

Multivariate Time Series Forecasting with NBEATS

Thesis Defence

Philip Kurzendörfer

Supervisors

Lars SCHMIDT-THIEME
Rafael RÊGO DRUMOND

University of Hildesheim, ISMLL

January 25, 2022

Outline

Introduction

Related Work

Methodology

Experiments

Conclusion

Introduction

Related Work

Methodology

Experiments

Conclusion

Introduction

- ▶ Predicting the future relevant task
- ▶ e.g. in business world: Optimize decision making
- ▶ Real-world applications often require the forecasting of multiple related time series
- ▶ Relevance in demand forecasting or stock price forecasting.
- ▶ Capturing dynamics between time series

Introduction

- ▶ N-BEATS published in ICLR 2020
- ▶ State-of-art in univariate time series forecasting (M3, M4, TOURISM) [5]
- ▶ Normalizing flow conditioned on transformer model [6]
- ▶ Published in ICLR 2021
- ▶ State-of-art in multivariate time series forecasting on data sets used in this thesis

Introduction

Research Question

- If and how N-BEATS can be extended for multivariate time series (M-N-BEATS)

Contributions

- (i) Naive approach of multivariate N-BEATS
- (ii) Next, M-N-BEATS, adjusts vanilla N-BEATS architecture at block level
- (iii) M-N-BEATS is combined with a normalizing flow model

Introduction

Related Work

Methodology

Experiments

Conclusion

Related Work

Notation and Problem Formulation I

- ▶ Time series $\mathbf{x} \in \mathbb{R}^S$
- ▶ $\mathbf{x} = (x_1, x_2, \dots, x_S)^T$
- ▶ Equally spaced time series
- ▶ Forecasts of the next consecutive H future values $\mathbf{y} \in \mathbb{R}^H$
- ▶ $\mathbf{y} = (y_1, y_2, \dots, y_H)^T$

Related Work

Notation and Problem Formulation II

- ▶ Multivariate time series $\mathbf{X} \in \mathbb{R}^{S \times D}$
- ▶ $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_S)^T$
- ▶ \mathbf{x}_i is a D -dimensional vector
- ▶ Forecast the next H consecutive future values
 $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_T)^T$
- ▶ $\hat{\mathbf{Y}} \in \mathbb{R}^{H \times D}$, $\hat{\mathbf{Y}} = (\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \dots, \hat{\mathbf{y}}_H)^T$

Related Work

N-BEATS

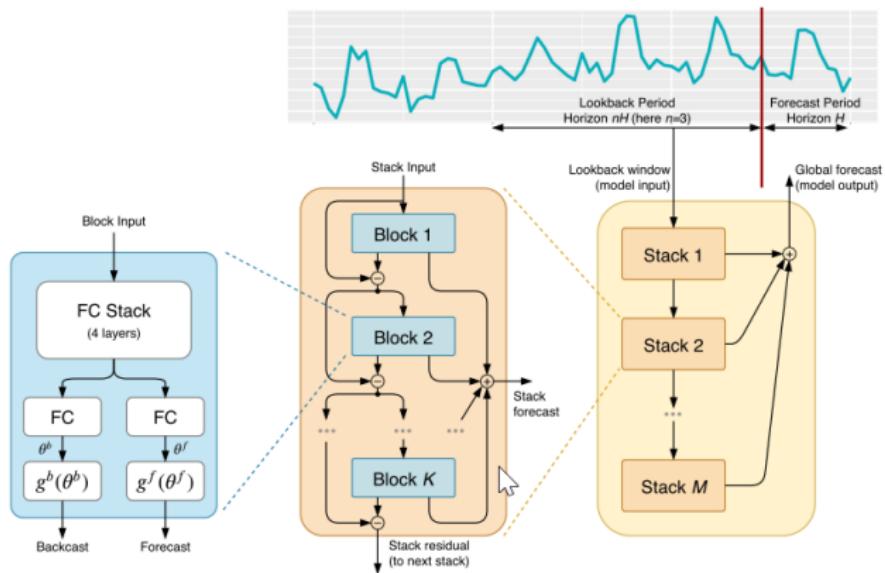


Figure 1: N-BEATS architecture.
source: [5].

Related Work

N-BEATS

Given block ℓ

- ▶ Basis parameters θ_ℓ^b and θ_ℓ^f
- ▶ Using the basis function g
- ▶ Forecast $\hat{\mathbf{y}}_\ell = g_\ell^f(\theta_\ell^f)$
- ▶ Backcast $\hat{\mathbf{x}}_\ell = g_\ell^b(\theta_\ell^b)$

Related Work

N-BEATS

Basis function g is given by

$$g_\ell^f(\theta_\ell^f) = \hat{\mathbf{y}}_\ell = \mathbf{V}_\ell^f \theta_\ell^f, \quad g_\ell^b(\theta_\ell^b) = \hat{\mathbf{x}}_\ell = \mathbf{V}_\ell^b \theta_\ell^b \quad (1)$$

with the basis $\mathbf{V}_\ell^f \in \mathbb{R}^{H \times \text{dim}(\theta^f)}$ and $\mathbf{V}_\ell^b \in \mathbb{R}^{S \times \text{dim}(\theta^b)}$.

Related Work

N-BEATS

- ▶ Basis vectors \mathbf{V} can be learnable parameters (generic variant)
- ▶ Can be fixed introducing an inductive bias (interpretable variant)
- ▶ Standard basis

Related Work

N-BEATS

Blocks are connected with so called Doubly residual connections defined by

$$\mathbf{x}_\ell = \mathbf{x}_{\ell-1} - \hat{\mathbf{x}}_{\ell-1}, \quad \hat{\mathbf{y}} = \sum_\ell \hat{\mathbf{y}}_\ell \quad (2)$$

- ▶ \mathbf{x}_ℓ block input of block ℓ
- ▶ $\mathbf{x}_{\ell-1}$ block input of block $\ell - 1$
- ▶ $\hat{\mathbf{x}}_{\ell-1}$ is the block forecast block $\ell - 1$

Related Work

Conditioned Normalizing Flow

Normalizing Flow

- ▶ Approximate an unknown distribution
- ▶ Transformer + normalizing flow
- ▶ LSTM + normalizing flow
- ▶ probabilistic forecasts
- ▶ Baseline for this thesis

Introduction

Related Work

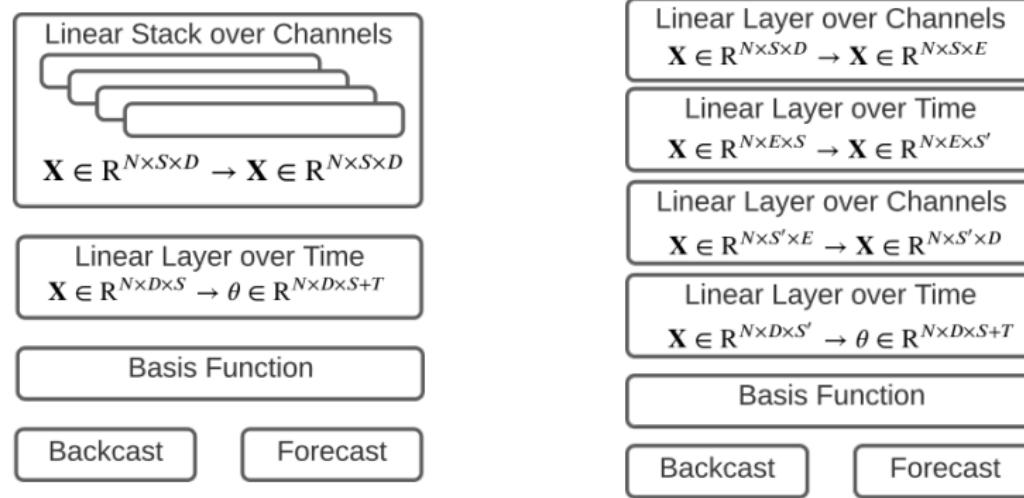
Methodology

Experiments

Conclusion

Methodology

M-N-BEATS Blocks



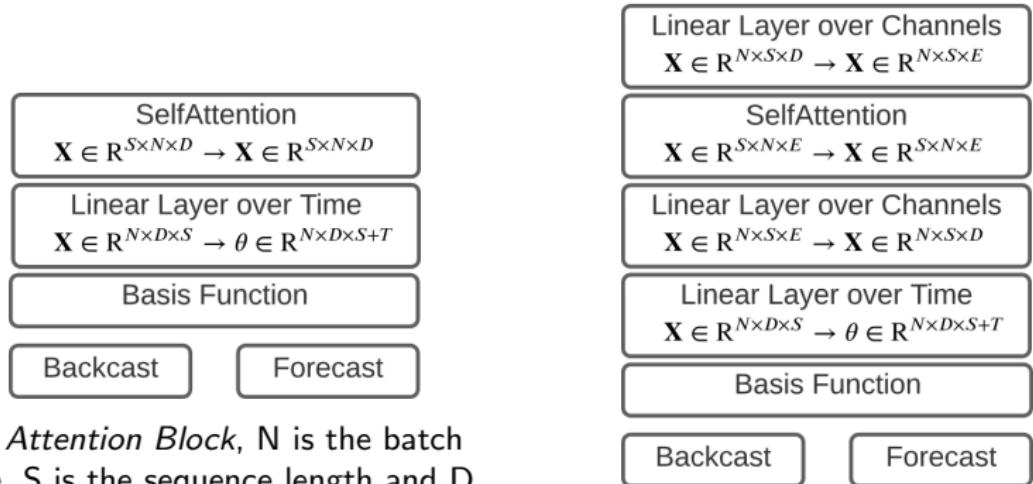
(a) *Simple Block*, N is the batch size, S is the sequence length and D is the number of time series.

(b) *Linear Block*, N is the batch size, S is the sequence length and D is the number of time series.

Figure 2: Multivariate N-BEATS blocks using linear layers.

Methodology

M-N-BEATS Blocks



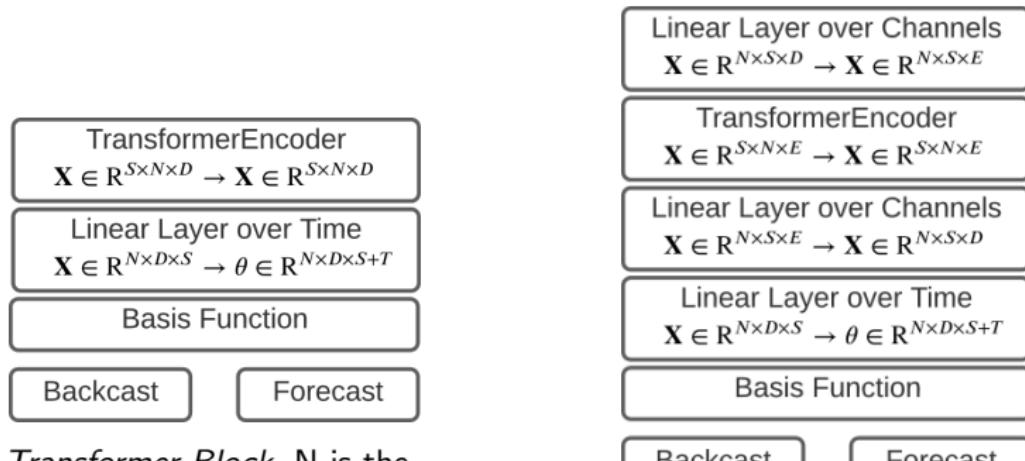
(a) *Attention Block*, N is the batch size, S is the sequence length and D is the number of time series.

(b) *Linear Attention Block*, N is the batch size, S is the sequence length and D is the number of time series.

Figure 3: Multivariate N-BEATS blocks using attention module.

Methodology

M-N-BEATS Blocks



(a) *Transformer Block*, N is the batch size, S is the sequence length and D is the number of time series.

(b) *Linear Transformer Block*, N is the batch size, S is the sequence length and D is the number of time series.

Figure 4: Multivariate N-BEATS blocks using transformer encoder layer.

Methodology

M-N-BEATS Blocks

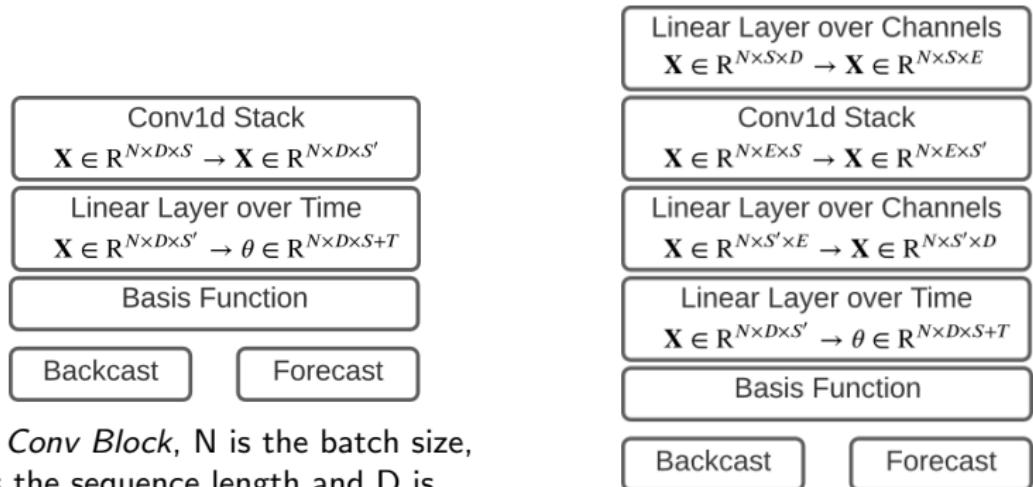


Figure 5: Multivariate N-BEATS blocks using convolutional layers.

Methodology

Normalizing Flow

- ▶ RV Z with simple base distribution p_Z
- ▶ Transform the simple distribution into a more complex target distribution p_X
- ▶ Mapping $f : \mathcal{X} \mapsto \mathcal{Z}$
- ▶ $z = f(x) = f_K \circ \dots \circ f_2 \circ f_1(x)$
- ▶ f is a composition of bijections or invertible functions (transformations)
- ▶ for any x there is exactly one z that could have been sampled to produce x
- ▶ By applying $x = f^{-1}(z)$
- ▶ should be easy to compute

Methodology

Normalizing Flow

- ▶ Apply change of variable formula
- ▶ Target distribution of X can be expressed in terms of the known base distribution $p_{\mathcal{Z}}$ and the Jacobian matrix of $f(\mathbf{x})$

$$p_{\mathcal{X}}(\mathbf{x}) = p_{\mathcal{Z}}(\mathbf{z}) \left| \det \left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \right) \right| = p_{\mathcal{Z}}(\mathbf{z}) \left| \det \left(\frac{\partial f^{-1}(\mathbf{z})}{\partial \mathbf{z}} \right) \right|^{-1} \quad (3)$$

Methodology

Normalizing Flow

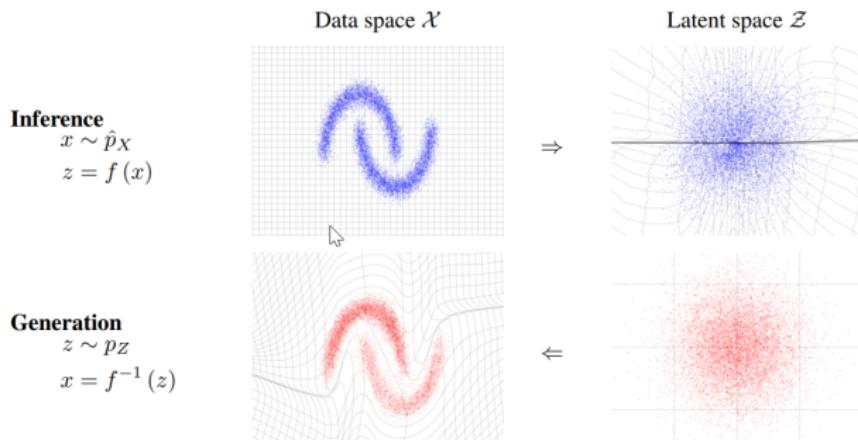


Figure 6: Normalizing Flow, RealNVP mapping learned on toy dataset.
Source: [1].

Introduction

Related Work

Methodology

Experiments

Conclusion

Experiments

Proposed Models and Baselines

- ▶ N-BEATS-Naive
- ▶ M-N-BEATS
- ▶ M-N-BEATS-Flow
- ▶ Transformer+Flow [6]
- ▶ LSTM+Flow [6]
- ▶ GP-Copula [7]

Experiments

Loss Functions

sMAPE for single time series

$$\text{sMAPE} = \frac{200}{H} \sum_{i=1}^H \frac{|y_{T+i} - \hat{y}_{T+i}|}{|y_{T+i}| + |\hat{y}_{T+i}|} \quad (4)$$

- ▶ y_{T+i} is the target value
- ▶ \hat{y}_{T+i} is the forecast for a single time step of a single time series
- ▶ For multivariate time series, mean of sMAPE of all channels

Experiments

Loss Functions

M-N-BEATS-Flow

- Maximize the average log likelihood

$$\mathcal{L} = \frac{1}{|\mathcal{D}|T} \sum_{\mathbf{x}_{1:T} \in \mathcal{D}} \sum_{t=1}^T \log p_{\mathcal{X}}(\mathbf{x}_t | \mathbf{h}_t; \theta) \quad (5)$$

Experiments

Metrics

- ▶ Mean Squared Error (*MSE*)
- ▶ Continuous Ranked Probability Score (*CRPS*)
- ▶ $CRPS_{sum}$

$$MSE(\mathbf{Y}, \hat{\mathbf{Y}}) = \left\| \mathbf{Y} - \hat{\mathbf{Y}} \right\|_F^2 \quad (6)$$

where $\|\cdot\|_F$ is the Frobenius norm.

Experiments

Metrics

CRPS for a given fixed data point x at time step t is given as

$$\text{CRPS}(F, x) = \int_{\mathbb{R}} (F(z) - \mathbb{I}\{x \leq z\})^2 dz \quad (7)$$

- ▶ $\mathbb{I}\{x \leq z\})^2$ is 1 if $x \leq z$ and 0 otherwise
- ▶ Measures how good the approximated distribution compares to the data using the CDF F

Experiments

Metrics

$CRPS_{sum}$ for a single time step t is given by

$$CRPS_{sum} = \mathbb{E}_t \left[CRPS \left(\hat{F}_{sum}(t), \sum_{i=0}^D x_t^i \right) \right] \quad (8)$$

- $\hat{F}_{sum}(t)$ is computed by summing up the individual \hat{F} over all D time series

Experiments

Data Sets

Table 4.1: Data Set Description

DATA SET	DIMENSION D	DOMAIN	FREQ.	TOTAL TIME STEPS	PREDICTION LENGTH
<i>Exchange</i>	8	\mathbb{R}^+	DAILY	6, 071	30
<i>Solar</i>	137	\mathbb{R}^+	HOURLY	7, 009	24
<i>Electricity</i>	370	\mathbb{R}^+	HOURLY	5, 790	24
<i>Traffic</i>	963	(0, 1)	HOURLY	10, 413	24

Source: [Rasul et al., 2021], [Salinas et al., 2019]

Experiments

Data Pipeline

Additional features

- ▶ Static features (embedding of channel index)
- ▶ Time dependent features (**day of week**)
- ▶ **Lagged time series**

Scaling

- ▶ Magnitudes of time series differ strongly
- ▶ Divide each time series by its training window mean
- ▶ Target divided by the same factor

Train test split

- ▶ Official train test splits are available
- ▶ The training set is split into train and validation set (70%)

Experiments

N-BEATS-Naive

- ▶ Flattening the **multivariate input** into a very long one-dimensional vector
- ▶ Apply vanilla N-BEATS
- ▶ Mean Results of 10 runs are reported
- ▶ Results are not competitive (see Appendix)

Experiments

N-BEATS-Naive

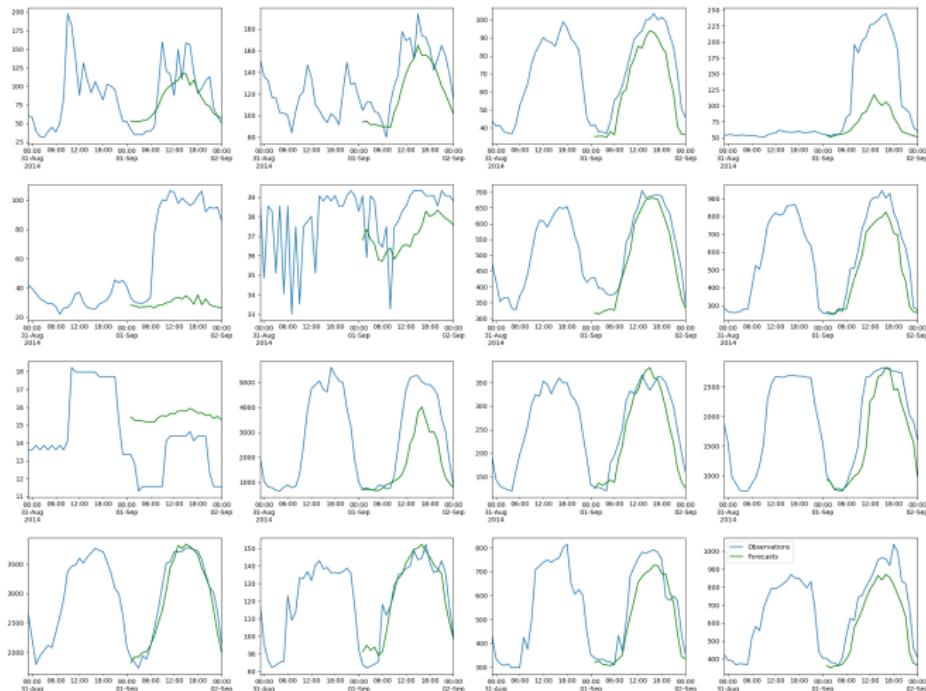


Figure 7: Forecasts produced by N-BEATS-Naive on the *Electricity* data set, only the forecasts of a single run are plotted.

Experiments

Hyperparameter Search

- ▶ M-N-BEATS
- ▶ M-N-BEATS-Flow
- ▶ Linear-Attention Block
- ▶ Linear-Transformer Block

Learning Rate	[0.0001, 0.00001]
Number of Blocks	[3, 5, 10]
Attention Heads	[4, 8]
Attention Embedding Size	[32, 64, 28]
Flow Type	[RealINVP, MAF]

Table 1: Hyperparameter Grid. Flow Type only applies for M-N-BEATS-Flow. The same grid has been applied for each data set independently.

Experiments

Main Results

- ▶ Best Model from HP-search is trained 10 times
- ▶ Average and standard deviation metrics on test set are reported
- ▶ For M-N-BEATS and M-N-BEATS-Flow,
linear-transformer-block performed better

Experiments

Main Results Baseline Comparison

MSE

- ▶ M-N-BEATS-Flow with *Linear-Transformer-Block*
- ▶ Mean and standard deviation of 10 runs
- ▶ Results taken from baselines papers (3 runs and 1 run)

Data Set	GP-Copula	LSTM-RealNVP	LSTM-MAF	Transformer-MAF	M-N-BEATS-Flow (Transformer)
Electricity	$240,000 \pm 55,000$	250,000	180,000	200,000	$155,810 \pm 26,344$
Traffic	0.00069 ± 0.000022	0.00069	0.00049	0.00050	0.00045 ± 0.0000
Solar	980 ± 52	910	980	930	925 ± 41
Exchange	0.00017 ± 0.000016	0.00024	0.00038	0.00034	0.000174 ± 0.0000

Table 2: *MSE* for baseline methods, M-N-BEATS-Flow with *Linear-Transformer-Block*

Experiments

Main Results Baseline Comparison

CRPS

- ▶ M-N-BEATS-Flow with *Linear-Transformer-Block*
- ▶ Baselines papers (3 runs and 3 runs)

Data Set	GP-Copula	LSTM-RealNVP	LSTM-MAF	Transformer-MAF	M-N-BEATS-Flow (Transformer)
<i>Electricity</i>	0.056 ± 0.002	0.059 ± 0.001	0.051 ± 0.000	0.052 ± 0.000	0.0511 ± 0.0019
<i>Traffic</i>	0.133 ± 0.001	0.172 ± 0.001	0.124 ± 0.002	0.134 ± 0.001	0.1148 ± 0.0039
<i>Solar</i>	0.371 ± 0.022	0.365 ± 0.02	0.378 ± 0.032	0.368 ± 0.001	0.3829 ± 0.0104
<i>Exchange</i>	0.008 ± 0.000	0.010 ± 0.001	0.012 ± 0.003	0.012 ± 0.003	0.0092 ± 0.0001

Table 3: CRPS for baseline methods, M-N-BEATS-Flow with *Linear-Transformer-Block*

Experiments

Main Results Baseline Comparison

$CRPS_{sum}$

- ▶ M-N-BEATS-Flow with *Linear-Transformer-Block*
- ▶ Baselines papers (3 runs and 20 runs)

Data Set	GP-Copula	LSTM-RealNVP	LSTM-MAF	Transformer-MAF	M-N-BEATS-Flow (Transformer)
Electricity	0.024 ± 0.002	0.024 ± 0.001	0.0208 ± 0.000	0.0207 ± 0.000	0.0223 ± 0.0021
Traffic	0.078 ± 0.002	0.078 ± 0.001	0.069 ± 0.002	0.056 ± 0.001	0.0485 ± 0.0049
Solar	0.337 ± 0.024	0.331 ± 0.02	0.315 ± 0.023	0.301 ± 0.014	0.3289 ± 0.0124
Exchange	0.007 ± 0.000	0.0064 ± 0.003	0.005 ± 0.003	0.005 ± 0.003	0.0062 ± 0.0001

Table 4: $CRPS_{sum}$ for baseline methods, M-N-BEATS-Flow with *Linear-Transformer-Block*

Experiments

Forecast Plots

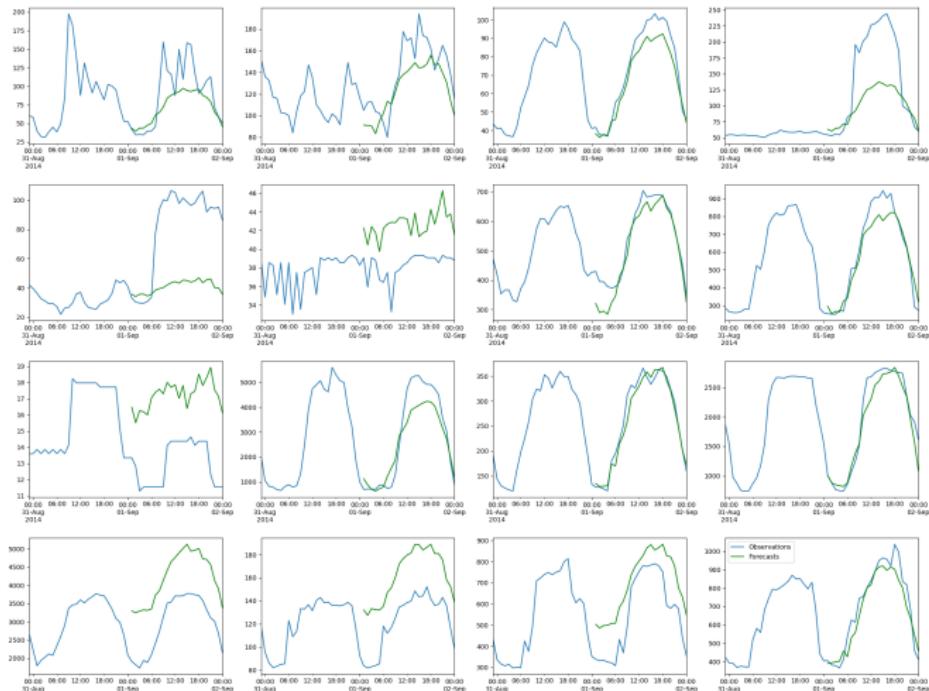


Figure 8: Forecasts produced by M-N-BEATS with Linear-Transformer-Block on the *Electricity* data set, only the forecasts of a single run are plotted.

Experiments

Forecast Plots

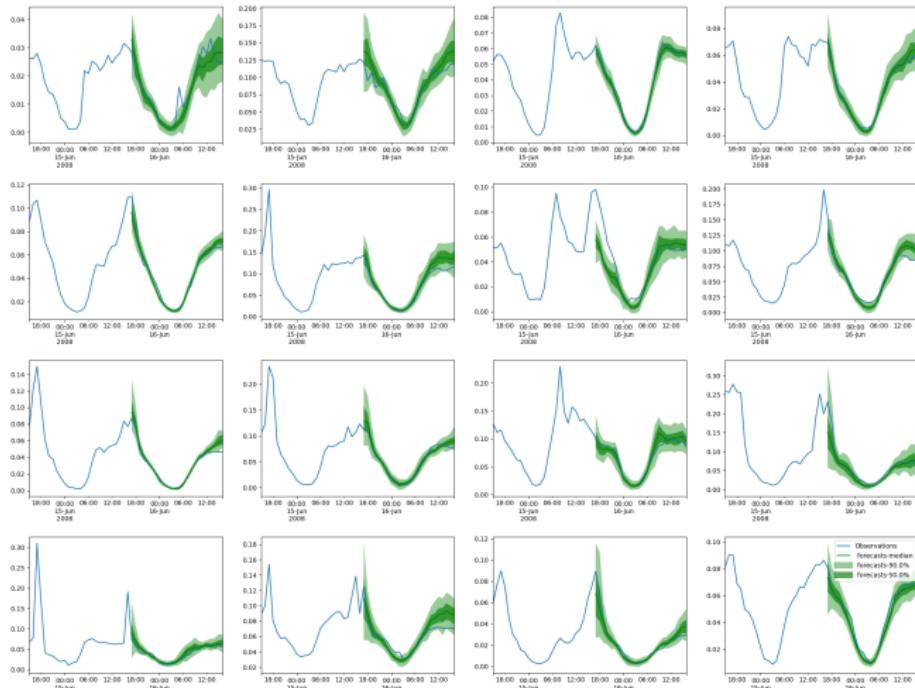


Figure 9: Forecasts produced by M-N-BEATS-Flow with
Linear-Transformer-Block on the *Traffic* data set, only the forecasts of a
single run are plotted.

Experiments

Further Experiments Results

Further Experiments

- ▶ Other block types
- ▶ Same hyperparameters as with *linear-transformer-block*

MSE

Data Set	Simple-Block	Linear-Block	Conv-Block	Linear-Conv-Block
Electricity	253864 ± 181095	162721 ± 13712	339995 ± 253724	135090 ± 10410
Traffic	0.0007 ± 0.0003	0.000435 ± 0.00002	None	0.0005 ± 0.0001
Solar	925 ± 68	746 ± 47	1014 ± 72	815 ± 61
Exchange Rate	0.00018 ± 0.0000	0.00016 ± 0.0000	0.00034 ± 0.0000	0.00019 ± 0.0000

Table 5: *MSE* for M-N-BEATS-Flow with other blocks, mean and standard deviation of 10 runs. Scores in bold beat the baseline models in table 2.

Experiments

Further Experiments Results

CRPS

Data Set	Simple-Block	Linear-Block	Conv-Block	Linear-Conv-Block
Electricity	0.0617 ± 0.0235	0.052 ± 0.0008	0.0617 ± 0.0127	0.0495 ± 0.0006
Traffic	0.1932 ± 0.1072	0.1117 ± 0.0078	None	0.1278 ± 0.0264
Solar	0.3717 ± 0.0129	0.3348 ± 0.0114	0.3934 ± 0.0153	0.3591 ± 0.0168
Exchange Rate	0.0092 ± 0.0002	0.0089 ± 0.0004	0.0108 ± 0.0008	0.0094 ± 0.0005

Table 6: CRPS for M-N-BEATS-Flow with other blocks, mean and standard deviation of 10 runs. Scores in bold beat the baseline models in table 3.

Experiments

Further Experiments Results

$CRPS_{sum}$

Data Set	Simple-Block	Linear-Block	Conv-Block	Linear-Conv-Block
Electricity	0.0321±0.0252	0.0206±0.0014	0.0314±0.0132	0.0203±0.0011
Traffic	0.1351±0.1047	0.0451±0.0102	None	0.0653±0.0285
Solar	0.3154±0.0173	0.2737±0.0143	0.3475±0.0211	0.3028±0.0217
Exchange Rate	0.0062±0.0003	0.0059±0.0003	0.0074±0.0014	0.0064±0.0006

Table 7: $CRPS_{sum}$ for M-N-BEATS-Flow with other blocks, mean and standard deviation of 10 runs. Scores in bold beat the baseline models in table 4.

Experiments

Forecast Plots

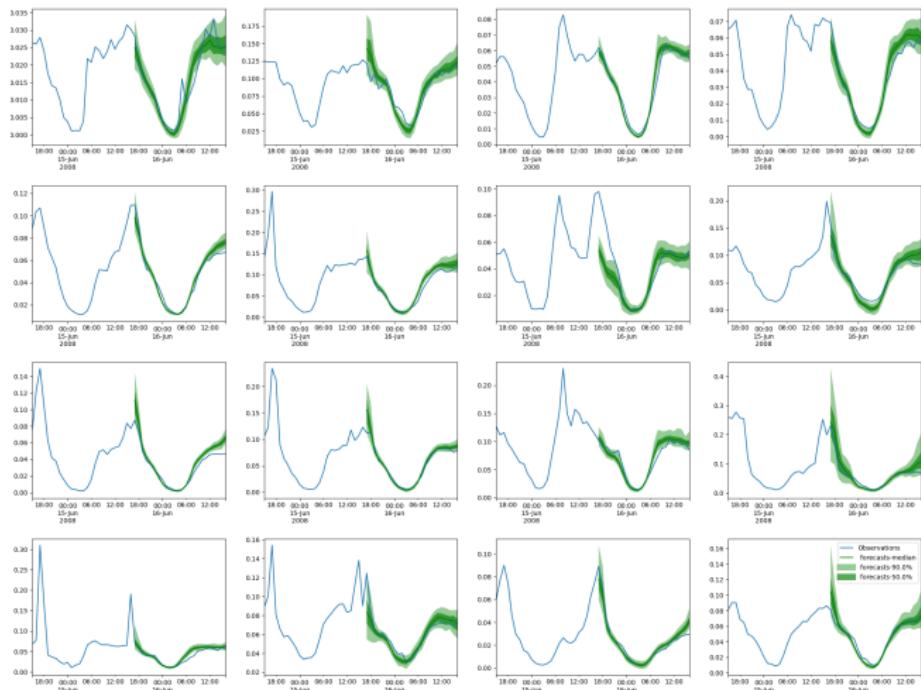


Figure 10: Forecasts produced by M-N-BEATS-Flow with *Linear-Block* on the *Traffic* data set, only the forecasts of a single run are plotted.

Introduction

Related Work

Methodology

Experiments

Conclusion

Conclusion

- ▶ Naive approach and M-N-BEATS not competitive
- ▶ By N-BEATS design due to residual connections model size explodes for high **dimensional data**
- ▶ Combine M-N-BEATS + normalizing flow
- ▶ competitive results
- ▶ Linear-Block performed very well on all data sets and outperformed all baselines
- ▶ This is likely to show the effectiveness of normalizing flows

References I

- [1] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. “Density estimation using Real NVP”. In: *CoRR* abs/1605.08803 (2016). arXiv: 1605.08803. URL: <http://arxiv.org/abs/1605.08803>.
- [2] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: <http://archive.ics.uci.edu/ml>.
- [3] Cheng Guo and Felix Berkhahn. “Entity Embeddings of Categorical Variables”. In: *CoRR* abs/1604.06737 (2016). arXiv: 1604.06737. URL: <http://arxiv.org/abs/1604.06737>.

References II

- [4] Guokun Lai et al. "Modeling Long- and Short-Term Temporal Patterns with Deep Neural Networks". In: *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '18. Ann Arbor, MI, USA: Association for Computing Machinery, 2018, pp. 95–104. ISBN: 9781450356572. DOI: 10.1145/3209978.3210006. URL: <https://doi.org/10.1145/3209978.3210006>.
- [5] Boris N Oreshkin et al. "N-BEATS: Neural basis expansion analysis for interpretable time series forecasting". In: *arXiv preprint arXiv:1905.10437* (2019).
- [6] Kashif Rasul et al. "Multivariate Probabilistic Time Series Forecasting via Conditioned Normalizing Flows". In: *International Conference on Learning Representations*. 2021. URL: <https://openreview.net/forum?id=WiGQBFuVRv>.

References III

- [7] David Salinas et al. “High-Dimensional Multivariate Forecasting with Low-Rank Gaussian Copula Processes”. In: *NeurIPS*. 2019.
- [8] Ashish Vaswani et al. “Attention is all you need”. In: *arXiv preprint arXiv:1706.03762* (2017).

APPENDIX

N-BEATS

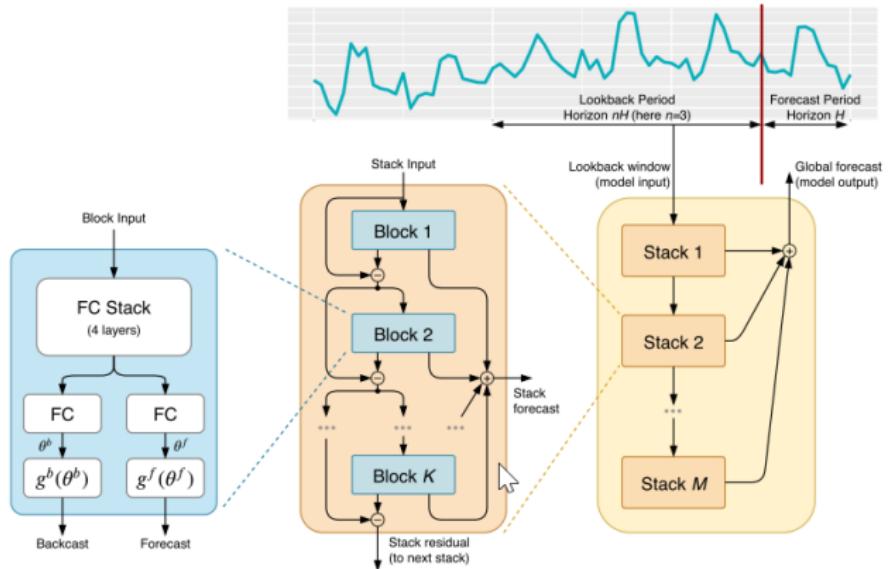


Figure 11: N-BEATS architecture. Illustration of the N-BEATS architecture, index ℓ was dropped for simplicity.
source: [5].

Appendix

N-BEATS

Stack of linear Layers

$$\mathbf{h}_{\ell,1} = \text{ReLU}(\mathbf{W}_{\ell,1}\mathbf{x}_\ell + \mathbf{b}_{\ell,1}) \quad (9)$$

$$\mathbf{h}_{\ell,2} = \text{ReLU}(\mathbf{W}_{\ell,2}\mathbf{h}_{\ell,1} + \mathbf{b}_{\ell,2}) \quad (10)$$

$$\mathbf{h}_{\ell,3} = \text{ReLU}(\mathbf{W}_{\ell,3}\mathbf{h}_{\ell,2} + \mathbf{b}_{\ell,3}) \quad (11)$$

$$\mathbf{h}_{\ell,4} = \text{ReLU}(\mathbf{W}_{\ell,4}\mathbf{h}_{\ell,3} + \mathbf{b}_{\ell,4}) \quad (12)$$

where \mathbf{x}_ℓ is the block input and \mathbf{W}_ℓ and \mathbf{b}_ℓ are weights and bias of the layers.

Appendix

N-BEATS

Basis vectors trend block

Given a time vector $\mathbf{t} = [0, 1, 2, \dots, H-2, H-1]^T/H$, the basis function is given by

$$\hat{\mathbf{y}}_{s,\ell} = \sum_{i=0}^p \theta_{s,\ell,i}^f t^i. \quad (13)$$

or

$$\hat{\mathbf{y}}_{s,\ell}^{tr} = \mathbf{T} \theta_{s,\ell}^f \quad (14)$$

with the basis $\mathbf{T} = [\mathbf{1}, \mathbf{t}, \dots, \mathbf{t}^p]$. The tr superscript in $\hat{\mathbf{y}}_{s,\ell}^{tr}$ indicates that the block returns a forecast of the trend. For the backcast, the time vector is $\mathbf{t} = [0, 1, 2, \dots, S-2, S-1]^T/S$ where S is the sequence length of the input time series. The number of basis parameters is $\dim(\theta^f) = p + 1$ for the forecast and $\dim(\theta^b) = p + 1$ for the backcast.

Appendix

N-BEATS

Basis vectors seasonality block

$$\hat{\mathbf{y}}_{s,\ell} = \sum_{i=0}^{\lfloor H/2-1 \rfloor} \theta_{s,\ell,i}^f \cos(2\pi it) + \theta_{s,\ell,i+\lfloor H/2 \rfloor}^f \sin(2\pi it) \quad (15)$$

which can be written in matrix form as

$$\hat{\mathbf{y}}_{s,\ell}^{\text{seas}} = \mathbf{S} \theta_{s,\ell}^f \quad (16)$$

with

$$\mathbf{S} = [\mathbf{1}, \cos(2\pi \mathbf{t}), \dots, \cos(2\pi \lfloor H/2-1 \rfloor \mathbf{t}), \sin(2\pi \mathbf{t}), \dots, \sin(2\pi \lfloor H/2-1 \rfloor \mathbf{t})] \quad (17)$$

The number of basis parameters is $\dim(\theta^f) = H$ for the forecast branch and $\dim(\theta^b) = H$ for the backcast branch.

Methodology

M-N-BEATS Blocks

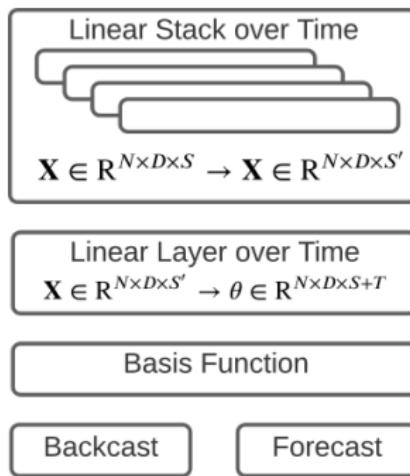


Figure 12: Univariate N-BEATS block. Vanilla N-BEATS block when applied to multivariate time series data applies the model identically to each time series.

Methodology

Transformer encoder layer

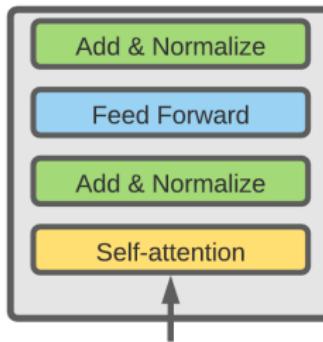


Figure 13: Transformer encoder layer. Source: [8].

Experiments

Metrics

- ▶ Mean Squared Error (MSE)
- ▶ Continuous Ranked Probability Score ($CRPS$)
- ▶ $CRPS_{sum}$

- ▶ MSE for M-N-BEATS
- ▶ MSE , $CRPS$, $CRPS_{sum}$ for M-N-BEATS-Flow

$$MSE(\mathbf{Y}, \hat{\mathbf{Y}}) = \left\| \mathbf{Y} - \hat{\mathbf{Y}} \right\|_F^2 \quad (18)$$

where $\|\cdot\|_F$ is the Frobenius norm.

Experiments

Metrics

CRPS

- ▶ CRPS score is computed for each time series
- ▶ Measures how good the approximated distribution compares to the data using the CDF F
- ▶ CDF for the predictive target distribution $F(z) = \mathbf{P}[X \leq z]$
- ▶ Measures the probability that the normalizing flow samples a x smaller than z

Experiments

Metrics

CRPS for a given fixed data point at time step t is given as

$$\text{CRPS}(F, x) = \int_{\mathbb{R}} (F(z) - \mathbb{I}\{x \leq z\})^2 dz \quad (19)$$

- ▶ x is a single data point (one time step of one channel)
- ▶ $\mathbb{I}\{x \leq z\})^2$ is 1 if $x \leq z$ and 0 otherwise
- ▶ Final score is the mean over the *CRPS* score of all data points and channels

Experiments

Metrics

- ▶ CDF $F(z) = \mathbf{P}[X \leq z]$ for a certain value of z can be approximated using the empirical CDF

$$\hat{F}(z) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{X_i \leq z\} \quad (20)$$

- ▶ n is the number of samples generated from the model using $X_i \sim F$
- ▶ In [6] 100 samples were used to compute the empirical CDF
- ▶ Lower values of $CRPS$ are better

Experiments

Metrics

$CRPS_{sum}$ for a single time step t is given by

$$CRPS_{sum} = \mathbb{E}_t \left[CRPS \left(\hat{F}_{sum}(t), \sum_{i=0}^D x_t^i \right) \right] \quad (21)$$

- ▶ $\hat{F}_{sum}(t)$ is computed by summing up the individual \hat{F} over all D time series
- ▶ The final score is obtained by taking the mean of the $CRPS_{sum}$ of all time steps

Experiments

Data Sets

- ▶ Exchange rate: daily exchange rate between 8 currencies as used in [4]
- ▶ Solar: hourly photo-voltaic production of 137 stations in Alabama State used in [4]
- ▶ Electricity: hourly time series of the electricity consumption of 370 customers [2]
- ▶ Traffic: hourly occupancy rate, between 0 and 1, of 963 San Francisco car lanes [2]

Source: [7]

Experiments

Data Sets

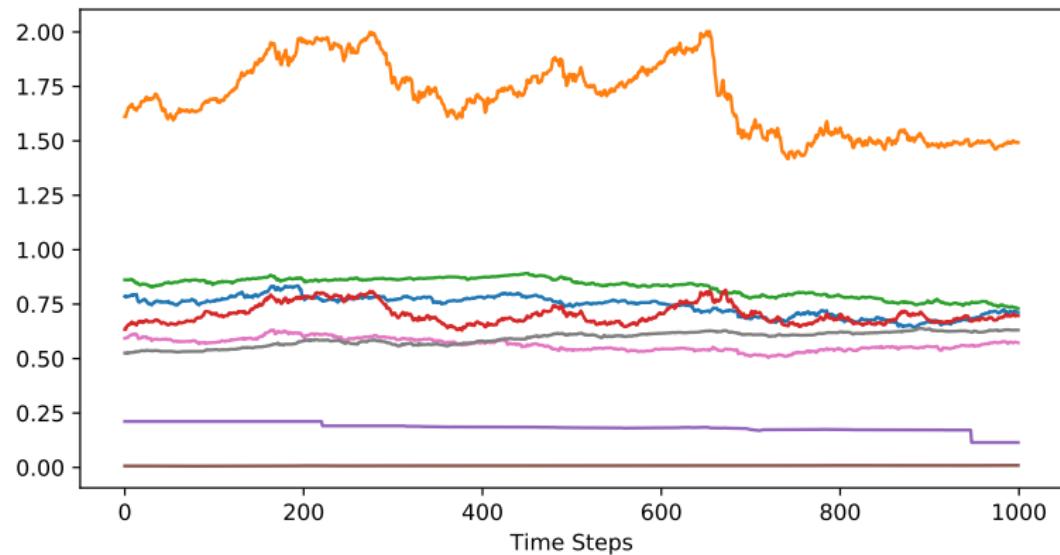


Figure 14: Exchange data set. The first 1000 time steps of the all time series in the *Exchange Rate* data set.

Experiments

Data Sets

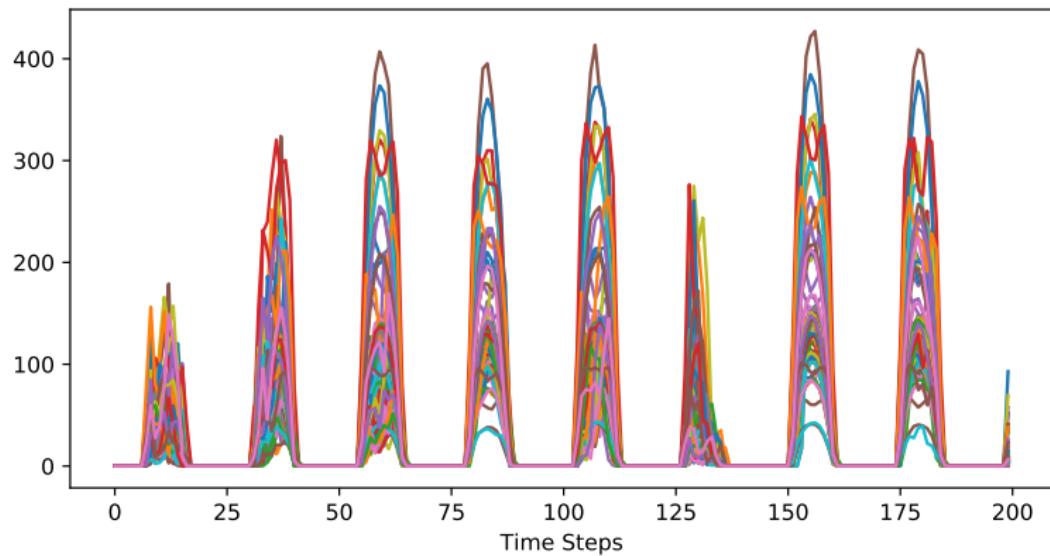


Figure 15: Solar data set. The first 200 time steps of the all time series in the Solar data set.

Experiments

Data Sets

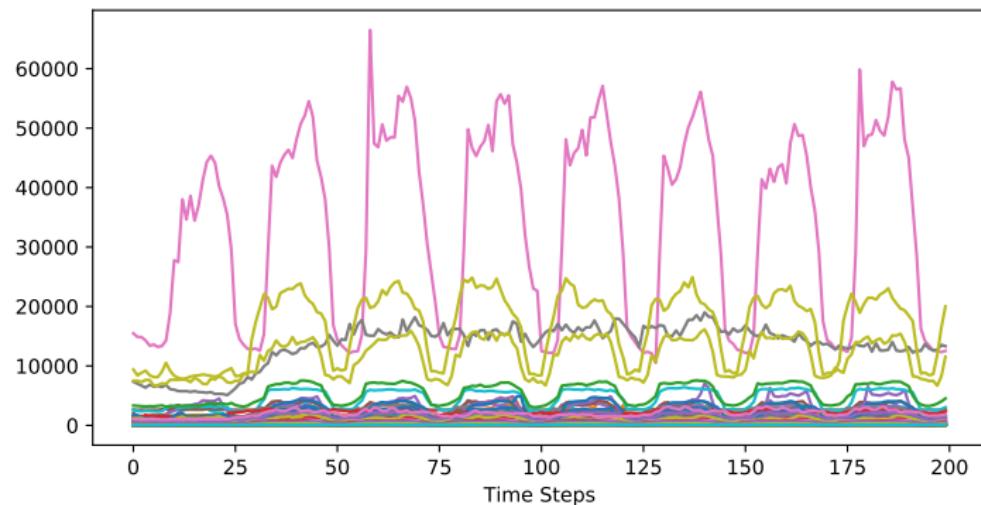


Figure 16: Electricity data set. The first 200 time steps of the all time series in the *Electricity* data set.

Experiments

Data Sets

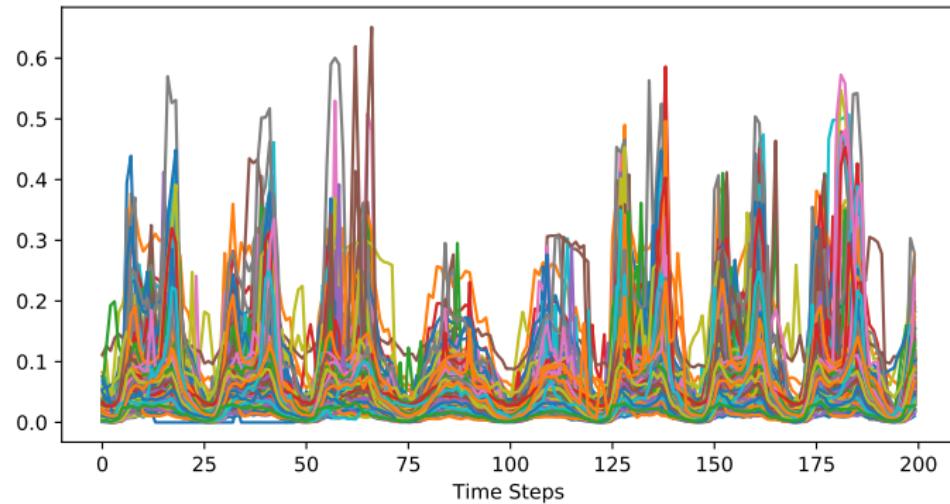


Figure 17: Traffic data set. The first 200 time steps of the all time series in the *Traffic* data set.

Experiments

Data Pipeline

Static features

- ▶ Channel index is used as categorical variable and an entity embedding like proposed in [3] is applied

Time dependent features

- ▶ Sinus cosinus from time features like **day of week** or **hour of day**

Lagged time series

- ▶ Lags are applied according to the data set's frequency
- ▶ 1, 24 and 168 for hourly time series
- ▶ 1, 2 For daily time series (*Exchange*, only business days)

Experiments

Data Pipeline

Scaling

- ▶ Magnitudes of time series differ strongly
- ▶ Divide each time series by its training window mean
- ▶ Target divided by the same factor
- ▶ After training for data generation, the output distribution is multiplied by the same factor.

Experiments

Data Pipeline

Example time dependent features

- ▶ Time features for *hour* and *dayofweek*
- ▶ 0 to 23, 0 to 6
- ▶ 2 features are computed for a single time step

The time features for a single time step are computed by

$$\sin(\text{dayofweek} \times 2.0 \times \pi/S), \quad \cos(\text{dayofweek} \times 2.0 \times \pi/S) \quad (22)$$

$$\sin(\text{hour} \times 2.0 \times \pi/S), \quad \cos(\text{hour} \times 2.0 \times \pi/S) \quad (23)$$

Experiments

Data Pipeline

- ▶ For hourly data, *hour* and *dayofweek* time features are used (*Traffic*, *Solar* and *Electricity*)
- ▶ For daily data, *dayofweek* and *dayofyear* (*Exchange*)
- ▶ Fourier time features are $(N \times S \times 4)$
- ▶ time dependent features since they vary over time.

Experiments

N-BEATS-Naive Results

Data Set	M-N-BEATS-Naive
<i>Electricity</i>	5005447 ± 2500474
<i>Traffic</i>	0.00088 ± 0.0000
<i>Solar</i>	8842 ± 9854
<i>Exchange Rate</i>	0.00037 ± 0.0000

Table 8: *MSE* for N-BEATS-Naive, mean and standard deviation of 10 runs are reported.

Experiments

Hyperparameter Search

Max Learning Rate	0.001
Batches per Epoch	100
Epochs	25
Batch Size	64
p (Dropout)	0.5
Weight Decay	0.01

Table 9: Fixed hyperparameters for M-N-BEATS and M-N-BEATS-Flow.

Experiments

Hyperparameter Search Results M-N-BEATS

Learning Rate	Block Type	N Blocks	N Heads	Embedding Size	MSE
0.000010	<i>Linear-Transformer-Block</i>	10	4	32	794697
0.000010	<i>Linear-Attention-Block</i>	10	8	32	940344
0.000010	<i>Linear-Attention-Block</i>	10	8	128	944887
0.000010	<i>Linear-Transformer-Block</i>	10	8	32	963443
0.000010	<i>Linear-Transformer-Block</i>	3	4	128	974484

Table 10: Grid search best 5 combinations for M-N-BEATS on *Electricity* data set. *N Heads* is number of attention heads, *Embedding Size* is the attention embedding size and *N Blocks* is the number of blocks.

Experiments

Hyperparameter Search Results M-N-BEATS

Learning Rate	Block Type	N Blocks	N Heads	Embedding Size	MSE
0.000100	<i>Linear-Transformer-Block</i>	3	8	64	0.000837
0.000100	<i>Linear-Transformer-Block</i>	5	8	64	0.001060
0.000010	<i>Linear-Attention-Block</i>	3	4	128	0.001088
0.000010	<i>Linear-Transformer-Block</i>	10	8	128	0.001097
0.000010	<i>Linear-Transformer-Block</i>	3	8	128	0.001112

Table 11: Grid search best 5 combinations for M-N-BEATS on *Traffic* data set. *N Heads* is number of attention heads, *Embedding Size* is the attention embedding size and *N Blocks* is the number of blocks.

Experiments

Hyperparameter Search Results M-N-BEATS

Learning Rate	Block Type	N Blocks	N Heads	Embedding Size	MSE
0.000100	<i>Linear-Attention-Block</i>	5	4	128	6051.420232
0.000010	<i>Linear-Attention-Block</i>	10	4	128	6778.967792
0.000010	<i>Linear-Transformer-Block</i>	3	8	64	6788.116120
0.000100	<i>Linear-Transformer-Block</i>	3	8	32	6788.814785
0.000010	<i>Linear-Attention-Block</i>	10	8	128	6796.519731

Table 12: Grid search best 5 combinations for M-N-BEATS on *solar* data set. *N Heads* is number of attention heads, *Embedding Size* is the attention embedding size and *N Blocks* is the number of blocks.

Experiments

Hyperparameter Search Results M-N-BEATS

Learning Rate	Block Type	N Blocks	N Heads	Embedding Size	MSE
0.000010	<i>Linear-Attention-Block</i>	5	8	32	0.001
0.000100	<i>Linear-Attention-Block</i>	5	4	64	0.0011
0.000100	<i>Linear-Attention-Block</i>	3	8	32	0.0011
0.000010	<i>Linear-Attention-Block</i>	10	4	128	0.0011
0.000100	<i>Linear-Attention-Block</i>	3	4	64	0.0012

Table 13: Grid search best 5 combinations for M-N-BEATS on *Exchange Rate* data set. *N Heads* is number of attention heads, *Embedding Size* is the attention embedding size and *N Blocks* is the number of blocks.

Experiments

Hyperparameter Search Results M-N-BEATS-Flow

Learning Rate	Block Type	N Blocks	N Heads	Embedding Size	Flow Type	MSE
0.000010	Linear-Transformer-Block	3	8	128	RealNVP	295682
0.000010	Linear-Transformer-Block	10	4	128	MAF	307175
0.000010	Linear-Transformer-Block	5	4	32	RealNVP	312071
0.000100	Linear-Transformer-Block	10	4	64	MAF	326673
0.000010	Linear-Transformer-Block	3	8	32	RealNVP	340831

Table 14: Grid search best 5 combinations for M-N-BEATS-Flow on *Electricity* data set. *N Heads* is number of attention heads, *Embedding Size* is the attention embedding size and *N Blocks* is the number of blocks.

Experiments

Hyperparameter Search Results M-N-BEATS-Flow

Learning Rate	Block Type	N Blocks	N Heads	Embedding Size	Flow Type	MSE
0.000010	<i>Linear-Attention-Block</i>	10	8	128	RealNVP	0.00066
0.000010	<i>Linear-Transformer-Block</i>	5	4	128	MAF	0.00068
0.000010	<i>Linear-Attention-Block</i>	10	4	64	RealNVP	0.00069
0.000100	<i>Linear-Transformer-Block</i>	3	8	128	MAF	0.00076
0.000100	<i>Linear-Attention-Block</i>	10	8	64	MAF	0.00077

Table 15: Grid search best 5 combinations for M-N-BEATS-Flow on *Traffic* data set. *N Heads* is number of attention heads, *Embedding Size* is the attention embedding size and *N Blocks* is the number of blocks.

Experiments

Hyperparameter Search Results M-N-BEATS-Flow

Learning Rate	Block Type	N Blocks	N Heads	Embedding Size	Flow Type	MSE
0.000010	<i>Linear-Attention-Block</i>	5	4	32	RealNVP	309
0.000100	<i>Linear-Attention-Block</i>	3	4	64	RealNVP	319
0.000010	<i>Linear-Attention-Block</i>	10	8	32	RealNVP	320
0.000100	<i>Linear-Transformer-Block</i>	10	4	32	RealNVP	327
0.000100	<i>Linear-Transformer-Block</i>	5	8	128	MAF	327

Table 16: Grid search best 5 combinations for M-N-BEATS-Flow on *Solar* data set. *N Heads* is number of attention heads, *Embedding Size* is the attention embedding size and *N Blocks* is the number of blocks.

Experiments

Hyperparameter Search Results M-N-BEATS-Flow

Learning Rate	Block Type	N Blocks	N Heads	Embedding Size	Flow Type	MSE
0.000100	Linear-Transformer-Block	10	4	64	RealNVP	0.00036
0.000010	Linear-Transformer-Block	10	4	32	RealNVP	0.00047
0.000010	Linear-Transformer-Block	3	4	64	MAF	0.00050
0.000100	Linear-Attention-Block	10	8	128	RealNVP	0.00052
0.000010	Linear-Attention-Block	3	8	32	RealNVP	0.00052

Table 17: Grid search best 5 combinations for M-N-BEATS-Flow on *Exchange Rate* data set. *N Heads* is number of attention heads, *Embedding Size* is the attention embedding size and *N Blocks* is the number of blocks.

Experiments

Main Results M-N-BEATS

M-N-BEATS MSE

Data Set	Linear-Attention-Block	Linear-Transformer-Block
<i>Electricity</i>	1865557 ± 2686855	1110571 ± 797587
<i>Traffic</i>	0.0019 ± 0.0000	0.0017 ± 0.0000
<i>Solar</i>	25157 ± 55542	4384 ± 417
<i>Exchange Rate</i>	0.00037 ± 0.00001	0.00030 ± 0.00006

Table 18: *MSE* for M-N-BEATS with *Linear-Transformer-Block* and *Linear-Attention-Block*, mean and standard deviation of 10 runs. Scores in bold beat the baseline models in table 2.

Experiments

Main Results M-N-BEATS-Flow

M-N-BEATS-Flow MSE

Data Set	Linear-Attention-Block	Linear-Transformer-Block
<i>Electricity</i>	6085982 ± 3535539	155810 ± 26344
<i>Traffic</i>	0.00089 ± 0.0008	0.00045 ± 0.0000
<i>Solar</i>	1329 ± 803	925 ± 41
<i>Exchange Rate</i>	0.000174 ± 0.0000	0.000174 ± 0.0000

Table 19: *MSE* for M-N-BEATS-Flow with *Linear-Transformer-Block* and *Linear-Attention-Block*, mean and standard deviation of 10 runs. Scores in bold beat the baseline models in table 2.

Experiments

Main Results M-N-BEATS-Flow

M-N-BEATS-Flow CRPS

Data Set	Linear-Attention-Block	Linear-Transformer-Block
<i>Electricity</i>	0.2247 ± 0.1013	0.0511 ± 0.0019
<i>Traffic</i>	0.2202 ± 0.1767	0.1148 ± 0.0039
<i>Solar</i>	0.4686 ± 0.1780	0.3829 ± 0.0104
<i>Exchange Rate</i>	0.0092 ± 0.0002	0.0092 ± 0.0001

Table 20: *CRPS* for M-N-BEATS-Flow with *Linear-Transformer-Block* and *Linear-Attention-Block*, mean and standard deviation of 10 runs. Scores in bold beat the baseline models in table 3.

Experiments

Main Results M-N-BEATS-Flow

M-N-BEATS-Flow $CRPS_{sum}$

Data Set	Linear-Attention-Block	Linear-Transformer-Block
Electricity	0.2172 ± 0.1129	0.0223 ± 0.0021
Traffic	0.1776 ± 0.2012	0.0485 ± 0.0049
Solar	0.4183 ± 0.1948	0.3289 ± 0.0124
Exchange Rate	0.0061 ± 0.0001	0.0062 ± 0.0001

Table 21: $CRPS_{sum}$ for M-N-BEATS-Flow with *Linear-Transformer-Block* and *Linear-Attention-Block*, mean and standard deviation of 10 runs.
Scores in bold beat the baseline models in table 4.

Experiments

Forecast Plots

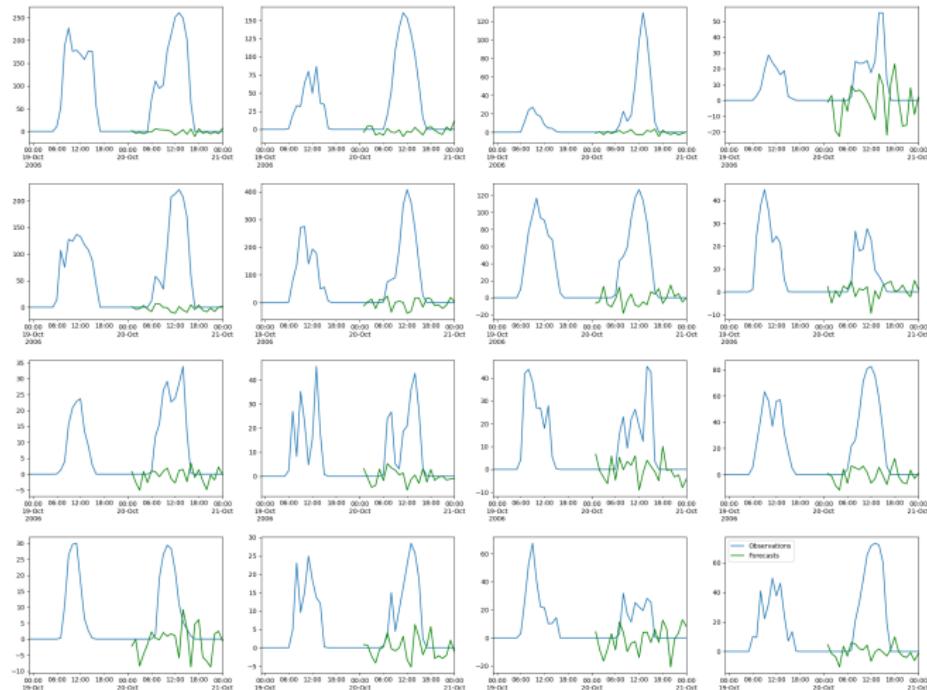


Figure 18: Forecasts produced by M-N-BEATS with
Linear-Transformer-Block on the *Solar* data set, only the forecasts of a
single run are plotted.

Experiments

Forecast Plots

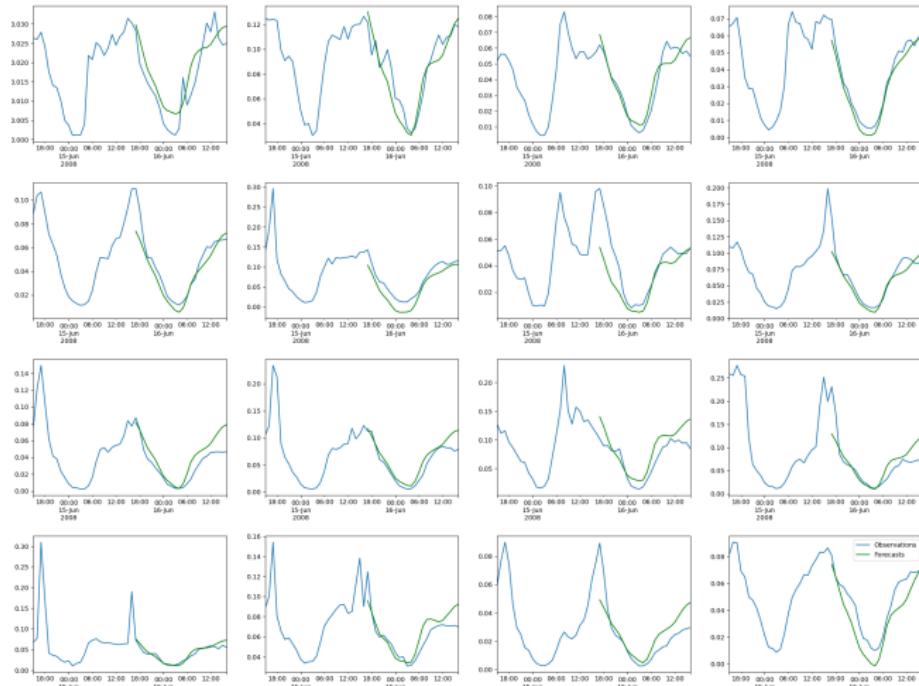


Figure 19: Forecasts produced by M-N-BEATS with *Linear-Transformer-Block* on the *Traffic* data set, only the forecasts of a single run are plotted.

Experiments

Forecast Plots

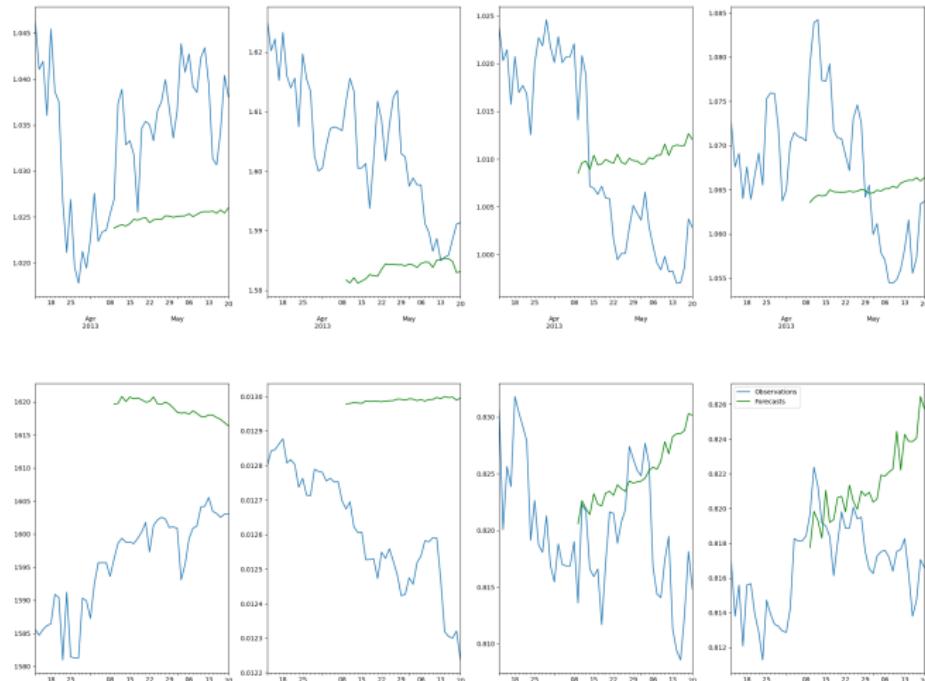


Figure 20: Forecasts produced by M-N-BEATS with *Linear-Transformer-Block* on the *Exchange Rate* data set, only the forecasts of a single run are plotted.

Experiments

Forecast Plots

Experiments

Forecast Plots

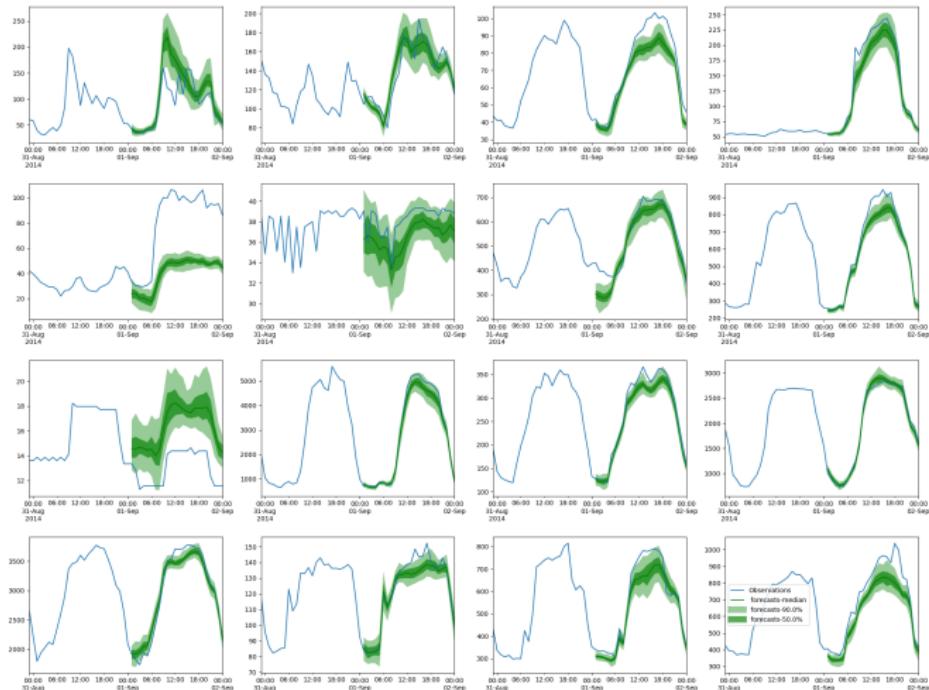


Figure 21: Forecasts produced by M-N-BEATS-Flow with Linear-Transformer-Block on the *Electricity* data set, only the forecasts of a single run are plotted.

Experiments

Forecast Plots

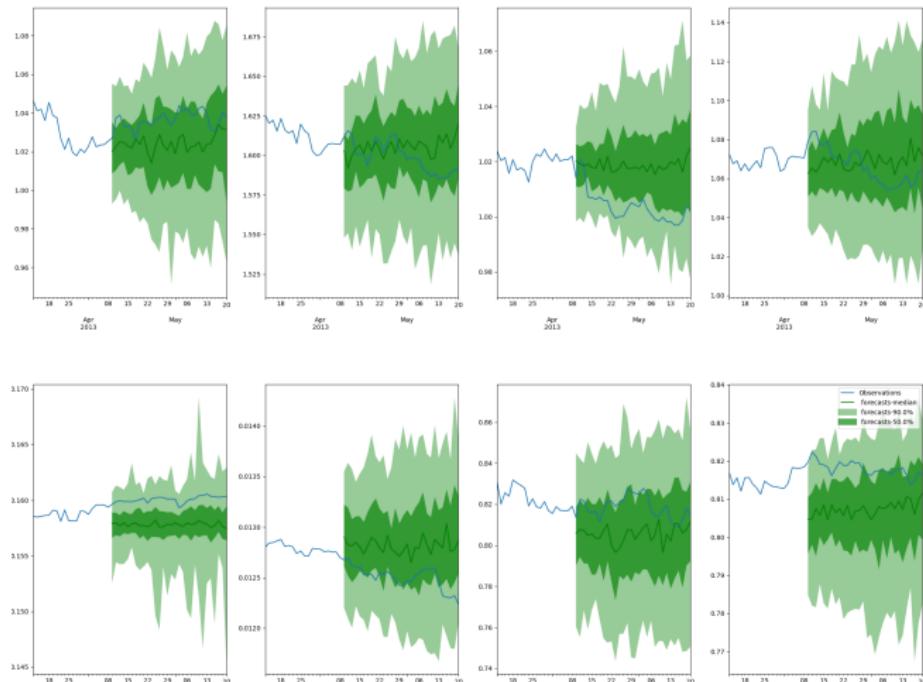


Figure 22: Forecasts produced by M-N-BEATS-Flow with *Linear-Transformer-Block* on the *Exchange Rate* data set, only the forecasts of a single run are plotted.

Experiments

Forecast Plots

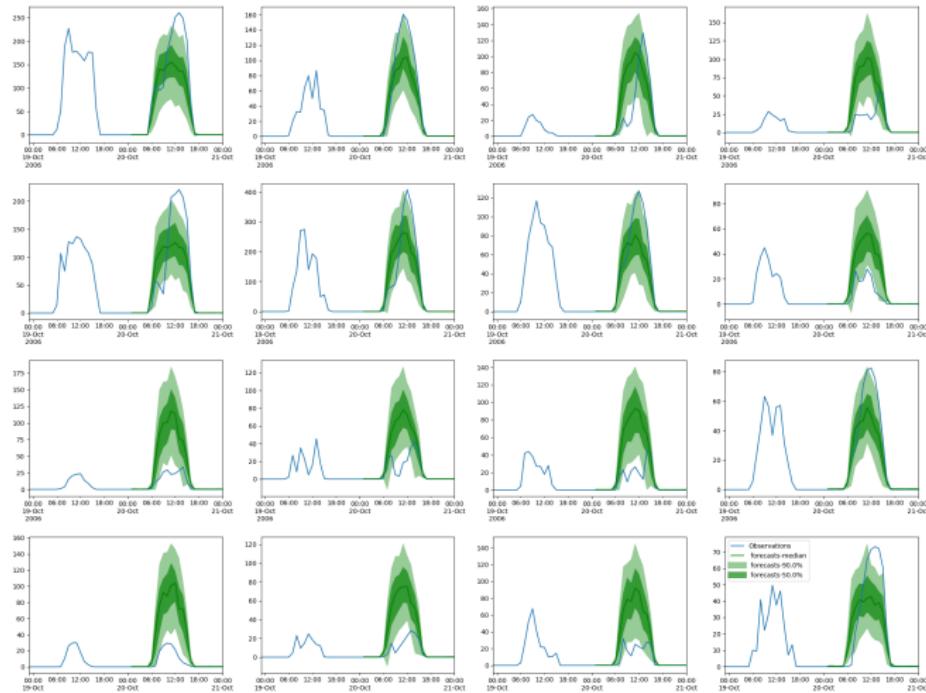


Figure 23: Forecasts produced by M-N-BEATS-Flow with Linear-Transformer-Block on the *Solar* data set, only the forecasts of a single run are plotted.

Experiments

Forecast Plots

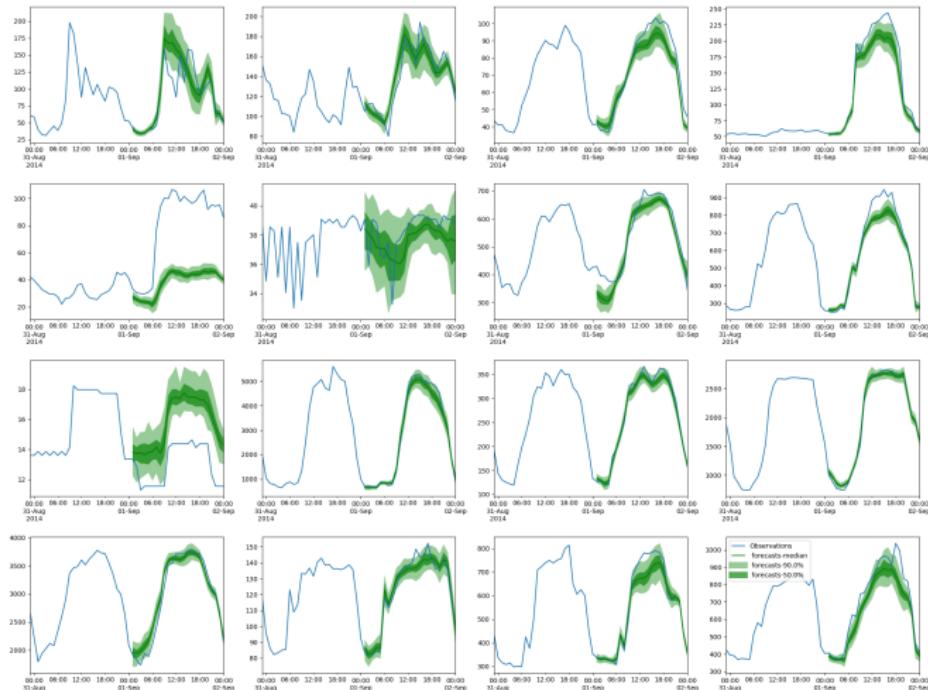


Figure 24: Forecasts produced by M-N-BEATS-Flow with *Linear-Block* on the *Electricity* data set, only the forecasts of a single run are plotted.

Experiments

Forecast Plots

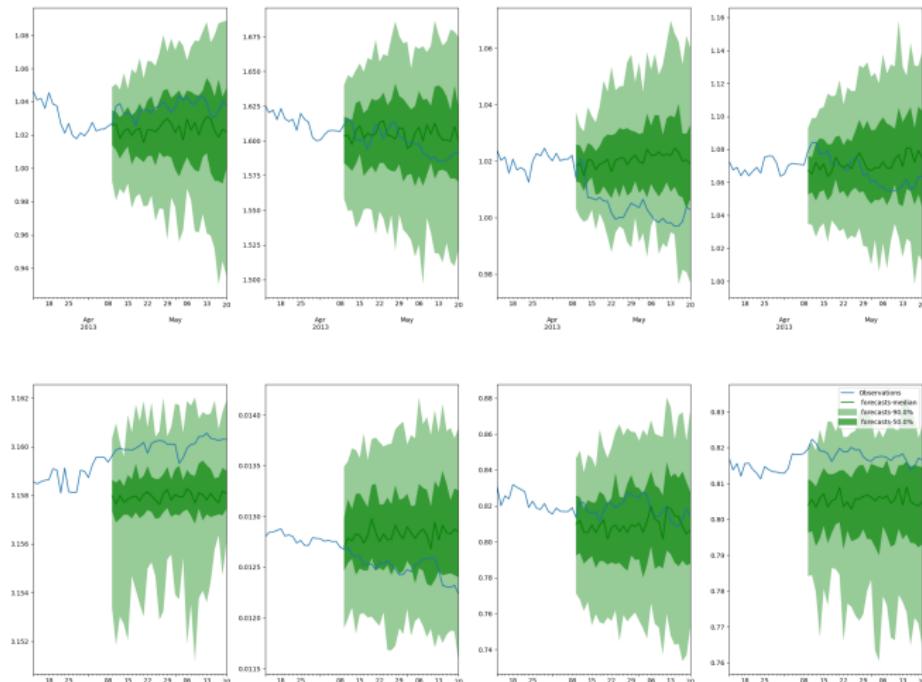


Figure 25: Forecasts produced by M-N-BEATS-Flow with *Linear-Block* on the *Exchange Rate* data set, only the forecasts of a single run are plotted.

Experiments

Forecast Plots

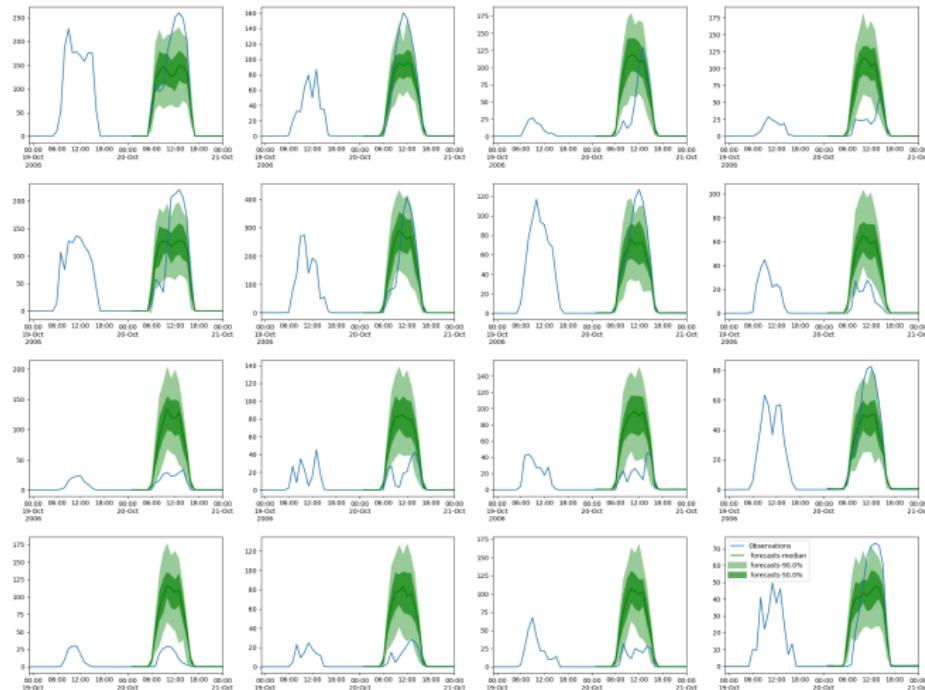


Figure 26: Forecasts produced by M-N-BEATS-Flow with *Linear-Block* on the *Solar* data set, only the forecasts of a single run are plotted.

Experiments

Forecast Plots

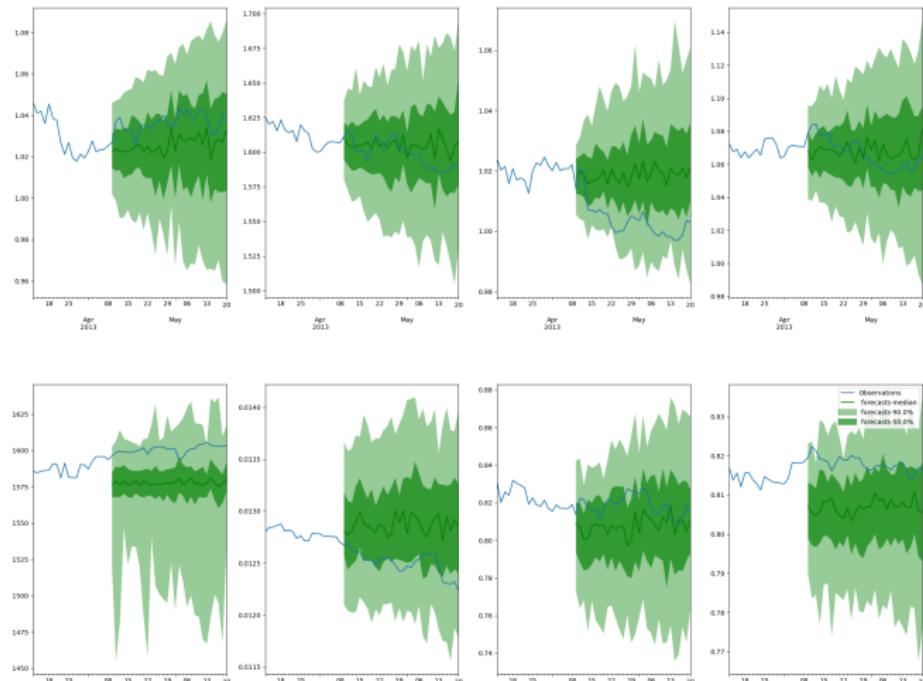


Figure 27: Forecasts produced by M-N-BEATS-Flow with *Linear-Conv-Block* on the *Exchange Rate* data set, only the forecasts of a single run are plotted.

Experiments

Forecast Plots

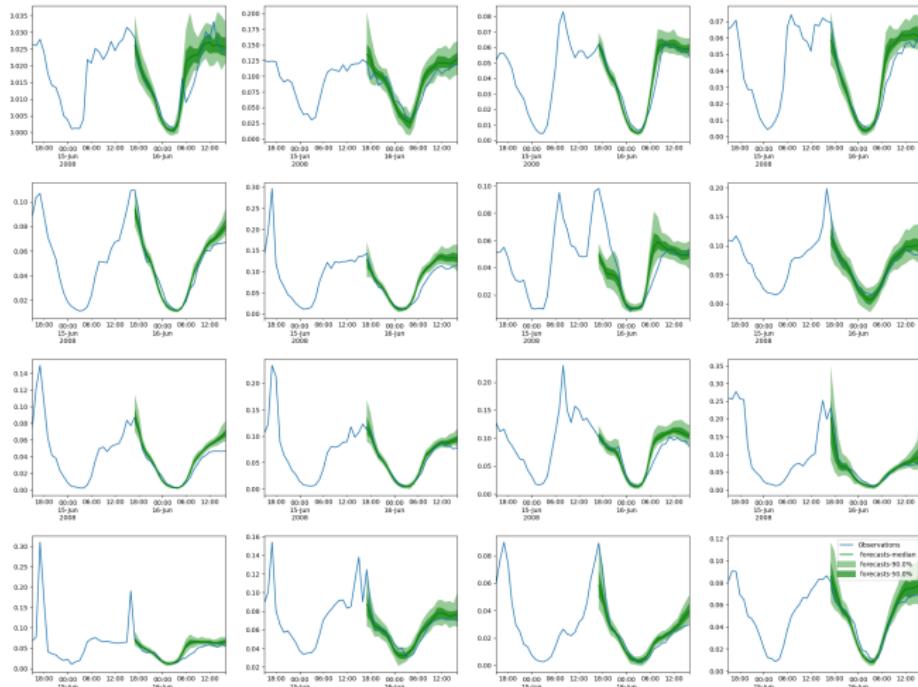


Figure 28: Forecasts produced by M-N-BEATS-Flow with Linear-Conv-Block on the *Traffic* data set, only the forecasts of a single run are plotted.

Experiments

Forecast Plots

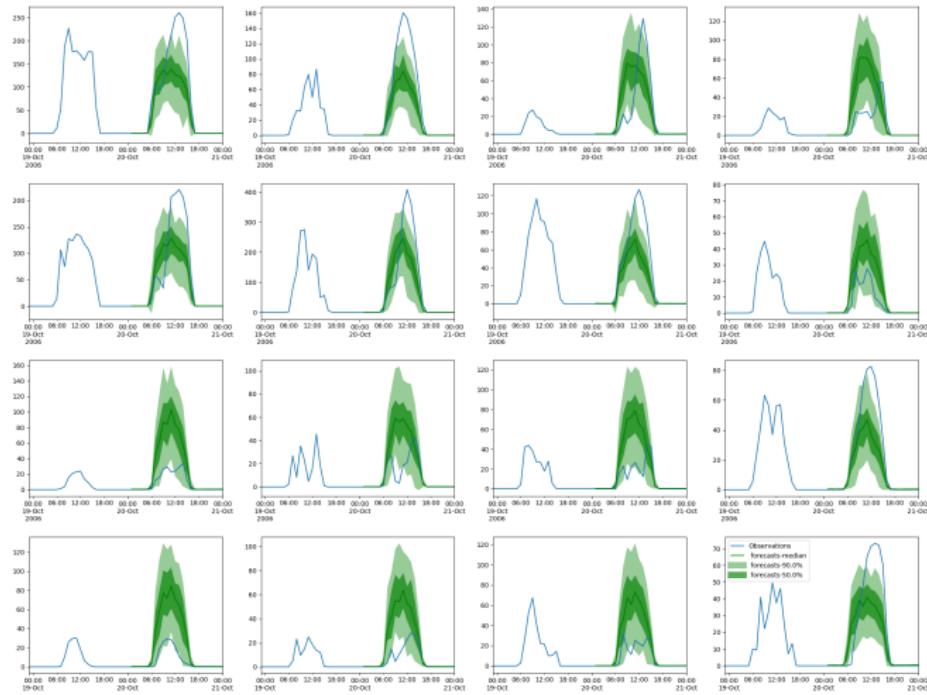


Figure 29: Forecasts produced by M-N-BEATS-Flow with Linear-Conv-Block on the Solar data set, only the forecasts of a single run are plotted.

Experiments

Forecast Plots

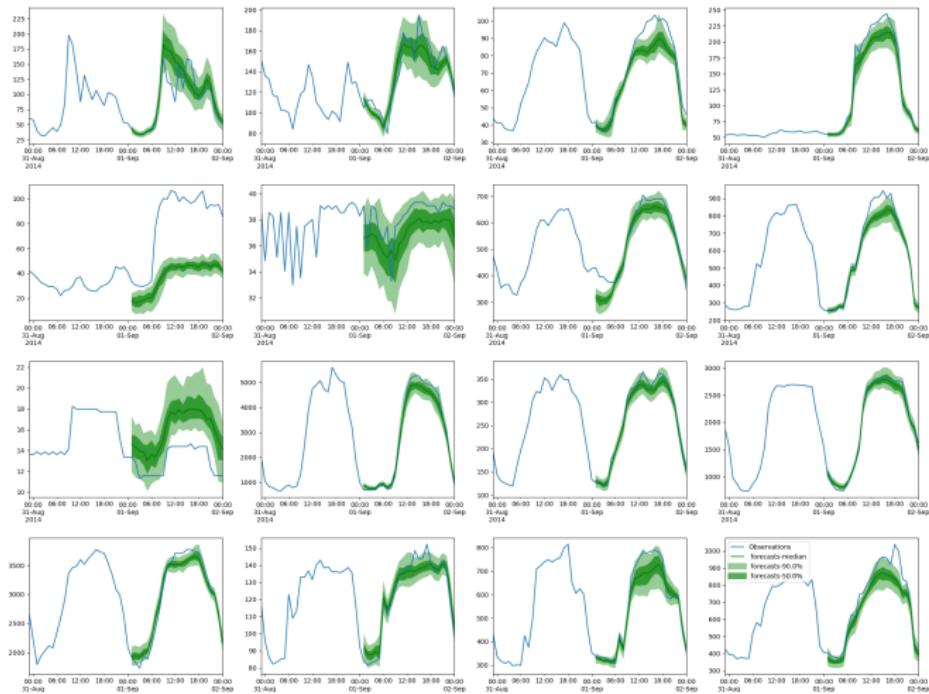


Figure 30: Forecasts produced by M-N-BEATS-Flow with Linear-Conv-Block on the *Electricity* data set, only the forecasts of a single run are plotted.

Experiments

Forecast Plots

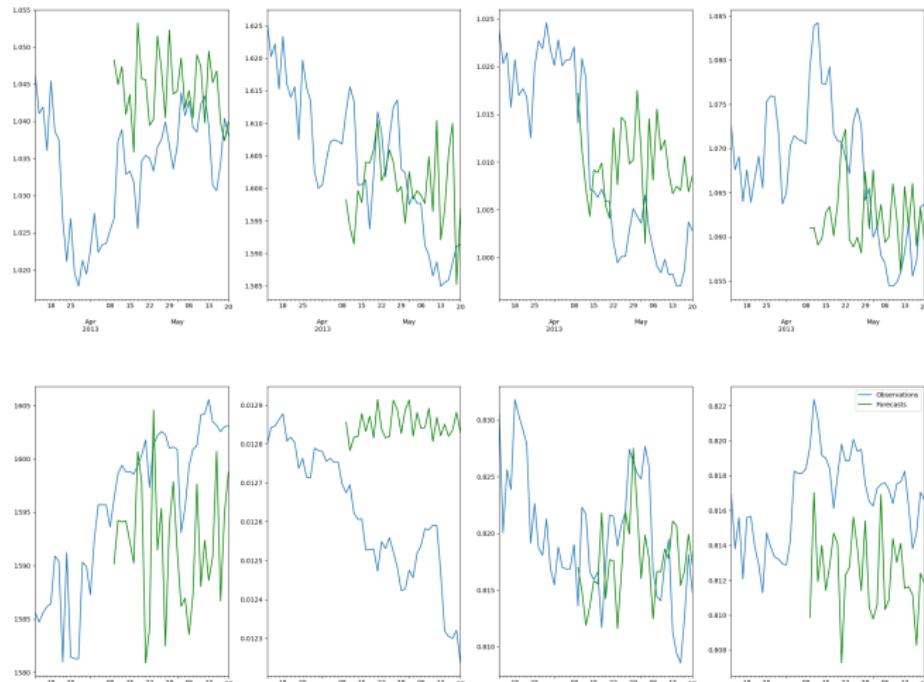


Figure 31: Forecasts produced by N-BEATS-Naive on the *Exchange Rate* data set, only the forecasts of a single run are plotted.

Experiments

Forecast Plots

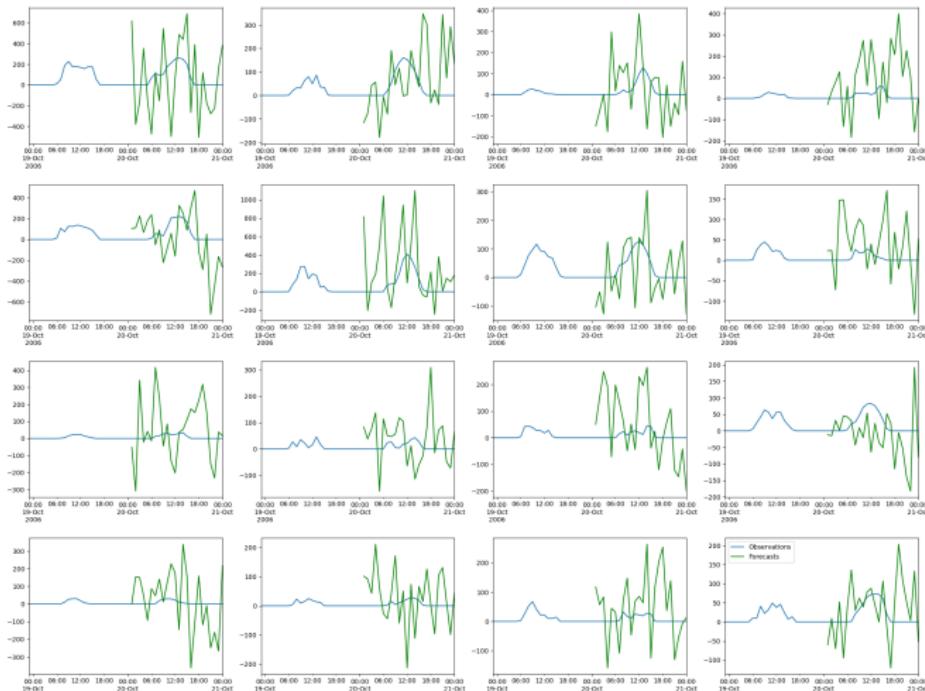


Figure 32: Forecasts produced by N-BEATS-Naive on the *Solar* data set, only the forecasts of a single run are plotted.

Experiments

N-BEATS-Naive

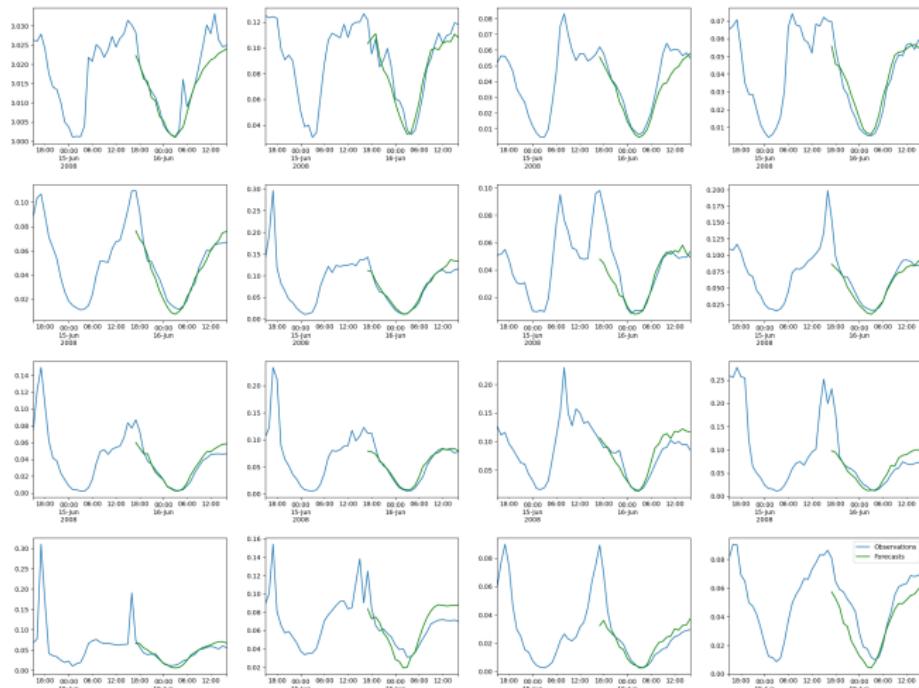


Figure 33: Forecasts produced by N-BEATS-Naive on the *Traffic* data set, only the forecasts of a single run are plotted.

Experiments

Learning Curves

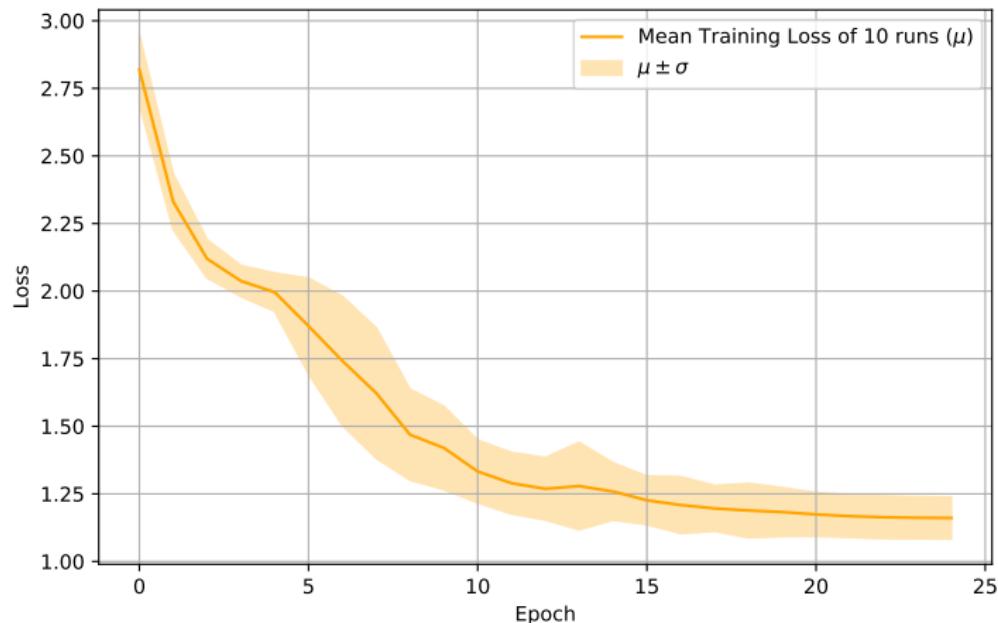


Figure 34: Learning curves for M-N-BEATS on *Electricity* data set. Mean plus minus one standard deviation is plotted.

Experiments

Learning Curves

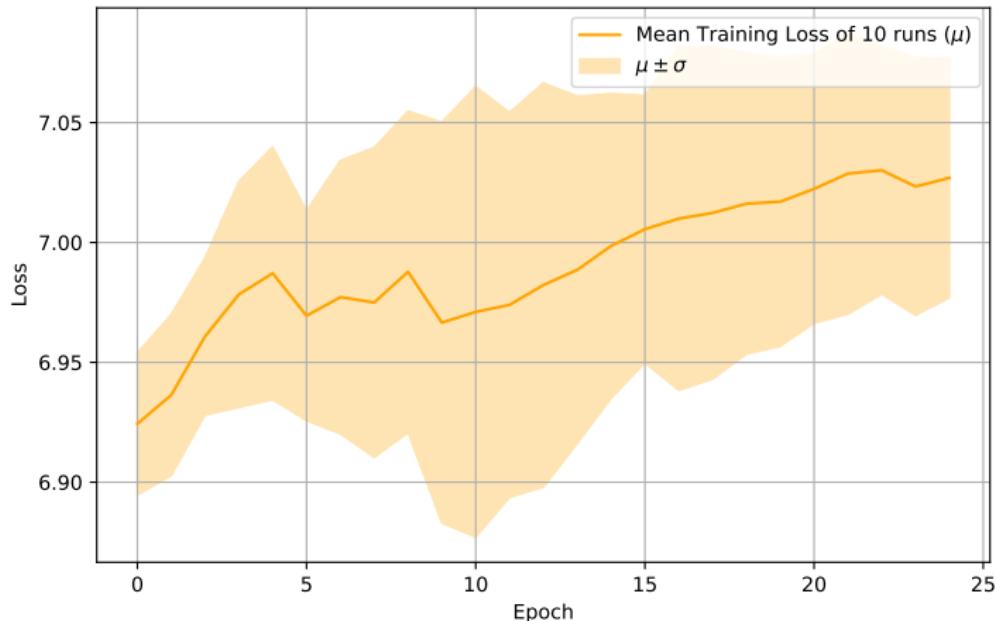


Figure 35: Learning curves for M-N-BEATS on *Solar* data set. Mean plus minus one standard deviation is plotted.

Experiments

Learning Curves

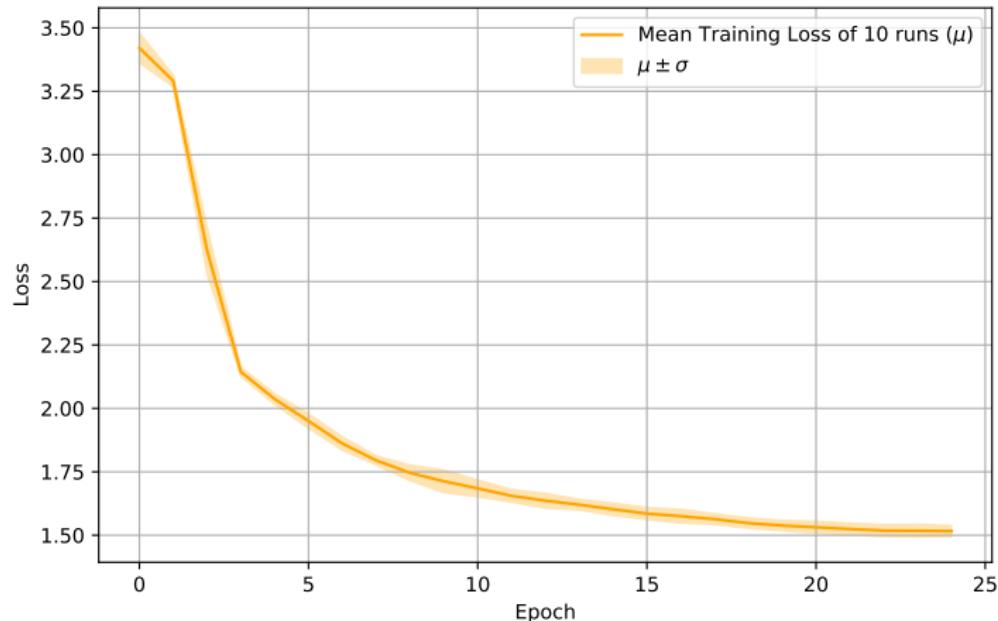


Figure 36: Learning curves for M-N-BEATS on *Traffic* data set. Mean plus minus one standard deviation is plotted.

Experiments

Learning Curves

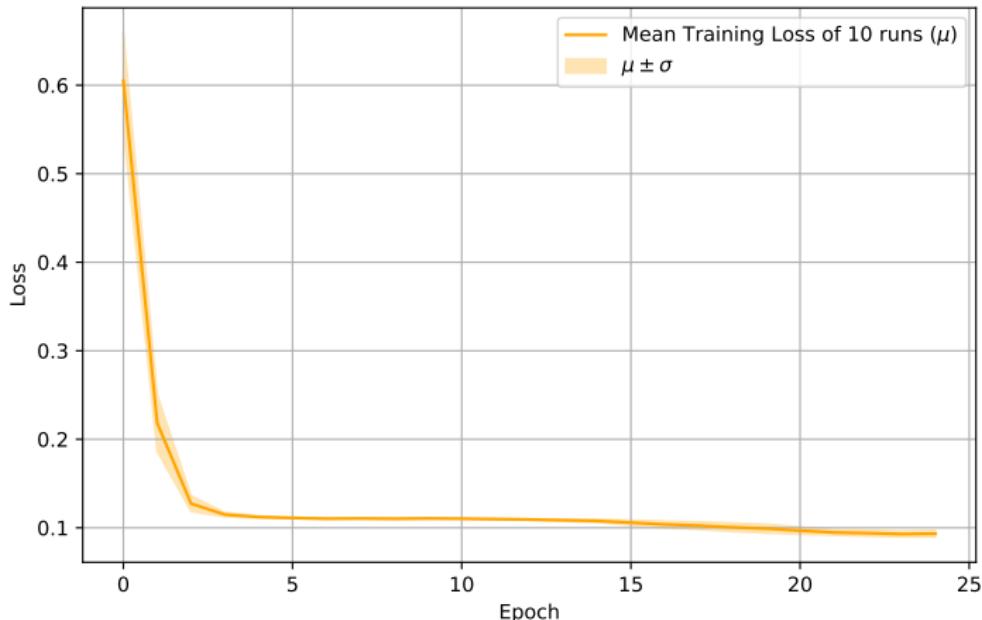


Figure 37: Learning curves for M-N-BEATS on *Exchange Rate* data set.
Mean plus minus one standard deviation is plotted.

Experiments

Learning Curves

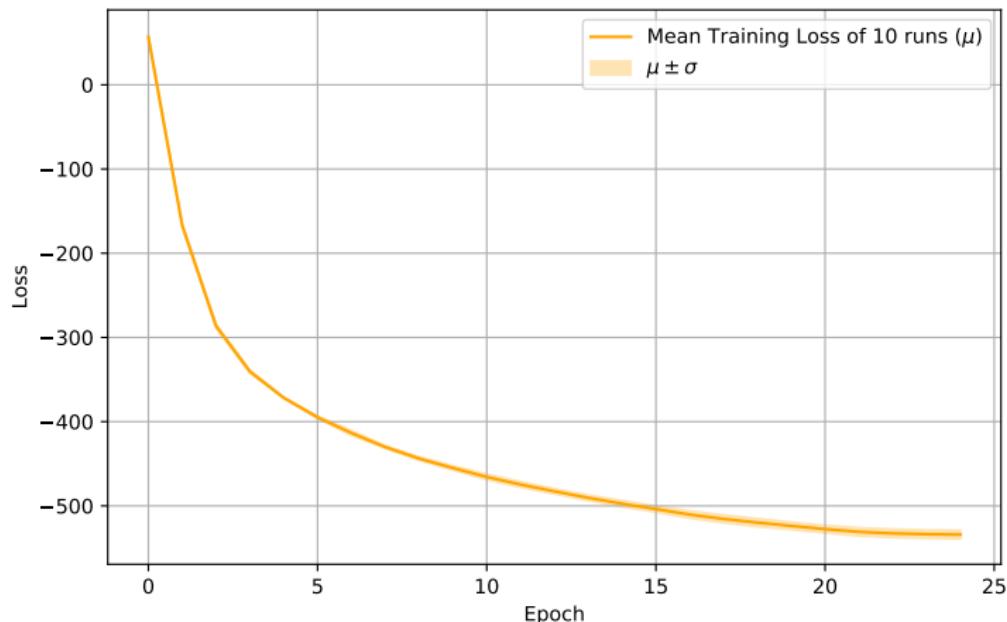


Figure 38: Learning curves for M-N-BEATS-Flow on *Electricity* data set.
Mean plus minus one standard deviation is plotted.

Experiments

Learning Curves

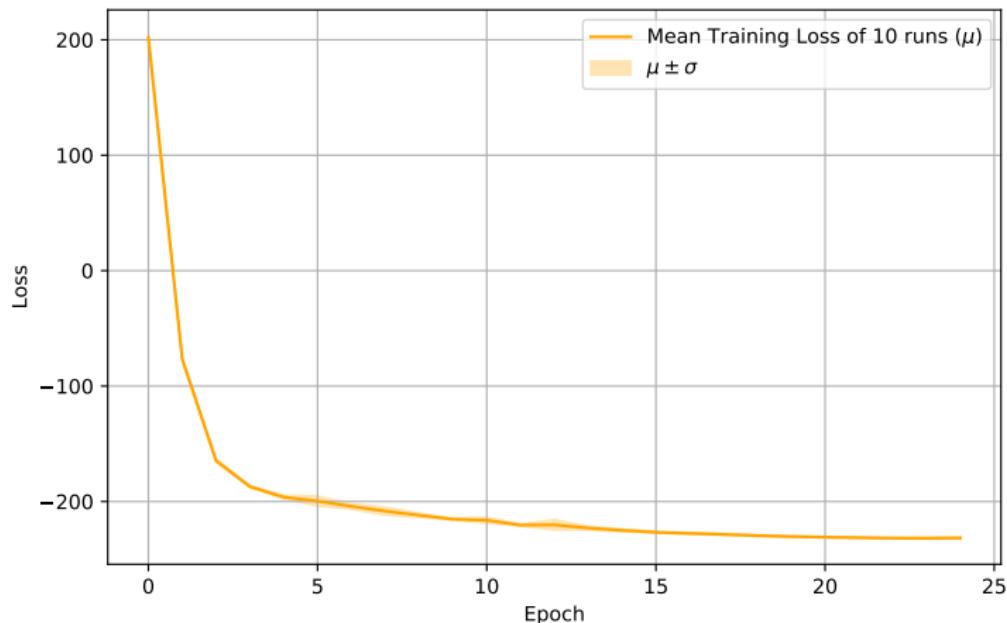


Figure 39: Learning curves for M-N-BEATS-Flow on *Solar* data set.
Mean plus minus one standard deviation is plotted.

Experiments

Learning Curves

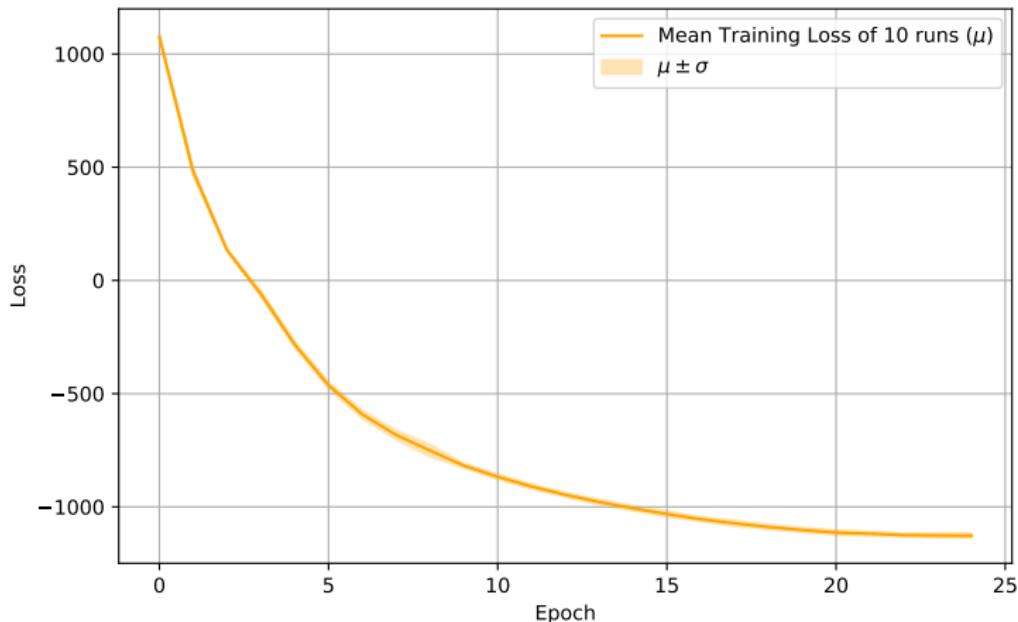


Figure 40: Learning curves for M-N-BEATS-Flow on *Traffic* data set.
Mean plus minus one standard deviation is plotted.

Experiments

Learning Curves

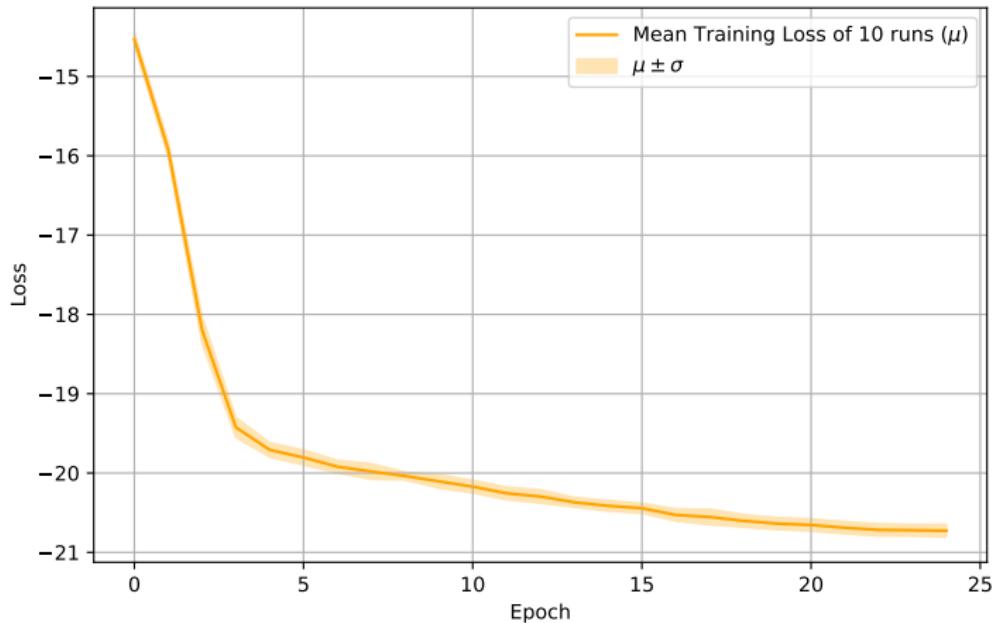


Figure 41: Learning curves for M-N-BEATS-Flow on *Exchange Rate* data set. Mean plus minus one standard deviation is plotted.