4 ± 1.0	48.7 ± 0.7	18.1 ± 1.6	62.0 ± 5.4	72.2 ± 3.1
cVAE DILATE	11.6 ± 1.8	**28.3 ± 2.9**	22.2 ± 2.5	67.8 ± 7.8	62.2 ± 4.2
variety loss [55] DILATE	14.9 ± 3.3	33.5 ± 1.9	23.8 ± 3.9	61.6 ± 1.9	62.6 ± 3.0
Entropy regul. [13] DILATE	12.7 ± 2.6	29.9 ± 3.2	23.5 ± 2.6	65.1 ± 4.5	62.4 ± 3.9
Diverse DPP [65] DILATE	**11.1 ± 1.6**	30.2 ± 2.9	20.7 ± 2.3	62.6 ± 11.3	**60.7 ± 1.6**
GDPP [17] kernel DILATE	**10.6 ± 1.6**	**28.7 ± 4.1**	21.7 ± 2.1	47.7 ± 9.0	63.4 ± 6.4
STRIPE S+T (ours)	**10.8 ± 0.4**	**30.7 ± 0.9**	**14.5 ± 0.6**	**35.5 ± 1.1**	**60.5 ± 0.4**

Results We compare our method to 4 recent competing diversification mechanisms (variety loss [55], entropy regularisation [13], diverse DPP [65] and GDPP [17]) based two different forecasting backbones: a conditional variational autoencoder (cVAE) trained with MSE and with DILATE. Results in Table 1 show that our model STRIPE S+T based on a cVAE DILATE obtains the global best performances by improving the diversity by a large margin ($H_{diversity}$(DILATE) = 35.5 vs. 67.8), significantly outperforming other methods. This highlights the relevance of the structured shape and time diversity in STRIPE. It is worth mentioning that STRIPE also presents the best performances in quality. In contrast, other diversification mechanisms (variety loss, entropy regularisation, diverse DPP) based on the same backbone (cVAE DILATE) improve the diversity in DILATE but at the cost of a loss in quality in MSE and/or DILATE. Although GDPP does not deteriorate quality, it is significantly worse than STRIPE in diversity, and the approach requires full future distribution supervision, which it not applicable in in real dataset (see section 2).

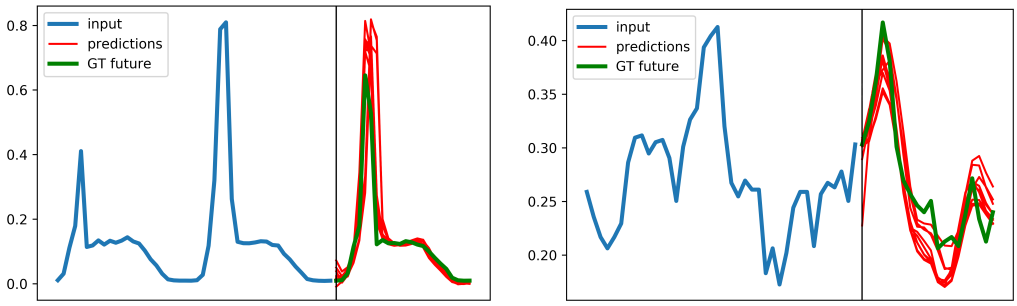
Similar conclusions can be drawn for the cVAE MSE backbone: the different diversity mechanisms improve the diversity but at the cost of a loss of quality. For example, Diverse DPP MSE [65] improves diversity ($H_{diversity}(DILATE) = 59.6$ vs. 85.4) but looses in quality ($H_{quality}(DILATE) = 56.6$ vs. 48.8). In contrast, STRIPE S+T again both improves diversity ($H_{diversity}(DILATE) = 62.0$ vs. 85.4) with equivalent quality ($H_{quality}(DILATE) = 48.7$ vs. 48.8). We further highlight that STRIPE S+T gets the best results evaluated in CPRS, confirming its ability to better recover the true future distribution.

4.2 Ablation study

To analyze the respective roles of the quality and diversity losses, we perform an ablation study on the synthetic dataset with the cVAE backbone trained with the quality loss DILATE and different DPP

Table 2: Ablation study on the synthetic dataset. We train a backbone cVAE with the DILATE quality loss and compare different DPP kernels for diversity. Metrics are scaled for readability. Results averaged over 5 runs (mean ± std). Best equivalent method(s) (Student t-test) shown in bold.

cVAE DILATE	$H_{quality}$(.) (↓)		$H_{diversity}$(.) (↓)				CRPS (↓)
diversity	MSE	DILATE	MSE	DTW	TDI	DILATE	
None	11.6 ± 1.8	**28.3 ± 2.9**	22.2 ± 2.5	18.8 ± 1.3	48.6 ± 2.2	67.8 ± 7.8	62.2 ± 4.2
DILATE	**11.1 ± 1.6**	**30.2 ± 2.8**	20.7 ± 2.3	18.6 ± 1.6	42.8 ± 10.2	62.6 ± 11.3	60.7 ± 1.7
MSE	**10.9 ± 1.5**	**30.2 ± 2.9**	20.1 ± 2.2	18.5 ± 1.3	41.9 ± 8.8	61.7 ± 9.5	62.1 ± 0.9
shape (ours)	**11.0 ± 1.4**	**30.2 ± 1.2**	15.5 ± 1.04	**16.4 ± 1.5**	15.4 ± 4.2	37.8 ± 3.7	63.2 ± 1.6
time (ours)	11.9 ± 0.5	**31.2 ± 1.3**	16.1 ± 0.70	17.6 ± 0.5	**15.1 ± 3.1**	38.9 ± 3.3	62.3 ± 1.4
S+T (ours)	**10.8 ± 0.4**	**30.7 ± 0.9**	**14.5 ± 0.6**	**16.1 ± 1.1**	**13.2 ± 1.7**	**35.5 ± 1.1**	**60.5 ± 0.4**



Traffic Electricity

Figure 4: Qualitative predictions for Traffic and Electricity datasets. Input series in blue are not shown entirely for readability. We display 10 future predictions of STRIPE S+T that are both sharp and accurate compared to the ground truth (GT) future in green.

diversity losses. For a finer analysis, we report in Table 2 the shape (DTW, computed with Tslearn [54]) and time (TDI) components of the DILATE loss [33].

Results presented in Table 2 first reveal the crucial importance to define different criteria for quality and diversity. With the same loss for quality and diversity (as this is the case in [65]), we observe here that the DILATE DPP kernel does not bring a statistically significant diversity gain compared to the cVAE DILATE baseline (without diversity loss). By choosing the MSE kernel instead, we even get a small diversity and quality improvement.

In contrast, our introduced shape and time kernels $\mathcal{K}^{shape}$ and $\mathcal{K}^{time}$ largely improve the diversity in DILATE without deteriorating precision. As expected, each kernel brings its own benefits: $\mathcal{K}^{shape}$ brings the best improvement in the shape metric DTW ($\text{H}_{diversity}(\text{DTW}) = 16.4$ vs. 18.8) and $\mathcal{K}^{shape}$ the best improvement in the time metric TDI ($\text{H}_{diversity}(\text{TDI}) = 15.1$ vs. 48.6). With our sequential shape and time sampling sheme described in section 3.2, STRIPE S+T gathers the benefits of both criteria and gets the global best results in diversity and equivalent results in quality.

4.3 State-of-the-art comparison on real-world datasets

We evaluate here the performances of STRIPE on two challenging real-world datasets commonly used as benchmarks in the time series forecasting literature [63, 52, 31, 45, 33, 53]: **Traffic**: consisting in hourly road occupancy rates (between 0 and 1) from the California Department of Transportation, and **Electricity**: consisting in hourly electricity consumption measurements (kWh) from 370 customers. For both datasets, models predict the 24 future points given the past 168 points (past week). Although these datasets present daily, weakly, yearly periodic patterns, we are more interested here in modeling finer intraday temporal scales, where these signals present sharp fluctuations that are crucial for many applications, e.g. short-term renewable energy forecasts for load adjustment in smart-grids [34].

Contrary to the synthetic dataset, we only dispose of one future trajectory sample $\mathbf{y}^{(0)}_{T+1:T+\tau}$ for each input series $\mathbf{x}_{1:T}$. In this case, the metrics $\text{H}_{quality}$ (resp. $\text{H}_{diversity}$) defined in section 4.1 reduce to the mean sample (resp. best sample), which are common for evaluating stochastic forecasting models [3, 19]. We also report the CRPS in supplementary 3.2.

Table 3: Forecasting results on the Traffic and Electricity datasets, averaged over 5 runs (mean $\pm$ std). Metrics are scaled for readability. Best equivalent method(s) (Student t-test) shown in bold.

	Traffic				Electricity			
	MSE		DILATE		MSE		DILATE	
Method	mean	best	mean	best	mean	best	mean	best
Nbeats [42] MSE	-	7.8 ± 0.3	-	22.1 ± 0.8	-	24.6 ± 0.9	-	29.3 ± 1.3
Nbeats [42] DILATE	-	17.1 ± 0.8	-	17.8 ± 0.3	-	38.9 ± 1.9	-	20.7 ± 0.5
Deep AR [52]	15.1 ± 1.7	$\mathbf{6.6 \pm 0.7}$	30.3 ± 1.9	16.9 ± 0.6	67.6 ± 5.1	25.6 ± 0.4	59.8 ± 5.2	17.2 ± 0.3
cVAE DILATE	$\mathbf{10.0 \pm 1.7}$	8.8 ± 1.6	$\mathbf{19.1 \pm 1.2}$	17.0 ± 1.1	$\mathbf{28.9 \pm 0.8}$	27.8 ± 0.8	24.6 ± 1.4	22.4 ± 1.3
Variety loss [55]	$\mathbf{9.8 \pm 0.8}$	7.9 ± 0.8	$\mathbf{18.9 \pm 1.4}$	15.9 ± 1.2	29.4 ± 1.0	27.7 ± 1.0	24.7 ± 1.1	21.6 ± 1.0
Entropy regul. [13]	11.4 ± 1.3	10.3 ± 1.4	$\mathbf{19.1 \pm 1.4}$	16.8 ± 1.3	34.4 ± 4.1	32.9 ± 3.8	29.8 ± 3.6	25.6 ± 3.1
Diverse DPP [65]	11.2 ± 1.8	6.9 ± 1.0	20.5 ± 1.0	14.7 ± 1.0	31.5 ± 0.8	25.8 ± 1.3	26.6 ± 1.0	19.4 ± 1.0
STRIPE S+T	$\mathbf{10.1 \pm 0.4}$	$\mathbf{6.5 \pm 0.2}$	$\mathbf{19.2 \pm 0.8}$	$\mathbf{14.2 \pm 0.2}$	29.7 ± 0.3	$\mathbf{23.4 \pm 0.2}$	$\mathbf{24.4 \pm 0.3}$	$\mathbf{16.9 \pm 0.2}$

Results in Table 3 reveal that STRIPE S+T outperforms all other methods in terms of the best sample trajectory evaluated in MSE and DILATE for both datasets, while being equivalent in the mean sample in 3/4 cases. Interestingly, STRIPE S+T provides better best trajectories (evaluated in MSE and DILATE) than the recent state-of-the-art N-Beats algorithm [42] (either trained with MSE or DILATE), which is dedicated to producing high quality deterministic forecasts. This confirms that STRIPE's structured quality and diversity framework enables to obtain very accurate best predictions. Finally when compared to the state-of-the art probabilistic deep AR method [52], STRIPE S+T is consistently better in diversity and quality.

We display a few qualitative forecasting examples of STRIPE S+T on Figure 4 and additional ones in supplementary 3.3. We observe that STRIPE predictions are both sharp and accurate: both the shape diversity (amplitude of the peaks) and temporal diversity match the ground truth future.

4.4 Model analysis

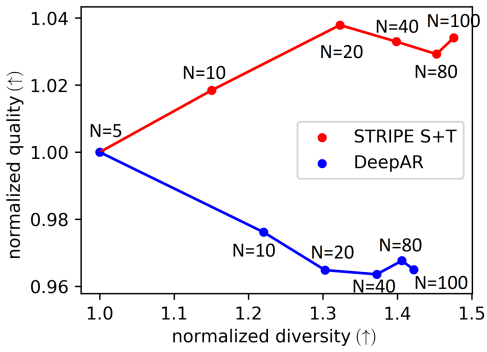


Figure 5: Influence of the number N of trajectories on quality (higher is better) and diversity for the synthetic dataset.

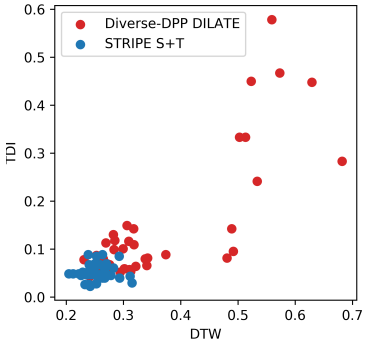


Figure 6: Scatterplot of 50 predictions in the plane (DTW,TDI), comparing STRIPE S+T v.s. Diverse DPP DILATE [65].

We analyze in Figure 5 for the synthetic dataset the evolution of performances when increasing the number N of sampled future trajectories from 5 to 100: we observe that this results in higher normalized DILATE diversity ($\mathrm{H}_{diversity}(5)/\mathrm{H}_{diversity}(N)$) for STRIPE S+T without deteriorating quality (which even increases slightly). In contrast, deepAR [52], which does not have control over the targetted diversity, increases diversity with N but at the cost of a loss in quality. This again confirms the relevance of our approach that effectively combines an adequate quality loss function and a structured diversity mechanism.

We provide an additional analysis to highlight the importance to separate the criteria for enforcing quality and diversity. In Figure 6, we represent 50 predictions from the models Diverse DPP DILATE [65] and STRIPE S+T in the plane (DTW,TDI). Diverse DPP DILATE [65] uses a DPP diversity loss based on the DILATE kernel, which is the same than for quality. We clearly see that the two objectives conflict: this model increases the DILATE diversity (by increasing the variance in the shape (DTW) or the time TDI) components) but a lot of these predictions have a high DILATE loss (worse quality). In contrast, STRIPE S+T predictions are diverse in DTW and TDI, and maintain an overall low DILATE loss. STRIPE S+T succeeds in recovering a set of good tradeoffs between shape and time leading a low DILATE loss.

5 Conclusion and perspectives

We present STRIPE, a probabilistic time series forecasting method that introduces structured shape and temporal diversity based on determinantal point processes. Diversity is controlled via two proposed differentiable positive semi-definite kernels for shape and time and exploits a forecasting model with a disentangled latent space. Experiments on synthetic and real-world datasets confirm that STRIPE leads to more diverse forecasts without sacrificing on quality. Ablation studies also reveal the crucial importance to decouple the criteria used for quality and diversity.

A future perspective would be to incorporate seasonality and extrinsic prior knowledge (such as special events) [32, 42] to better model the non-stationary abrupt changes and their impact on diversity

and model confidence [11]. Other appealing directions include diversity-promoting forecasting for exploration in reinforcement learning [43, 18, 36], and extension of structured diversity to spatio-temporal or video prediction tasks [62, 19, 25].

Broader Impact

Probabilistic time series forecasting, especially in the non-stationary contexts, is a paramount research problem with immediate and large impacts in the society. A wide range of sensitive applications heavily rely on accurate forecasts of uncertain events with potentially sharp variations for making crucial decisions: in weather and climate science, better anticipating floods, hurricanes, earthquakes or other extreme events evolution could help taking emergency measures on time and save lives; in medicine, better predictions of an outbreak's evolution is a particularly actual topic. We believe that introducing meaningful criteria such as shape and time, which are more related to application-specific evaluation metrics, is an important step toward more reliable and interpretable forecasts for decision makers.

References

[1] Ahmed M Alaa and Mihaela van der Schaar. Attentive state-space modeling of disease progression. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11334–11344, 2019.

[2] Samaneh Azadi, Jiashi Feng, and Trevor Darrell. Learning detection with diverse proposals. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7149–7157, 2017.

[3] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. *International Conference on Learning Representations (ICLR)*, 2018.

[4] Mathieu Blondel, Arthur Mensch, and Jean-Philippe Vert. Differentiable divergences between time series. *arXiv preprint arXiv:2010.08354*, 2020.

[5] Anastasia Borovykh, Sander Bohte, and Cornelis W Oosterlee. Conditional time series forecasting with convolutional neural networks. *arXiv preprint arXiv:1703.04691*, 2017.

[6] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[7] Philippe Chatigny, Jean-Marc Patenaude, and Shengrui Wang. Financial time series representation learning. *arXiv preprint arXiv:2003.12194*, 2020.

[8] Sucheta Chauhan and Lovekesh Vig. Anomaly detection in ECG time signals via deep long short-term memory networks. In *International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–7. IEEE, 2015.

[9] Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in neural information processing systems (NeurIPS)*, pages 6571–6583, 2018.

[10] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[11] Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2902–2913, 2019.

[12] Marco Cuturi and Mathieu Blondel. Soft-DTW: a differentiable loss function for time-series. In *International Conference on Machine Learning (ICML)*, pages 894–903, 2017.

[13] Adji B Dieng, Francisco JR Ruiz, David M Blei, and Michalis K Titsias. Prescribed generative adversarial networks. *arXiv preprint arXiv:1910.04302*, 2019.

[14] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. Deep learning for event-driven stock prediction. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

[15] Jérémie Donà, Jean-Yves Franceschi, Sylvain Lamprier, and Patrick Gallinari. PDE-driven spatiotemporal disentanglement. *arXiv preprint arXiv:2008.01352*, 2020.

[16] Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. Augmented neural ODEs. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3134–3144, 2019.

[17] Mohamed Elfeki, Camille Couprie, Morgane Riviere, and Mohamed Elhoseiny. GDPP: learning diverse generations using determinantal point process. *International Conference on Machine Learning (ICML)*, 2019.

[18] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *International Conference on Learning Representations (ICLR)*, 2019.

[19] Jean-Yves Franceschi, Edouard Delasalles, Mickaël Chen, Sylvain Lamprier, and Patrick Gallinari. Stochastic latent residual video prediction. *International Conference on Machine Learning (ICML)*, 2020.

[20] Laura Frías-Paredes, Fermín Mallor, Martín Gastón-Romeo, and Teresa León. Assessing energy forecasting inaccuracy by simultaneously considering temporal and absolute errors. *Energy Conversion and Management*, 142:533–546, 2017.

[21] Jan Gasthaus, Konstantinos Benidis, Yuyang Wang, Syama Sundar Rangapuram, David Salinas, Valentin Flunkert, and Tim Januschowski. Probabilistic forecasting with spline quantile function RNNs. In *The 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1901–1910, 2019.

[22] Jennifer A Gillenwater, Alex Kulesza, Emily Fox, and Ben Taskar. Expectation-maximization for learning determinantal point processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3149–3157, 2014.

[23] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

[24] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. In *Advances in neural information processing systems (NeurIPS)*, pages 2069–2077, 2014.

[25] Vincent Le Guen, Yuan Yin, Jérémie Dona, Ibrahim Ayed, Emmanuel de Bézenac, Nicolas Thome, and Patrick Gallinari. Augmenting physical models with deep networks for complex dynamics forecasting. *arXiv preprint arXiv:2010.04456*, 2020.

[26] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social GAN: Socially acceptable trajectories with generative adversarial networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2255–2264, 2018.

[27] Rob Hyndman, Anne B Koehler, J Keith Ord, and Ralph D Snyder. *Forecasting with exponential smoothing: the state space approach*. Springer Science & Business Media, 2008.

[28] Roger Koenker and Kevin F Hallock. Quantile regression. *Journal of economic perspectives*, 15(4):143–156, 2001.

[29] Alireza Koochali, Andreas Dengel, and Sheraz Ahmed. If you like it, gan it. probabilistic multivariate times series forecast with gan. *arXiv preprint arXiv:2005.01181*, 2020.

[30] Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286, 2012.

[31] Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 95–104, 2018.

[32] Nikolay Laptev, Jason Yosinski, Li Erran Li, and Slawek Smyl. Time-series extreme event forecasting with neural networks at Uber. In *International Conference on Machine Learning (ICML)*, volume 34, pages 1–5, 2017.

[33] Vincent Le Guen and Nicolas Thome. Shape and time distortion loss for training deep time series forecasting models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4191–4203. 2019.

[34] Vincent Le Guen and Nicolas Thome. A deep physical model for solar irradiance forecasting with fisheye images. In *CVPR 2020 OmniCV workshop*, 2020.

[35] Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[36] Edouard Leurent, Denis Efimov, and Odalric-Ambrym Maillard. Robust estimation, prediction and control with linear dynamics and generic costs. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

[37] Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5244–5254, 2019.

[38] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations (ICLR)*, 2018.

[39] Yisheng Lv, Yanjie Duan, Wenwen Kang, Zhengxi Li, and Fei-Yue Wang. Traffic flow prediction with big data: a deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):865–873, 2015.

[40] Zelda E Mariet, Yaniv Ovadia, and Jasper Snoek. DPPNet: Approximating determinantal point processes with deep networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3218–3229, 2019.

[41] Shamsul Masum, Ying Liu, and John Chiverton. Multi-step time series forecasting of electric load using machine learning models. In *International Conference on Artificial Intelligence and Soft Computing*, pages 148–159. Springer, 2018.

[42] Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *International Conference on Learning Representations (ICLR)*, 2020.

[43] Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 16–17, 2017.

[44] Yao Qin, Dongjin Song, Haifeng Cheng, Wei Cheng, Guofei Jiang, and Garrison W Cottrell. A dual-stage attention-based recurrent neural network for time series prediction. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2627–2633. AAAI Press, 2017.

[45] Syama Sundar Rangapuram, Matthias W Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski. Deep state space models for time series forecasting. In *Advances in neural information processing systems (NeurIPS)*, pages 7785–7794, 2018.

[46] Kashif Rasul, Abdul-Saboor Sheikh, Ingmar Schuster, Urs Bergmann, and Roland Vollgraf. Multi-variate probabilistic time series forecasting via conditioned normalizing flows. *arXiv preprint arXiv:2002.06103*, 2020.

[47] François Rivest and Richard Kohar. A new timing error cost function for binary time series prediction. *IEEE transactions on neural networks and learning systems*, 2019.

[48] Joshua Robinson, Suvrit Sra, and Stefanie Jegelka. Flexible modeling of diversity with strongly log-concave distributions. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 15199–15209, 2019.

[49] Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3538–3548, 2019.

[50] Yulia Rubanova, Tian Qi Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5321–5331, 2019.

[51] David Salinas, Michael Bohlke-Schneider, Laurent Callot, Roberto Medico, and Jan Gasthaus. High-dimensional multivariate forecasting with low-rank gaussian copula processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6824–6834, 2019.

[52] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020.

[53] Rajat Sen, Hsiang-Fu Yu, and Inderjit S Dhillon. Think globally, act locally: A deep neural network approach to high-dimensional time series forecasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4838–4847, 2019.

[54] Romain Tavenard, Johann Faouzi, Gilles Vandewiele, Felix Divo, Guillaume Androz, Chester Holtz, Marie Payne, Roman Yurchak, Marc Rußwurm, Kushal Kolar, et al. Tslearn, a machine learning toolkit for time series data. *Journal of Machine Learning Research*, 21(118):1–6, 2020.

[55] Luca Anthony Thiede and Pratik Prabhanjan Brahma. Analyzing the variety loss in the context of probabilistic trajectory prediction. In *International Conference on Computer Vision (ICCV)*, pages 9954–9963, 2019.

[56] Loïc Vallance, Bruno Charbonnier, Nicolas Paul, Stéphanie Dubost, and Philippe Blanc. Towards a standardized procedure to assess solar forecast accuracy: A new ramp and time alignment metric. *Solar Energy*, 150:408–422, 2017.

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems (NIPS)*, pages 5998–6008, 2017.

[58] Titouan Vayer, Laetitia Chapel, Nicolas Courty, Rémi Flamary, Yann Soullard, and Romain Tavenard. Time series alignment with global invariances. *arXiv preprint arXiv:2002.03848*, 2020.

[59] Dilin Wang and Qiang Liu. Nonlinear Stein variational gradient descent for learning diversified mixture models. In *International Conference on Machine Learning (ICML)*, pages 6576–6585, 2019.

[60] Ruofeng Wen and Kari Torkkola. Deep generative quantile-copula models for probabilistic forecasting. *ICML Time Series Workshop*, 2019.

[61] Ruofeng Wen, Kari Torkkola, Balakrishnan Narayanaswamy, and Dhruv Madeka. A multi-horizon quantile recurrent forecaster. *NeurIPS Time Series Workshop*, 2017.

[62] Shi Xingjian, Zhourong Chen, and Hao et al Wang. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[63] Hsiang-Fu Yu, Nikhil Rao, and Inderjit S Dhillon. Temporal regularized matrix factorization for high-dimensional time series prediction. In *Advances in neural information processing systems (NIPS)*, pages 847–855, 2016.

[64] Rose Yu, Stephan Zheng, Anima Anandkumar, and Yisong Yue. Long-term forecasting using tensor-train RNNs. *arXiv preprint arXiv:1711.00073*, 2017.

[65] Ye Yuan and Kris Kitani. Diverse trajectory forecasting with determinantal point processes. *International Conference on Learning Representations (ICLR)*, 2020.

[66] Jian Zheng, Cencen Xu, Ziang Zhang, and Xiaohua Li. Electric load forecasting in smart grids using long-short-term-memory based recurrent neural network. In *51st Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2017.

[67] Lingxue Zhu and Nikolay Laptev. Deep and confident prediction for time series at Uber. In *International Conference on Data Mining Workshops (ICDMW)*, pages 103–110. IEEE, 2017.

Probabilistic Time Series Forecasting with Structured Shape and Temporal Diversity Supplementary material

1 Proof of Proposition 1

We define the following kernels for comparing two trajectories $\mathbf{y} \in \mathbb{R}^{d\times\tau}$ and $\mathbf{z} \in \mathbb{R}^{d\times\tau}$:

$$\mathcal{K}^{shape}(\mathbf{y},\mathbf{z}) = e^{-DTW_\gamma(\mathbf{y},\mathbf{z})} \tag{1}$$

$$\mathcal{K}^{time}(\mathbf{y},\mathbf{z}) = \frac{1}{Z}\sum_{\mathbf{A}\in\mathcal{A}_{\tau,\tau}} \langle \mathbf{A}, \mathbf{\Omega}\rangle \exp^{-\frac{\langle \mathbf{A},\mathbf{\Delta}(\mathbf{y},\mathbf{z})\rangle}{\gamma}} \tag{2}$$

where $DTW_\gamma(\mathbf{y_1},\mathbf{y_2}) := -\gamma \log\left(\sum_{\mathbf{A}\in\mathcal{A}_{\tau,\tau}} \exp^{-\frac{\langle \mathbf{A},\mathbf{\Delta}(\mathbf{y_1},\mathbf{y_2})\rangle}{\gamma}}\right)$.

Proposition 1. *Providing that κ is a positive semi-definite (PSD) kernel κ such that $\frac{\kappa}{1+\kappa}$ is also PSD, if we define the cost matrix Δ with general term $\delta(y_i, z_j) = -\gamma \log \kappa(y_i, z_j)$, then $\mathcal{K}^{shape}$ and $\mathcal{K}^{time}$ defined respectively in Equations (1) and (2) are PSD kernels.*

Proof. The proof for $\mathcal{K}^{shape}$ is a direct consequence of Theorem 1 in [CVBM07]. Under the conditions that κ and $\frac{\kappa}{1+\kappa}$ are PSD kernels, Theorem 1 [CVBM07] states that for any alignment $\pi = (\pi_1, \pi_2)$ that respects the warping conditions, the following kernel K is also PSD:

$$\begin{aligned}
K(\mathbf{y},\mathbf{z}) &= \sum_{\pi}\prod_{i=1}^{|\pi|} \kappa\left(y_{\pi_1(i)}, z_{\pi_2(i)}\right) \\
&= \sum_{\pi}\prod_{i=1}^{|\pi|} \exp^{-\frac{\delta\left(y_{\pi_1(i)}, z_{\pi_2(i)}\right)}{\gamma}} \\
&= \sum_{\pi} \exp^{-\sum_{i=1}^{|\pi|}\frac{\delta\left(y_{\pi_1(i)}, z_{\pi_2(i)}\right)}{\gamma}} \\
&= \sum_{\mathbf{A}\in\mathcal{A}_{\tau,\tau}} \exp^{-\frac{\langle A,\Delta(\mathbf{y},\mathbf{z})\rangle}{\gamma}} \\
&= \exp^{-DTW_\gamma(\mathbf{y},\mathbf{z})} \\
&= \mathcal{K}^{shape}(\mathbf{y},\mathbf{z})
\end{aligned}$$

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Let $a_1, ..., a_N \in \mathbb{R}$ and $\mathbf{y}_1, ..., \mathbf{y}_N \in \mathbb{R}^{d \times \tau}$. If Ω is non-zero on the diagonal (*e.g.* $\Omega(a,b) = \mu + \frac{(a-b)^2}{k^2}$ with $\mu > 0$), then there exists $\varepsilon > 0$ such that $\frac{\langle \mathbf{A}, \Omega \rangle}{Z} \geq \varepsilon \ \ \forall \mathbf{A} \in \mathcal{A}_{\tau,\tau}$. Then:

$$\begin{aligned}
\sum_i \sum_j a_i a_j \, \mathcal{K}^{time}(\mathbf{y}_i, \mathbf{y}_j) &= \sum_i \sum_j a_i a_j \, \frac{1}{Z} \sum_{\mathbf{A} \in \mathcal{A}_{\tau,\tau}} \langle \mathbf{A}, \mathbf{\Omega} \rangle \exp^{-\frac{\langle \mathbf{A}, \mathbf{\Delta}(\mathbf{y}_i, \mathbf{y}_j) \rangle}{\gamma}} \\
&\geq \sum_i \sum_j a_i a_j \sum_{\mathbf{A} \in \mathcal{A}_{\tau,\tau}} \varepsilon \exp^{-\frac{\langle \mathbf{A}, \mathbf{\Delta}(\mathbf{y}_i, \mathbf{y}_j) \rangle}{\gamma}} \\
&= \varepsilon \sum_i \sum_j a_i a_j \, \mathcal{K}^{shape}(\mathbf{y}_i, \mathbf{y}_j) \geq 0
\end{aligned}$$

The last inequality holds since we have already proven that $\mathcal{K}^{shape}$ is a PSD kernel. This proves that $\mathcal{K}^{time}$ is a PSD kernel. □

The particular choice $\kappa(u,v) = \frac{1}{2} e^{-\frac{(u-v)^2}{\sigma^2}} (1 - \frac{1}{2} e^{-\frac{(u-v)^2}{\sigma^2}})^{-1}$ fullfills Prop 1 requirements: κ is indeed PSD as the infinite limit of a sequence of PSD kernels $\sum_{i=1}^{\infty} k^i = \frac{k}{1-k} = \kappa$, where k is a halved Gaussian PSD kernel: $k(u,v) = \frac{1}{2} e^{-\frac{(u-v)^2}{\sigma^2}}$.

2 Derivation of $\mathcal{L}_{diversity}$

Determinantal Point Processes (DPPs) [KT+12] are a probabilistic tool for describing the diversity of a ground set of items $\mathcal{S} = \{\mathbf{y}_1, ..., \mathbf{y}_N\}$. Diversity is controlled via the choice of a positive semi-definite (PSD) kernel $\mathcal{K}$ for comparing items. A DPP is a probability distribution over all subsets of $\mathcal{S}$ that assigns the following probability to a random subset $\mathbf{Y}$:

$$\mathcal{P}_{\mathbf{K}}(\mathbf{Y} = Y) = \frac{\det(\mathbf{K}_Y)}{\sum_{Y' \subseteq \mathcal{S}} \det(\mathbf{K}'_Y)} = \frac{\det(\mathbf{K}_Y)}{\det(\mathbf{K} + \mathbf{I})} \tag{3}$$

where $\mathbf{K}$ denotes the kernel in matrix form and $\mathbf{K}_A$ is its restriction to the elements indexed by A : $\mathbf{K}_A = [\mathbf{K}_{i,j}]_{i,j \in A}$.

Intuitively, a DPP encourages the selection of diverse elements from the ground set $\mathcal{Y}$. If $\mathcal{Y}$ is more diverse, a random subset $Y \sim DPP(\mathcal{K})$ sampled from the DPP will select more items, *i.e.* will have a larger cardinality. This idea is embedded into the diversity loss $\mathcal{L}_{diversity}$ proposed in [YK20]:

$$\mathcal{L}_{diversity}(\mathcal{K}) = -\mathbb{E}_{Y \sim DPP(\mathcal{K})} |Y| = -Trace(\mathbf{I} - (\mathbf{K} + \mathbf{I})^{-1}) \tag{4}$$

3 Experiments

3.1 Datasets and implementation details

Synthetic dataset We use a synthetic dataset similar to [LGT19] that consists in predicting sudden changes (step functions) based on a two-peaks input signal. For each time series, the 20 first timesteps are the inputs, and the last 20 steps the targets to forecast. In each series, the input range is composed of 2 peaks at random temporal positions i_1 and i_2 and random amplitudes j_1 and j_2 between 0 and 1, and the target range is composed of a step of amplitude $j_2 - j_1$ at stochastic position $i_2 + (i_2 - i_1) + randint(-3; 3)$. All time series are corrupted by an additive Gaussian white noise of variance 0.01.

The difference with [LGT19] is that for each input series, we generate 10 different future series of length 20 by adding noise on the step amplitude and localisation. The dataset is composed of $100 \times 10 = 1000$ time series for each train/valid/test split.

Neural network architectures For the synthetic dataset, we use a stochastic predictive model based on a conditional variational autoencoder (cVAE).The encoder of the cVAE is a RNN with 1

layer of 128 GRU units, followed by a MLP which outputs the mean and variance of the latent state Gaussian distribution. We fixed by cross-validation the size of the latent state to $k = 16$. The decoder is another RNN with $128 + 16 = 144$ GRU units responsible for producing the future trajectory.

For the real-world datasets, we use a deterministic predictive Seq2Seq model with 1 layer of 128 GRU units for the encoder, and $128 + 16 = 144$ units for the decoder.

In all experiments, the STRIPE proposal modules (STRIPE-shape and STRIPE-time) are composed of a RNN with a layer of 128 GRU units followed by an MLP with 3 layers of 512 neurons (with BatchNormalization and LeakyReLU activations) and a final linear layer to produce $N = 10$ latent codes of dimension $k/2 = 8$ (corresponding to the proposals for z_s or z_t).

STRIPE hyperparameters We cross-validated the relevant hyperparameters of STRIPE:

- λ : tradeoff between $\mathcal{L}_{quality}$ and $\mathcal{L}_{diversity}$. When increasing λ (see Figure 1), the diversity increases and stabilizes starting from 10^{-3}, without loosing on quality. We fixed $\lambda = 1$ in all experiments.
- k: dimension of the diversifying latent variables z. This dimension should be chosen relatively to the hidden size of the RNN encoders and decoders (128 in our experiments). We fixed $k = 16$ in all cases.
- N: the number of future trajectories to sample. We fixed $N = 10$. We performed a sensibility analysis to this parameter in paper section 4.4.

For computing the DILATE loss, we used the parameters recommended in paper [LGT19] ($\gamma = 0.01, \alpha = 0.5$).

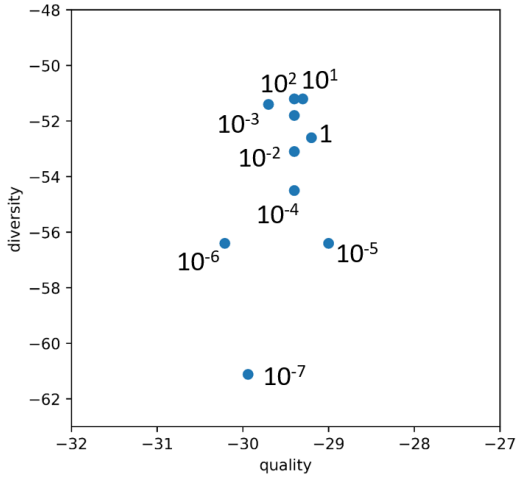


Figure 1: Influence of the hyperparameter λ balancing $\mathcal{L}_{quality}$ and $\mathcal{L}_{diversity}$ for the synthetic dataset. Quality (resp. diversity) are represented by $-\text{H}_{quality}(DILATE)$ (resp. $-\text{H}_{diversity}(DILATE)$), higher is better. When λ increases, diversity increases without deteriorating quality.

3.2 Full state-of-the-art comparison results

We provide here (Table 1) the full results of the state-of-the-art comparison (Table **??** in paper). We report the additional CRPS metric. We observe that STRIPE S+T obtains the best results evaluated in CRPS on the Electricity dataset (equivalent to DeepAR [SFGJ20]), and the second best results on the Traffic dataset (only behind DeepAR that is otherwise far worse in diversity and quality).

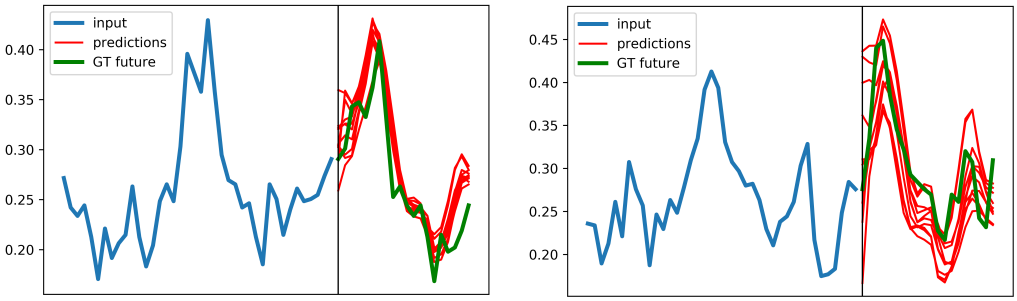
Table 1: Forecasting results on the Traffic and Electricity datasets, averaged over 5 runs (mean ± std). Metrics are scaled for readability. Best equivalent method(s) (Student t-test) shown in bold.

	Traffic					Electricity				
	MSE (× 1000)		DILATE (× 100)			MSE		DILATE		
Method	mean	best	mean	best	CRPS	mean	best	mean	best	CRPS
Nbeats MSE [OCCB20]	-	7.8 ± 0.3	-	22.1 ± 0.8	37.1 ± 0.9	-	24.6 ± 0.9	-	29.3 ± 1.3	36.3 ± 0.6
Nbeats DILATE	-	17.1 ± 0.8	-	17.8 ± 0.3	51.0 ± 2.6	-	38.9 ± 1.9	-	20.7 ± 0.5	47.5 ± 0.5
Deep AR [**?**]	15.1 ± 1.7	**6.6 ± 0.7**	30.3 ± 1.9	16.9 ± 0.6	**24.6 ± 1.1**	67.6 ± 5.1	25.6 ± 0.4	59.8 ± 5.2	17.2 ± 0.3	**34.5 ± 0.3**
cVAE DILATE	**10.0 ± 1.7**	8.8 ± 1.6	**19.1 ± 1.2**	17.0 ± 1.1	34.4 ± 2.5	**28.9 ± 0.8**	27.8 ± 0.8	24.6 ± 1.4	22.4 ± 1.3	39.2 ± 0.5
Variety loss [TB19]	**9.8 ± 0.8**	7.9 ± 0.8	**18.9 ± 1.4**	15.9 ± 1.2	32.4 ± 1.4	29.4 ± 1.0	27.7 ± 1.0	24.7 ± 1.1	21.6 ± 1.0	39.5 ± 0.8
Entropy regul. [DRBT19]	11.4 ± 1.3	10.3 ± 1.4	**19.1 ± 1.4**	16.8 ± 1.3	37.0 ± 2.7	34.4 ± 4.1	32.9 ± 3.8	29.8 ± 3.6	25.6 ± 3.1	42.4 ± 2.3
Diverse DPP [YK20]	11.2 ± 1.8	6.9 ± 1.0	20.5 ± 1.0	14.7 ± 1.0	30.9 ± 2.0	31.5 ± 0.8	25.8 ± 1.3	26.6 ± 1.0	19.4 ± 1.0	36.6 ± 0.9
STRIPE S+T	**10.1 ± 0.4**	**6.5 ± 0.2**	**19.2 ± 0.8**	**14.2 ± 0.2**	29.8 ± 0.3	29.7 ± 0.3	**23.4 ± 0.2**	**24.4 ± 0.3**	**16.9 ± 0.2**	**34.8 ± 0.4**

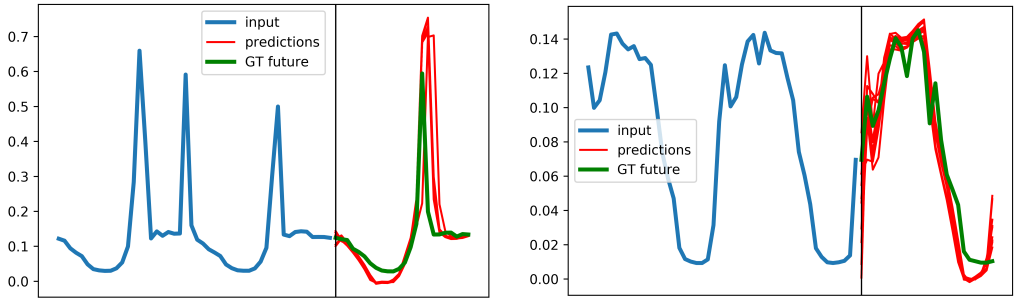
3.3 Additional visus

Wee provide additional visualizations for the Traffic and Electricity datasets that confirm that STRIPE S+T predictions are both diverse and sharp.

3.3.1 Electricity



3.3.2 Traffic



References

[CVBM07] Marco Cuturi, Jean-Philippe Vert, Oystein Birkenes, and Tomoko Matsui, *A kernel for time series based on global alignments*, 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, vol. 2, IEEE, 2007, pp. II–413.

[DRBT19] Adji B Dieng, Francisco JR Ruiz, David M Blei, and Michalis K Titsias, *Prescribed generative adversarial networks*, arXiv preprint arXiv:1910.04302 (2019).

[KT+12] Alex Kulesza, Ben Taskar, et al., *Determinantal point processes for machine learning*, Foundations and Trends in Machine Learning **5** (2012), no. 2–3, 123–286.

[LGT19] Vincent Le Guen and Nicolas Thome, *Shape and time distortion loss for training deep time series forecasting models*, Advances in Neural Information Processing Systems (NeurIPS), 2019, pp. 4191–4203.

[OCCB20] Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio, N-BEATS*: Neural basis expansion analysis for interpretable time series forecasting*, International Conference on Learning Representations (ICLR) (2020).

[SFGJ20] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski, *Deep*AR*: Probabilistic forecasting with autoregressive recurrent networks*, International Journal of Forecasting **36** (2020), no. 3, 1181–1191.

[TB19] Luca Anthony Thiede and Pratik Prabhanjan Brahma, *Analyzing the variety loss in the context of probabilistic trajectory prediction*, International Conference on Computer Vision (ICCV), 2019, pp. 9954–9963.

[YK20] Ye Yuan and Kris Kitani, *Diverse trajectory forecasting with determinantal point processes*, International Conference on Learning Representations (ICLR) (2020).