NHITS: Neural Hierarchical Interpolation for Time Series Forecasting

Cristian Challu[1]*, Kin G. Olivares[1]*, Boris N. Oreshkin[2], Federico Garza Ramirez[3], Max Mergenthaler-Canseco[3], Artur Dubrawski[1]

[1]Auton Lab, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
[2]Unity Technologies, Labs, Montreal, QC, Canada
[3]Nixtla, Pittsburgh, PA, USA
{cchallu, kdgutier, awd}@cs.cmu.edu, boris.oreshkin@unity3d.com, {federico, max}@nixtla.io

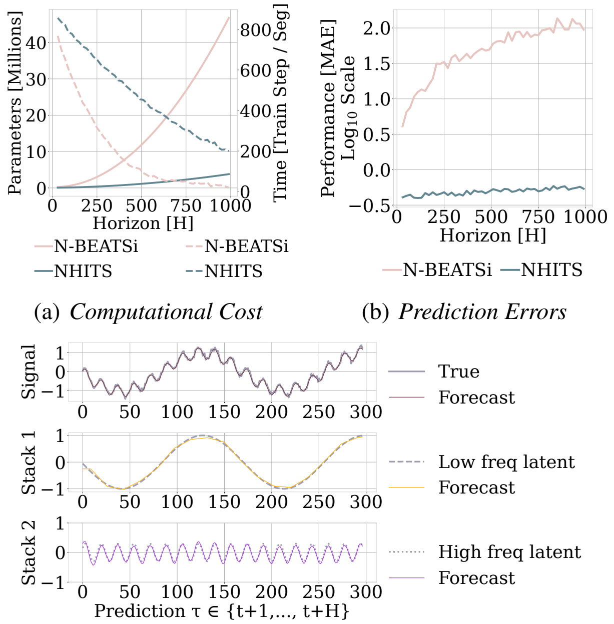
Abstract

Recent progress in neural forecasting accelerated improvements in the performance of large-scale forecasting systems. Yet, long-horizon forecasting remains a very difficult task. Two common challenges afflicting the task are the volatility of the predictions and their computational complexity. We introduce NHITS, a model which addresses both challenges by incorporating novel hierarchical interpolation and multi-rate data sampling techniques. These techniques enable the proposed method to assemble its predictions sequentially, emphasizing components with different frequencies and scales while decomposing the input signal and synthesizing the forecast. We prove that the hierarchical interpolation technique can efficiently approximate arbitrarily long horizons in the presence of smoothness. Additionally, we conduct extensive large-scale dataset experiments from the long-horizon forecasting literature, demonstrating the advantages of our method over the state-of-the-art methods, where NHITS provides an average accuracy improvement of almost 20% over the latest Transformer architectures while reducing the computation time by an order of magnitude (50 times). Our code is available at https://github.com/Nixtla/neuralforecast.

Introduction

Long-horizon forecasting is critical in many important applications, including risk management and planning. Notable examples include power plant maintenance scheduling (Hyndman and Fan 2009) and planning for infrastructure construction (Ziel and Steinert 2018), as well as early warning systems that help mitigate vulnerabilities due to extreme weather events (Basher 2006; Field et al. 2012). In healthcare, predictive monitoring of vital signs enables the detection of preventable adverse outcomes and application of life-saving interventions (Churpek, Adhikari, and Edelson 2016).

Recently, neural time series forecasting has progressed in a few promising directions. First, the architectural evolution included adopting the attention mechanism and the rise of Transformer-inspired approaches (Li et al. 2019; Fan et al. 2019; Alaa and van der Schaar 2019; Lim et al. 2021), as well as the introduction of attention-free architectures composed of deep stacks of fully connected layers (Oreshkin et al. 2020; Olivares et al. 2021a). Both approaches are relatively easy to scale up in terms of capacity, compared to LSTMs, and have proven capable of capturing long-range dependencies. The attention-based approaches are generic as they can explicitly model direct interactions between every pair of input-output elements. Unsurprisingly, they happen to be the most computationally expensive. The architectures based on fully connected stacks implicitly capture input-output relationships and tend to be more compute-efficient. Second, both approaches have replaced the recurrent forecast generation strategy with the multi-step prediction strategy. Aside from



(a) *Computational Cost* (b) *Prediction Errors*

(c) *Neural Hierarchical Interpolation*

Figure 1: (a) The computational costs in time and memory (b) and *mean absolute errors* (MAE) of the predictions of a high capacity fully connected model exhibit evident deterioration with growing forecast horizons. (c) Specializing a flexible model's outputs in the different signal frequencies through hierarchical interpolation combined with multi-rate input processing offers a solution.

*These authors contributed equally.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

its convenient bias-variance benefits and robustness (Marcellino, Stock, and Watson 2006; Atiya and Taieb 2016), the multi-step strategy has enabled the models to efficiently predict long sequences in a single forward pass (Wen et al. 2017; Zhou et al. 2020; Lim et al. 2021).

Despite all the recent progress, long-horizon forecasting remains challenging for neural networks because their unbounded expressiveness translates directly into *excessive computational complexity* and *forecast volatility*, both of which become especially pronounced in this context. For instance, both attention and fully connected layers scale quadratically in memory and computational cost with respect to the forecasting horizon length. Fig. 1 illustrates how forecasting errors and computation costs inflate dramatically with the growing forecasting horizon in the case of the fully connected architecture electricity consumption predictions. Attention-based predictions show similar behavior.

Neural long-horizon forecasting research has mostly focused on attention efficiency making self-attention sparse (Child et al. 2019; Li et al. 2019; Zhou et al. 2020) or local (Li et al. 2019). In the same vein, attention has been cleverly redefined through locality-sensitive hashing (Kitaev, Łukasz Kaiser, and Levskaya 2020) or FFT (Wu et al. 2021). Although that research has led to incremental improvements in computing cost and accuracy, the silver bullet long-horizon forecasting solution is yet to be found. In this paper, we make a bold step in this direction by developing a novel forecasting approach that cuts long-horizon compute cost by an order of magnitude while simultaneously offering 16% accuracy improvements on a large array of multi-variate forecasting datasets compared to existing state-of-the-art Transformer-based techniques. We redefine existing fully-connected `N-BEATS` architecture (Oreshkin et al. 2020) by enhancing its input decomposition via multi-rate data sampling and its output synthesizer via multi-scale interpolation. Our extensive experiments show the importance of the proposed novel architectural components and validate significant improvements in the accuracy and computational complexity of the proposed algorithm.

Our contributions are summarized below:

1. **Multi-Rate Data Sampling**: We incorporate subsampling layers in front of fully-connected blocks, significantly reducing the memory footprint and the amount of computation needed, while maintaining the ability to model long-range dependencies.
2. **Hierarchical Interpolation**: We enforce smoothness of the multi-step predictions by reducing the dimensionality of neural network's prediction and matching its time scale with that of the final output via multi-scale hierarchical interpolation. This novel technique is not unique to our proposed model, and can be incorporated into different architectures.
3. **NHITS architecture**: A novel way of hierarchically synchronizing the rate of input sampling with the scale of output interpolation across blocks, which induces each block to specialize in forecasting its own frequency band of the time-series signal.
4. **State-of-the-art results** on six large-scale benchmark datasets from the long-horizon forecasting literature: electricity transformer temperature, exchange rate, electricity consumption, San Francisco bay area highway traffic, weather, and influenza-like illness.

The remainder of this paper is structured as follows. First, we review the relevant literature. Second, we introduce notation and describe the methodology. After it, we describe and analyze our empirical findings. The last section concludes the paper.

Related Work

Neural forecasting. Over the past few years, deep forecasting methods have become ubiquitous in industrial forecasting systems, with examples in optimal resource allocation and planning in transportation (Laptev et al. 2017), large e-commerce retail (Wen et al. 2017; Olivares et al. 2021b; Paria et al. 2021; Rangapuram et al. 2021), or financial trading (Sezer, Gudelek, and Ozbayoglu 2020). The evident success of the methods in recent forecasting competitions (Makridakis, Spiliotis, and Assimakopoulos 2020, 2021) has renovated the interest within the academic community (Benidis et al. 2020). In the context of multi-variate long-horizon forecasting, Transformer-based approaches have dominated the landscape in recent years, including `Autoformer` (Wu et al. 2021), an encoder-decoder model with decomposition capabilities and an approximation to attention based on Fourier transform, `Informer` (Zhou et al. 2020), Transformer with MLP based multi-step prediction strategy, that approximates self-attention with sparsity, `Reformer` (Kitaev, Łukasz Kaiser, and Levskaya 2020), Transformer that approximates attention with locality-sensitive hashing and `LogTrans` (Li et al. 2019), Transformer with local/log-sparse attention.

Multi-step forecasting. Investigations of the bias/variance trade-off in multi-step forecasting strategies reveal that the *direct* strategy, which allocates a different model for each step, has low bias and high variance, avoiding error accumulation across steps, exhibited by the classical *recursive* strategy, but losing in terms of net model parsimony. Conversely, in the *joint* forecasting strategy, a single model produces forecasts for all steps in one shot, striking the perfect balance between variance and bias, avoiding error accumulation and leveraging shared model parameters (Bao, Xiong, and Hu 2014; Atiya and Taieb 2016; Wen et al. 2017).

Multi-rate input sampling. Previous forecasting literature recognized challenges of extremely long horizon predictions, and proposed *mixed data sampling regression* (`MIDAS`; Ghysels, Sinko, and Valkanov 2007; Armesto, Engemann, and Owyang 2010) to ameliorate the problem of parameter proliferation while preserving high-frequency temporal information. `MIDAS` regressions maintained the classic *recursive* forecasting strategy of linear auto-regressive models but defined a parsimonious fashion of feeding the inputs.

Interpolation. Interpolation has been extensively used to augment the resolution of modeled signals in many fields such as signal and image processing (Meijering 2002). In time-series forecasting, its applications range from com-

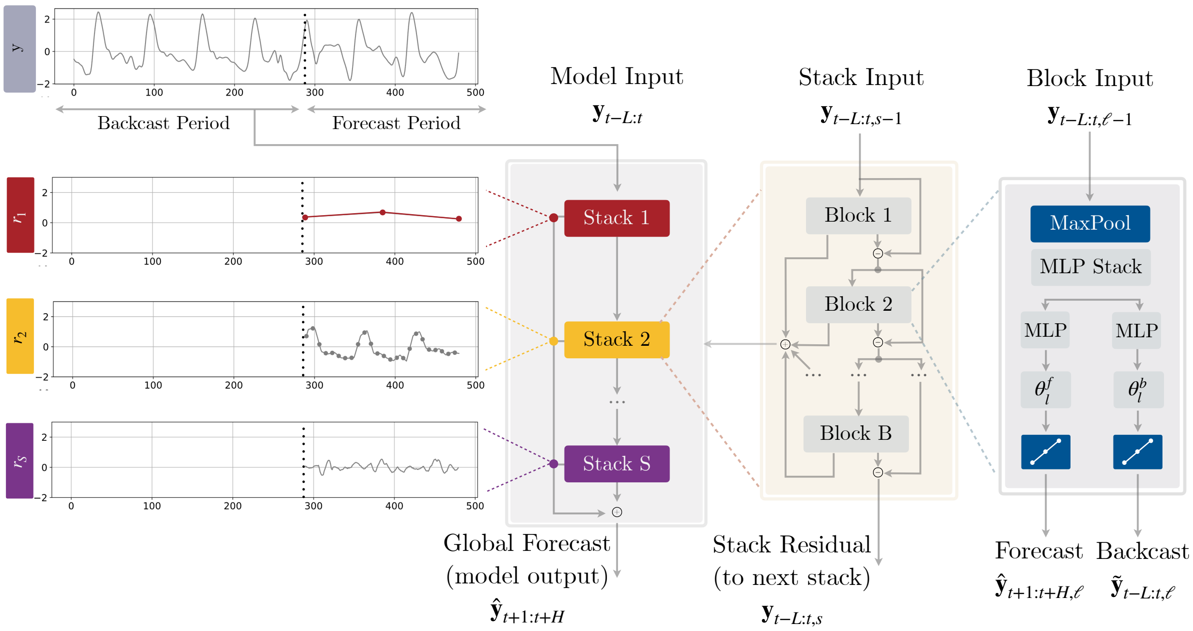


Figure 2: NHITS architecture. The model is composed of several MLPs with ReLU nonlinearities. Blocks are connected via doubly residual stacking principle with the backcast $\tilde{\mathbf{y}}_{t-L:t,\ell}$ and forecast $\hat{\mathbf{y}}_{t+1:t+H,\ell}$ outputs of the ℓ-th block. Multi-rate input pooling, hierarchical interpolation and backcast residual connections together induce the specialization of the additive predictions in different signal bands, reducing memory footprint and compute time, improving architecture parsimony and accuracy.

pleting unevenly sampled data and noise filters (Chow and loh Lin 1971; Fernandez 1981; Shukla and Marlin 2019; Rubanova, Chen, and Duvenaud 2019) to fine-grained quantile-regressions with recurrent networks (Gasthaus et al. 2019). To our knowledge, temporal interpolation has not been used to induce multi-scale hierarchical time-series forecasts.

NHITS Methodology

In this section, we describe our proposed approach, NHITS, whose high-level diagram and main principles of operation are depicted in Fig. 2. Our method extends the *Neural Basis Expansion Analysis* approach (N-BEATS; Oreshkin et al. 2020) in several important respects, making it more accurate and computationally efficient, especially in the context of long-horizon forecasting. In essence, our approach uses multi-rate sampling of the input signal and multi-scale synthesis of the forecast, resulting in a hierarchical construction of forecast, greatly reducing computational requirements and improving forecasting accuracy.

Similarly to N-BEATS, NHITS performs local nonlinear projections onto basis functions across multiple blocks. Each block consists of a *multilayer perceptron* (MLP), which learns to produce coefficients for the backcast and forecast outputs of its basis. The backcast output is used to clean the inputs of subsequent blocks, while the forecasts are summed to compose the final prediction. The blocks are grouped in stacks, each specialized in learning a different characteristic of the data using a different set of basis functions. The overall network input, $\mathbf{y}_{t-L:t}$, consists of L lags.

NHITS is composed of S stacks, B blocks each. Each block contains an MLP predicting forward and backward basis coefficients. The next subsections describe the novel components of our architecture. Note that in the following, we skip the stack index s for brevity.

Multi-Rate Signal Sampling

At the input to each block ℓ, we propose to use a MaxPool layer with kernel size k_ℓ to help it focus on analyzing components of its input with a specific scale. Larger k_ℓ will tend to cut more high-frequency/small-time-scale components from the input of the MLP, forcing the block to focus on analyzing large scale/low frequency content. We call this *multi-rate signal sampling*, referring to the fact that the MLP in each block faces a different effective input signal sampling rate. Intuitively, this helps the blocks with larger pooling kernel size k_ℓ focus on analyzing large scale components critical for producing consistent long-horizon forecasts.

Additionally, multi-rate processing reduces the width of the MLP input for most blocks, limiting the memory footprint and the amount of computation as well as reducing the number of learnable parameters and hence alleviating the effects of overfitting, while maintaining the original receptive field. Given block ℓ input $\mathbf{y}_{t-L:t,\ell}$ (the input to the first block $\ell = 1$ is the network-wide input, $\mathbf{y}_{t-L:t,1} \equiv \mathbf{y}_{t-L:t}$), this operation can be formalized as follows:

$$\mathbf{y}^{(p)}_{t-L:t,\ell} = \mathbf{MaxPool}\left(\mathbf{y}_{t-L:t,\ell},\ k_\ell\right) \tag{1}$$

Non-Linear Regression

Following subsampling, block ℓ looks at its input and nonlinearly regresses forward θ^f_ℓ and backward θ^b_ℓ interpolation

MLP coefficients that learns hidden vector $\mathbf{h}_\ell \in \mathbb{R}^{N_h}$, which is then linearly projected:

$$\begin{aligned}\mathbf{h}_\ell &= \mathbf{MLP}_\ell\left(\mathbf{y}^{(p)}_{t-L:t,\ell}\right)\\ \boldsymbol{\theta}^f_\ell &= \mathbf{LINEAR}^f\left(\mathbf{h}_\ell\right) \quad \boldsymbol{\theta}^b_\ell = \mathbf{LINEAR}^b\left(\mathbf{h}_\ell\right)\end{aligned} \tag{2}$$

The coefficients are then used to synthesize backcast $\tilde{\mathbf{y}}_{t-L:t,\ell}$ and forecast $\hat{\mathbf{y}}_{t+1:t+H,\ell}$ outputs of the block, via the process described below.

Hierarchical Interpolation

In most multi-horizon forecasting models, the cardinality of the neural network prediction equals the dimensionality of the horizon, H. For example, in N-BEATSi $|\boldsymbol{\theta}^f_\ell| = H$; in Transformer-based models, decoder attention layer cross-correlates H output embeddings with L encoded input embeddings (L tends to grow with growing H). This leads to quick inflation in compute requirements and unnecessary explosion in model expressiveness as horizon H increases.

We propose to use *temporal interpolation* to combat these issues. We define the dimensionality of the interpolation coefficients in terms of the *expressiveness ratio* r_ℓ that controls the number of parameters per unit of output time, $|\boldsymbol{\theta}^f_\ell| = \lceil r_\ell H \rceil$. To recover the original sampling rate and predict all H points in the horizon, we use temporal interpolation via the interpolation function g:

$$\begin{aligned}\hat{y}_{\tau,\ell} &= g(\tau, \boldsymbol{\theta}^f_\ell), \quad \forall \tau \in \{t+1, \dots, t+H\},\\ \tilde{y}_{\tau,\ell} &= g(\tau, \boldsymbol{\theta}^b_\ell), \quad \forall \tau \in \{t-L, \dots, t\}.\end{aligned} \tag{3}$$

Interpolation can vary in *smoothness*, $g \in \mathcal{C}^0, \mathcal{C}^1, \mathcal{C}^2$. In Appendix G we explore the nearest neighbor, piece-wise linear, and cubic alternatives. For concreteness, the linear interpolator $g \in \mathcal{C}^1$, along with the time partition $\mathcal{T} = \{t+1, t+1+1/r_\ell, \dots, t+H-1/r_\ell, t+H\}$, is defined as

$$\begin{aligned}g(\tau, \theta) &= \theta[t_1] + \left(\frac{\theta[t_2] - \theta[t_1]}{t_2 - t_1}\right)(\tau - t_1)\\ t_1 &= \arg\min_{t \in \mathcal{T}: t \le \tau} \tau - t, \quad t_2 = t_1 + 1/r_\ell.\end{aligned} \tag{4}$$

The *hierarchical* interpolation principle is implemented by distributing expressiveness ratios across blocks in a manner synchronized with multi-rate sampling. Blocks closer to the input have smaller r_ℓ and larger k_ℓ, implying that input blocks generate low-granularity signals via more aggressive interpolation, being also forced to look at more aggressively sub-sampled (and smoothed) signals. The resulting hierarchical forecast $\hat{\mathbf{y}}_{t+1:t+H}$ is assembled by summing the outputs of all blocks, essentially composing it out of interpolations at different time-scale hierarchy levels.

Since each block specializes on its own scale of input and output signal, this induces a clearly structured hierarchy of interpolation granularity, the intuition conveyed in Fig. 1 and 3. We propose to use *exponentially increasing expressiveness ratios* to handle a wide range of frequency bands while controlling the number of parameters. Alternatively, each stack can specialize in modeling a different known cycle of the time-series (weekly, daily etc.) using a matching r_ℓ (see Table A.3). Finally, the backcast residual formed at previous hierarchy scale is subtracted from the input of the next hierarchy level to amplify the focus of the next level block on signals outside of the band that has already been handled by the previous hierarchy members.

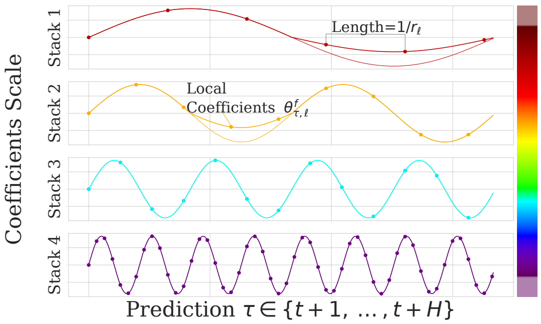


Figure 3: NHITS composes its predictions hierarchically using blocks specializing on different frequencies, through *expressiveness ratios*, and interpolation. The coefficients are locally determined along the horizon, allowing NHITS to reconstruct non-periodic/stationary signals, beyond constant Fourier transform.

$$\hat{\mathbf{y}}_{t+1:t+H} = \sum_{l=1}^{L} \hat{\mathbf{y}}_{t+1:t+H,\ell}$$

$$\mathbf{y}_{t-L:t,\ell+1} = \mathbf{y}_{t-L:t,\ell} - \tilde{\mathbf{y}}_{t-L:t,\ell}$$

Hierarchical interpolation has advantageous theoretical guarantees. We show in Appendix A, that it can approximate infinitely/dense horizons. As long as the interpolating function g is characterized by projections to informed multi-resolution functions V_w, and the forecast relationships are smooth.

Neural Basis Approximation Theorem. Let a forecast mapping be $\mathcal{Y}(\cdot \mid \mathbf{y}_{t-L:t}) : [0,1]^L \to \mathcal{F}$, where the forecast functions $\mathcal{F} = \{\mathcal{Y}(\tau) : [0,1] \to \mathbb{R}\} = \mathcal{L}^2([0,1])$ representing a infinite/dense horizon, are square integrable. If the multi-resolution functions $V_w = \{\phi_{w,h}(\tau) = \phi(2^w(\tau - h)) \mid w \in \mathbb{Z}, h \in 2^{-w} \times [0, \dots, 2^w]\}$ can arbitrarily approximate $\mathcal{L}^2([0,1])$. And the projection $\mathrm{Proj}_{V_w}(\mathcal{Y}(\tau))$ varies smoothly on $\mathbf{y}_{t-L:t}$. Then the forecast mapping $\mathcal{Y}(\cdot \mid \mathbf{y}_{t-L:t})$ can be arbitrarily approximated by a neural basis expansion learning a finite number of multi-resolution coefficients $\hat{\theta}_{w,h}$. That is $\forall \epsilon > 0$,

$$\int |\mathcal{Y}(\tau \mid \mathbf{y}_{t-L:t}) - \sum_{w,h} \hat{\theta}_{w,h}(\mathbf{y}_{t-L:t}) \phi_{w,h}(\tau)| d\tau \le \epsilon \tag{5}$$

Examples of multi-resolution functions $V_w = \{\phi_{w,h}(\tau) = \phi(2^w(\tau - h)) \mid w \in \mathbb{Z}, h \in 2^{-w} \times [0, \dots, 2^w]\}$ include piece-wise constants, piece-wise linear functions and splines with arbitrary approximation capabilities.

Experimental Results

We follow the experimental settings from (Wu et al. 2021; Zhou et al. 2020) (NeurIPS 2021 and AAAI 2021 Best Paper Award). We first describe datasets, baselines and metrics used for the quantitative evaluation of our model. Table 1 presents our key results, demonstrating SoTA performance of our method relative to existing work. We then carefully describe the details of training and evaluation setups. We conclude the section by describing ablation studies.

Datasets

All large-scale datasets used in our empirical studies are publicly available and have been used in neural forecasting literature, particularly in the context of long-horizon (Lai et al. 2017; Zhou et al. 2019; Li et al. 2019; Wu et al. 2021). Table A1 summarizes their characteristics. Each set is normalized with the train data mean and standard deviation.

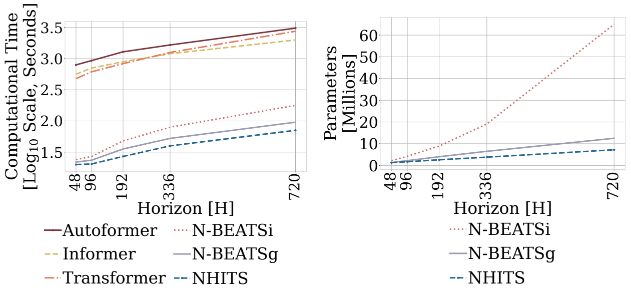
Electricity Transformer Temperature. The $\texttt{ETTm}_2$ dataset measures an electricity transformer from a region of a province of China including oil temperature and variants of load (such as high useful load and high useless load) from July 2016 to July 2018 at a fifteen minutes frequency. **Exchange-Rate.** The `Exchange` dataset is a collection of daily exchange rates of eight countries relative to the US dollar. The countries include Australia, UK, Canada, Switzerland, China, Japan, New Zealand and Singapore from 1990 to 2016. **Electricity.** The `ECL` dataset reports the fifteen minute electricity consumption (KWh) of 321 customers from 2012 to 2014. For comparability, we aggregate it hourly. **San Francisco Bay Area Highway Traffic.** This `TrafficL` dataset was collected by the California Department of Transportation, it reports road hourly occupancy rates of 862 sensors, from January 2015 to December 2016. **Weather.** This `Weather` dataset contains the 2020 year of 21 meteorological measurements recorded every 10 minutes from the Weather Station of the Max Planck Biogeochemistry Institute in Jena, Germany. **Influenza-like illness.** The `ILI` dataset reports weekly recorded influenza-like illness (ILI) patients from Centers for Disease Control and Prevention of the United States from 2002 to 2021. It is a ratio of ILI patients vs. the week's total.

Evaluation Setup

We evaluate the accuracy of our approach using *mean absolute error* (MAE) and *mean squared error* (MSE) metrics, which are well-established in the literature (Zhou et al. 2020; Wu et al. 2021), for varying horizon lengths H:

$$\mathrm{MSE} = \frac{1}{H}\sum_{\tau=t}^{t+H}\left(\mathbf{y}_{\tau} - \hat{\mathbf{y}}_{\tau}\right)^{2}, \qquad \mathrm{MAE} = \frac{1}{H}\sum_{\tau=t}^{t+H}\left|\mathbf{y}_{\tau} - \hat{\mathbf{y}}_{\tau}\right| \tag{6}$$

Note that for multivariate datasets, our algorithm produces forecast for each feature in the dataset and metrics are averaged across dataset features. Since our model is univariate, each variable is predicted using only its own history, $\mathbf{y}_{t-L:t}$, as input. Datasets are partitioned into train, validation and test splits. Train split is used to train model parameters, validation split is used to tune hyperparameters, and test split is used to compute metrics reported in Table 1. Appendix C shows partitioning into train, validation and test splits: seventy, ten, and twenty percent of the available observations respectively, with the exception of $\texttt{ETTm}_2$ that uses twenty percent as validation.



(a) *Time Efficiency* (b) *Memory Efficiency*

Figure 4: Computational efficiency comparison. `NHITS` exhibits the best training time compared to Transformer-based and fully connected models, and smallest memory footprint.

Key Results

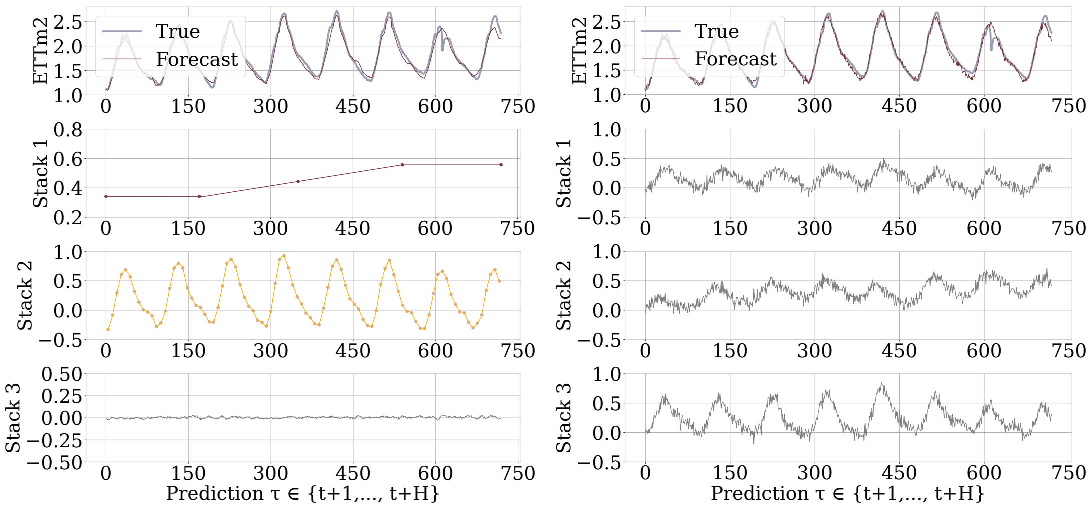
We compare `NHITS` to the following SoTA multivariate baselines: (1) `FEDformer` (Zhou et al. 2022), (2) `Autoformer` (Wu et al. 2021), (3) `Informer` (Zhou et al. 2020), (4) `Reformer` (Kitaev, Łukasz Kaiser, and Levskaya 2020) and (5) `LogTrans` (Li et al. 2019). Additionally, we consider the univariate baselines: (6) `DilRNN` (Chang et al. 2017) and (7) auto-`ARIMA` (Hyndman and Khandakar 2008).

Forecasting Accuracy. Table 1 summarizes the multivariate forecasting results. `NHITS` outperforms the best baseline, with average relative error decrease across datasets and horizons of 14% in MAE and 16% in MSE. `NHITS` maintains a comparable performance to other state-of-the-art methods for the shortest measured horizon (96/24), while for the longest measured horizon (720/60) decreases multivariate MAE by 11% and MSE by 17%. We complement the key results in Table 1, with the additional univariate forecasting experiments in Appendix F, again demonstrating state-of-the-art performance against baselines.

Computational Efficiency. We measure the computational training time of `NHITS`, `N-BEATS` and Transformer-based methods in the multivariate setting and show compare in Figure 4. The experiment monitors the whole training process for the $\texttt{ETTm}_2$ dataset. For the Transformer-based models we used hyperparameters reported in (Wu et al. 2021). Compared to the Transformer-based methods, `NHITS` is 45× faster than `Autoformer`. In terms of memory, `NHITS` has less than 26% of the parameters of the second-best alternative since it scales linearly with respect to the input's length. Compared to the original `N-BEATS`, our method is 1.26× faster and requires only 54% of the parameters. Finally, while `NHITS` is an univariate model, it has *global* (shared) parameters for all time-series in the dataset. Just like (Oreshkin et al. 2020), our experiments (Appendix I) show that `NHITS` maintains constant parameter/training computational complexity regarding dataset's size.

		NHITS		N-BEATS		FEDformer		Autoformer		Informer		LogTrans		DilRNN		ARIMA	
	H.	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTm2	96	**0.176**	**0.255**	0.184	0.263	0.203	0.287	0.255	0.339	0.365	0.453	0.768	0.642	0.343	0.401	0.225	0.301
	192	**0.245**	**0.305**	0.273	0.337	0.269	0.328	0.281	0.340	0.533	0.563	0.989	0.757	0.424	0.468	0.298	0.345
	336	**0.295**	**0.346**	0.309	0.355	0.325	0.366	0.339	0.372	1.363	0.887	1.334	0.872	0.632	1.083	0.370	0.386
	720	**0.401**	**0.413**	0.411	0.425	0.421	0.415	0.422	0.419	3.379	1.388	3.048	1.328	0.634	0.594	0.478	0.445
ECL	96	0.147	0.249	**0.145**	**0.247**	0.183	0.297	0.201	0.317	0.274	0.368	0.258	0.357	0.233	0.927	1.220	0.814
	192	**0.167**	**0.269**	0.180	0.283	0.195	0.308	0.222	0.334	0.296	0.386	0.266	0.368	0.265	0.921	1.264	0.842
	336	**0.186**	**0.290**	0.200	0.308	0.212	0.313	0.231	0.338	0.300	0.394	0.280	0.380	0.235	0.896	1.311	0.866
	720	0.243	**0.340**	0.266	0.362	**0.231**	0.343	0.254	0.361	0.373	0.439	0.283	0.376	0.322	0.890	1.364	0.891
Exchange	96	**0.092**	**0.202**	0.098	0.206	0.139	0.276	0.197	0.323	0.847	0.752	0.968	0.812	0.383	0.45	0.296	0.214
	192	**0.208**	**0.322**	0.225	0.329	0.256	0.369	0.300	0.369	1.204	0.895	1.040	0.851	1.123	0.834	1.056	0.326
	336	**0.301**	**0.403**	0.493	0.482	0.426	0.464	0.509	0.524	1.672	1.036	1.659	1.081	1.612	1.051	2.298	0.467
	720	**0.798**	**0.596**	1.108	0.804	1.090	0.800	1.447	0.941	2.478	1.310	1.941	1.127	1.827	1.131	20.666	0.864
TrafficL	96	0.402	**0.282**	**0.398**	**0.282**	0.562	0.349	0.613	0.388	0.719	0.391	0.684	0.384	0.580	0.308	1.997	0.924
	192	0.420	0.297	**0.409**	**0.293**	0.562	0.346	0.616	0.382	0.696	0.379	0.685	0.390	0.739	0.383	2.044	0.944
	336	**0.448**	**0.313**	0.449	0.318	0.570	0.323	0.622	0.337	0.777	0.420	0.733	0.408	0.804	0.419	2.096	0.960
	720	**0.539**	**0.353**	0.589	0.391	0.596	0.368	0.660	0.408	0.864	0.472	0.717	0.396	0.695	0.372	2.138	0.971
Weather	96	**0.158**	**0.195**	0.167	0.203	0.217	0.296	0.266	0.336	0.300	0.384	0.458	0.490	0.193	0.245	0.217	0.258
	192	**0.211**	**0.247**	0.229	0.261	0.276	0.336	0.307	0.367	0.598	0.544	0.658	0.589	0.255	0.306	0.263	0.299
	336	**0.274**	**0.300**	0.287	0.304	0.339	0.380	0.359	0.395	0.578	0.523	0.797	0.652	0.329	0.360	0.330	0.347
	720	**0.351**	**0.353**	0.368	0.359	0.403	0.428	0.419	0.428	1.059	0.741	0.869	0.675	0.521	0.495	0.425	0.405
ILI	24	**1.862**	**0.869**	1.879	0.886	2.203	0.963	3.483	1.287	5.764	1.677	4.480	1.444	4.538	1.449	5.554	1.434
	36	**2.071**	**0.934**	2.210	1.018	2.272	0.976	3.103	1.148	4.755	1.467	4.799	1.467	3.709	1.273	6.940	1.676
	48	**2.134**	**0.932**	2.440	1.088	2.209	0.981	2.669	1.085	4.763	1.469	4.800	1.468	3.436	1.238	7.192	1.736
	60	**2.137**	**0.968**	2.547	1.057	2.545	1.061	2.770	1.125	5.264	1.564	5.278	1.560	3.703	1.272	6.648	1.656

Table 1: Main empirical results in long-horizon forecasting setup, lower scores are better. Metrics are averaged over eight runs, best results are highlighted in bold. In Appendix E we complement the main results with standard deviations.



(a) hie. interp., multi-rate sampling

(b) No hie. interp., multi-rate sampling

Figure 5: ETTm2 and 720 ahead forecasts using NHITS (left panel), NHITS with hierarchical linear interpolation and multi-rate sampling removed (right panel). The top row shows the original signal and the forecast. The second, third and fourth rows show the forecast components for each stack. The last row shows the residuals, $y - \hat{y}$. In (a), each block shows scale specialization, unlike (b), in which signals are not interpretable.

H.	NHITS	NHITS$_2$	NHITS$_3$	NHITS$_4$	N-BEATSi
96	0.239	0.241	**0.237**	0.240	0.254
192	**0.290**	0.299	0.291	0.300	0.307
336	**0.338**	0.342	0.346	0.352	0.405
720	**0.439**	0.450	0.454	0.468	0.597

Table 2: Empirical evaluation of long multi-horizon multivariate forecasts for NHITS with/without enhancements. MAE for predictions averaged over eight runs, and five datasets.

Training and Hyperparameter Optimization

We consider a minimal search space. We tune the kernel size for multi-rate sampling from Equation (1) and the number of coefficients from Equation (2), some matching common seasonalities and others exponentially increasing. Additionally, we tune the random seed to escape underperforming local minima. Details are reported in Table A3 in Appendix D.

During the *hyperparameter optimization phase*, we measure MAE on the validation set and use a Bayesian optimization library (HYPEROPT; Bergstra et al. 2011), with 20 iterations. We use the optimal configuration based on the validation loss to make predictions on the test set. We refer to the combination of hyperparameter optimization and test prediction as a *run*. NHITS is implemented in PyTorch (Paszke et al. 2019) and trained using ADAM optimizer (Kingma and Ba 2014), MAE loss, batch size 256 and initial learning rate of 1e-3, halved three times across the training procedure. All our experiments are conducted on a GeForce RTX 2080 GPU.

Ablation Studies

We believe that the advantages of the NHITS architecture are rooted in its multi-rate hierarchical nature. Fig. 5 shows a qualitative comparison of NHITS with and without hierarchical interpolation/multi-rate sampling components. Unlike the control model, we clearly see NHITS developing the ability to produce interpretable forecast decomposition providing valuable information about trends and seasonality in separate channels. Appendix G presents the decomposition for the different interpolation techniques.

We support our qualitative conclusion with quantitative results. We define the following set of alternative models: NHITS$_2$ only hierarchical interpolation, NHITS$_3$ only multi-rate sampling, NHITS$_4$ no multi-rate sampling or interpolation (corresponds to the original N-BEATSg (Oreshkin et al. 2020)), finally N-BEATSi, the interpretable version of the N-BEATS ((Oreshkin et al. 2020)). Tab. 2 clearly shows that combining both proposed components results in the best performance, emphasizing their complementary nature in long-horizon forecasting. We see that the original N-BEATS is consistently worse, especially the N-BEATSi. The advantages of multi-rate sampling, and interpolation for long-horizon forecasting, are not limited to the NHITS architecture. In Appendix H we demonstrate how adding them to a DilRNN improves its performance.

Additional *ablation studies* are reported in Appendix G. The MaxPool multi-rate sampling wins over AveragePool. Linear interpolation wins over nearest neighbor and cubic. Finally and most importantly, we show that the order in which hierarchical interpolation is implemented matters significantly. The best configuration is to have the low-frequency/large-scale components synthesized and removed from analysis first, followed by more fine-grained modeling of high-frequency/intermittent signals.

Discussion of Findings

Our results indicate the complementarity and effectiveness of multi-rate sampling and hierarchical interpolation for long-horizon time-series forecasting. Table 2 indicates that these components enforce a useful inductive bias compared to both the free-form model NHITS$_4$ (plain fully connected architecture) and the parametric model N-BEATSi (polynomial trend and sinusoidal seasonality used as basis functions in two respective stacks). The latter provides a detrimental inductive bias for long-horizon forecasting. We barely scratched the surface in the right direction and further progress is possible using advanced multi-scale processing approaches in the forecasting context, motivating further research.

NHITS outperforms SoTA baselines and provides an interpretable non-linear decomposition. Fig. 1 and 5 showcase NHITS perfectly specializing and reconstructing latent harmonic signals from synthetic and real data respectively. This novel *interpretable* decomposition can provide insights to users, improving their confidence in high-stakes applications like healthcare. Finally, NHITS hierarchical interpolation is connected to Wavelet's multi-resolution analysis (Daubechies 1992). Replacing the interpolation functions with orthogonal Wavelet spaces is a possible research line.

Our study questions the effectiveness of existing long-horizon multi-variate forecasting approaches, as all of them are substantially outperformed by our univariate algorithm. If these approaches underperform due to overfitting problems at the level of marginals, the integration of our approach with Transformer-inspired architectures is a promising research direction as Appendix F results suggest. However, there is a chance that existing approaches underperform due to their inability to integrate information from multiple variables, which clearly hints at possibly untapped research potential. Whichever is the case, we believe our results provide a strong guidance signal and a valuable baseline for future research in the area of long-horizon multivariate forecasting.

Conclusions

We proposed a novel neural forecasting algorithm NHITS that combines two complementary techniques, multi-rate input sampling, and hierarchical interpolation, to produce drastically improved, interpretable, and computationally efficient long-horizon time-series predictions. Our model, operating in the univariate regime and accepting only the predicted time-series history, significantly outperforms all previous Transformer-based multi-variate models using an order of magnitude less computation. This sets a new baseline for all ensuing multivariate work on six popular datasets and motivates research to effectively use information across variables.

Acknowledgements

This work was partially supported by the Defense Advanced Research Projects Agency (award FA8750-17-2-0130), the National Science Foundation (grant 2038612), the Space Technology Research Institutes grant from NASA's Space Technology Research Grants Program, the U.S. Department of Homeland Security (award 18DN-ARI-00031), and by the U.S. Army Contracting Command (contracts W911NF20D0002 and W911NF22F0014 delivery order #4). Thanks to Mengfei Cao for in-depth discussion and comments on the method, and Kartik Gupta for his insights on the connection of `NHITS` with Wavelet's theory. The authors are also grateful to Stefania La Vattiata for her assistance in the upbeat visualization of the *Neural Hierarchical Interpolation for Time Series* method.

References

Alaa, A. M.; and van der Schaar, M. 2019. Attentive State-Space Modeling of Disease Progression. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, volume 32. Curran Associates, Inc.

Armesto, M. T.; Engemann, K. M.; and Owyang, M. T. 2010. Forecasting with Mixed Frequencies. *Federal Reserve Bank of St. Louis Review*, 92: 521–536.

Atiya, A.; and Taieb, B. 2016. A Bias and Variance Analysis for Multistep-Ahead Time Series Forecasting. *IEEE transactions on neural networks and learning systems*, 27(1): 2162–2388.

Bao, Y.; Xiong, T.; and Hu, Z. 2014. Multi-step-ahead time series prediction using multiple-output support vector regression. *Neurocomputing*, 129: 482–493.

Basher, R. 2006. Global early warning systems for natural hazards: Systematic and people-centred. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 364: 2167–82.

Benidis, K.; Rangapuram, S. S.; Flunkert, V.; Wang, B.; Maddix, D.; Turkmen, C.; Gasthaus, J.; Bohlke-Schneider, M.; Salinas, D.; Stella, L.; Callot, L.; and Januschowski, T. 2020. Neural forecasting: Introduction and literature overview. *Computing Research Repository*.

Bergstra, J.; Bardenet, R.; Bengio, Y.; and Kégl, B. 2011. Algorithms for Hyper-Parameter Optimization. In Shawe-Taylor, J.; Zemel, R.; Bartlett, P.; Pereira, F.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems*, volume 24, 2546–2554. Curran Associates, Inc.

Chang, S.; Zhang, Y.; Han, W.; Yu, M.; Guo, X.; Tan, W.; Cui, X.; Witbrock, M.; Hasegawa-Johnson, M. A.; and Huang, T. S. 2017. Dilated Recurrent Neural Networks. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Child, R.; Gray, S.; Radford, A.; and Sutskever, I. 2019. Generating Long Sequences with Sparse Transformers. *CoRR*, abs/1904.10509.

Chow, G. C.; and loh Lin, A. 1971. Best Linear Unbiased Interpolation, Distribution, and Extrapolation of Time Series by Related Series. *The Review of Economics and Statistics*, 53(4): 372–375.

Churpek, M. M.; Adhikari, R.; and Edelson, D. P. 2016. The value of vital sign trends for detecting clinical deterioration on the wards. *Resuscitation*, 102: 1–5.

Daubechies, I. 1992. *Ten lectures on wavelets*. SIAM.

Fan, C.; Zhang, Y.; Pan, Y.; Li, X.; Zhang, C.; Yuan, R.; Wu, D.; Wang, W.; Pei, J.; and Huang, H. 2019. Multi-Horizon Time Series Forecasting with Temporal Attention Learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, 2527–2535. New York, NY, USA: Association for Computing Machinery. ISBN 9781450362016.

Fernandez, R. B. 1981. A Methodological Note on the Estimation of Time Series. *The Review of Economics and Statistics*, 63(3): 471–476.

Field, C. B.; Barros, V.; Stocker, T. F.; and Dahe, Q. 2012. *Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change*. Cambridge University Press.

Gasthaus, J.; Benidis, K.; Wang, B.; Rangapuram, S. S.; Salinas, D.; Flunkert, V.; and Januschowski, T. 2019. Probabilistic Forecasting with Spline Quantile Function RNNs. In *AISTATS*.

Ghysels, E.; Sinko, A.; and Valkanov, R. 2007. MIDAS Regressions: Further Results and New Directions. *Econometric Reviews*, 26(1): 53–90.

Hyndman, R. J.; and Fan, S. 2009. Density forecasting for long-term peak electricity demand. *IEEE Transactions on Power Systems*, 25(2): 1142–1153.

Hyndman, R. J.; and Khandakar, Y. 2008. Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software, Articles*, 27(3): 1–22.

Kingma, D. P.; and Ba, J. 2014. ADAM: A Method for Stochastic Optimization. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations (ICLR), San Diego, 2015.

Kitaev, N.; Łukasz Kaiser; and Levskaya, A. 2020. Reformer: The Efficient Transformer. In *8th International Conference on Learning Representations, (ICLR 2020)*.

Lai, G.; Chang, W.; Yang, Y.; and Liu, H. 2017. Modeling Long- and Short-Term Temporal Patterns with Deep Neural Networks. *Special Interest Group on Information Retrieval Conference 2018 (SIGIR 2018)*, abs/1703.07015.

Laptev, N.; Yosinsk, J.; Erran, L. L.; and Smyl, S. 2017. Time-series extreme event forecasting with neural networks at UBER. In *34th International Conference on Machine Learning ICML 2017, Time Series Workshop*.

Li, S.; Jin, X.; Xuan, Y.; Zhou, X.; Chen, W.; Wang, Y.; and Yan, X. 2019. Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, volume 32. Curran Associates, Inc.

Lim, B.; Arık, S. Ö.; Loeff, N.; and Pfister, T. 2021. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*.

Makridakis, S.; Spiliotis, E.; and Assimakopoulos, V. 2020. The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1): 54–74. M4 Competition.

Makridakis, S.; Spiliotis, E.; and Assimakopoulos, V. 2021. Predicting/hypothesizing the findings of the M5 competition. *International Journal of Forecasting*.

Marcellino, M.; Stock, J. H.; and Watson, M. W. 2006. A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics*, 135(1): 499–526.

Meijering, E. 2002. A chronology of interpolation: from ancient astronomy to modern signal and image processing. *Proceedings of the IEEE*, 90(3): 319–342.

Olivares, K. G.; Challu, C.; Marcjasz, G.; Weron, R.; and Dubrawski, A. 2021a. Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with NBEATSx. *International Journal of Forecasting, submitted*, Working Paper version available at arXiv:2104.05522.

Olivares, K. G.; Meetei, N. O.; Ma, R.; Reddy, R.; Cao, M.; and Dicker, L. 2021b. Probabilistic Hierarchical Forecasting with Deep Poisson Mixtures. *International Journal of Forecasting (Hierarchical Forecasting special issue), submitted*, Working Paper version available at arXiv:2110.13179.

Oreshkin, B. N.; Carpov, D.; Chapados, N.; and Bengio, Y. 2020. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In *8th International Conference on Learning Representations, ICLR 2020*.

Paria, B.; Sen, R.; Ahmed, A.; and Das, A. 2021. Hierarchically Regularized Deep Forecasting. In *Submitted to Proceedings of the 39th International Conference on Machine Learning*. PMLR. Working Paper version available at arXiv:2106.07630.

Paszke et al. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32*, 8024–8035. Curran Associates, Inc.

Rangapuram, S. S.; Werner, L. D.; Benidis, K.; Mercado, P.; Gasthaus, J.; and Januschowski, T. 2021. End-to-End Learning of Coherent Probabilistic Forecasts for Hierarchical Time Series. In Balcan, M. F.; and Meila, M., eds., *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR.

Rubanova, Y.; Chen, R. T. Q.; and Duvenaud, D. 2019. Latent ODEs for Irregularly-Sampled Time Series. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2019)*.

Sezer, O. B.; Gudelek, M. U.; and Ozbayoglu, A. M. 2020. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied soft computing*, 90: 106181.

Shukla, S. N.; and Marlin, B. M. 2019. Interpolation-Prediction Networks for Irregularly Sampled Time Series. In *7th International Conference on Learning Representations, (ICLR 2019)*.

Wen, R.; Torkkola, K.; Narayanaswamy, B.; and Madeka, D. 2017. A Multi-Horizon Quantile Recurrent Forecaster. In *31st Conference on Neural Information Processing Systems NIPS 2017, Time Series Workshop*.

Wu, H.; Xu, J.; Wang, J.; and Long, M. 2021. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. In Ranzato, M.; Beygelzime, A.; Liang, P.; Vaughan, J.; and Dauphin, Y., eds., *Advances in Neural Information Processing Systems 35 (NeurIPS 2021)*.

Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2020. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *The Association for the Advancement of Artificial Intelligence Conference 2021 (AAAI 2021).*, abs/2012.07436.

Zhou, S.; Zhou, L.; Mao, M.; Tai, H.; and Wan, Y. 2019. An Optimized Heterogeneous Structure LSTM Network for Electricity Price Forecasting. *IEEE Access*, 7: 108161–108173.

Zhou, T.; Ma, Z.; Wen, Q.; Wang, X.; Sun, L.; and Jin, R. 2022. FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting. *Computing Research Repository*, abs/2201.12740.

Ziel, F.; and Steinert, R. 2018. Probabilistic mid- and long-term electricity price forecasting. *Renewable and Sustainable Energy Reviews*, 94: 251–266.