NHITS: Neural Hierarchical Interpolation for Time Series Forecasting

Cristian Challu\textsuperscript{1*}, Kin G. Olivares\textsuperscript{1*}, Boris N. Oreshkin\textsuperscript{2}, Federico Garza Ramirez\textsuperscript{3}, Max Mergenthaler-Canseco\textsuperscript{3}, Artur Dubrawski\textsuperscript{1}

\textsuperscript{1}Auton Lab, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
\textsuperscript{2}Unity Technologies, Labs, Montreal, QC, Canada
\textsuperscript{3}Nixtla, Pittsburgh, PA, USA

\{cchallu, kdgutier, awd\}@cs.cmu.edu, boris.oshchen@gmail.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
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, Transformer (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; Armstrong, 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-
work input, whose high-level diagram and main principles of operation
Expansion Analysis approach (forecast, greatly reducing computational requirements and
block consists of a data using a different set of basis functions. The overall net-
pose the final prediction. The blocks are grouped in stacks, resulting in a hierarchical construction of
projections onto basis functions across multiple blocks. Each
rate sampling of the input signal and multi-scale synthesis
long-horizon forecasting. In essence, our approach uses multi-
used to induce multi-scale hierarchical time-series forecasts.

2020) in several important respects, making it more accurate
are depicted in Fig. 2. Our method extends the
in different signal bands, reducing memory footprint and compute time, improving architecture parsimony and accuracy.

\[
\hat{y}_{t+1:t+H} \quad \text{Stack Residual} \quad \text{to next stack} \quad y_{t-L:t}
\]

\[
\hat{y}_{t-L:t,\ell} \quad \text{Stack Input} \quad y_{t-L:t,\ell+1}
\]

\[
\hat{y}_{t-L:t} \quad \text{Model Input} \quad y_{t-L:t}
\]

\[
\hat{y}_{t-L:t,\ell} \quad \text{Block Input} \quad \theta_{\ell,1} \quad \theta_{\ell,2} \quad \ldots \quad \theta_{\ell,B}
\]

\[
\text{MaxPool} \quad \text{MLP Stack}
\]

\[
\hat{y}_{t-L:t,\ell} \quad \text{Backcast}
\]

\[
\text{Forecast}
\]

\[
\hat{y}_{t+1:t+H,\ell} \quad \text{Stack 1}
\]

\[
\text{Stack 2}
\]

\[
\text{Stack 3}
\]

\[
\text{Global Forecast (model output)} \quad \hat{y}_{t+1:t+H}
\]

\[
\text{Block B}
\]

\[
\text{Block 2}
\]

\[
\text{Block 1}
\]

\[
\hat{y}_{t-L:t,\ell} \quad \text{Block Input}
\]

\[
y_{t-L:t,\ell}
\]

\[
y_{t-L:t,\ell-1}
\]

\[
y_{t-L:t-1}
\]

\[
y_{t-L:1}
\]

\[
y_{t-L:0}
\]

\[
y_{t-L:-1}
\]

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 \(\hat{y}_{t-L:t,\ell}\) and forecast \(\hat{y}_{t+1:t+H,\ell}\) outputs of the \(\ell\)-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.

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 com-
pose 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, \(y_{t-L:t}\), 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 \(\ell\), we propose to use a MaxPool layer with kernel size \(k_\ell\) to help it focus on analyzing compo-
nents of its input with a specific scale. Larger \(k_\ell\) 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 \(k_\ell\) 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 num-
ber of learnable parameters and hence alleviating the effects
of overfitting, while maintaining the original receptive field.

Given block \(\ell\) input \(y_{t-L:t,\ell}\) (the input to the first block \(\ell = 1\) is the network-wide input, \(y_{t-L:t,1} \equiv y_{t-L:t}\)), this operation

\[
y_{t-L:t,\ell}^{(p)} = \text{MaxPool} \left( y_{t-L:t,\ell}, k_\ell \right)
\]

Non-Linear Regression

Following subsampling, block \(\ell\) looks at its input and non-
linearly regresses forward \(\theta_\ell^f\) and backward \(\theta_\ell^b\) interpolation
MLP coefficients that learns hidden vector \( h_{\ell} \in \mathbb{R}^{N_h} \), which is then linearly projected:

\[
    h_{\ell} = \text{MLP}_{\ell}\left(y_{t-L:t,\ell}^{(p)}\right)
\]

\[
    \theta^l_{\ell} = \text{LINEAR}^l(h_{\ell}) \quad \theta^b_{\ell} = \text{LINEAR}^b(h_{\ell})
\]

The coefficients are then used to synthesize backcast \( \hat{y}_{t-L:t,\ell} \) and forecast \( \tilde{y}_{t+1:t+H,\ell} \) 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 \textsc{N-Beats} \( |\theta^f| = H \); 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* \( r_{\ell} \) that controls the number of parameters per unit of output time, \( |\theta^f| = \lfloor r_{\ell} H \rfloor \). To recover the original sampling rate and predict all \( H \) points in the horizon, we use temporal interpolation via the interpolation function \( g \):

\[
    \hat{y}_{\tau,\ell} = g(\tau, \theta^f_{\ell}), \quad \forall \tau \in \{t + 1, \ldots, t + H\},
\]

\[
    \tilde{y}_{\tau,\ell} = g(\tau, \theta^b_{\ell}), \quad \forall \tau \in \{t - L, \ldots, t\}.
\]

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 \( \mathcal{T} = \{t + 1, t + 1/r_{\ell}, \ldots, t + H - 1/r_{\ell}, t + H\} \), is defined as

\[
    g(\tau, \theta) = \theta[t_1] + \left( \frac{\theta[t_2] - \theta[t_1]}{t_2 - t_1} \right) (\tau - t_1)
\]

\[
    t_1 = \arg \min_{t \in \mathcal{T} : t \leq \tau} \tau - t, \quad t_2 = t_1 + 1/r_{\ell}.
\]

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 \( r_{\ell} \) and larger \( k_{\ell} \), 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 \( \hat{y}_{t+1:t+H} \) 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 \( r_{\ell} \) (see Table A.3). Finally, the backcast residual formed at previous hierarchy members.

\[
    \tilde{y}_{t+1:t+H} = \sum_{l=1}^{L} \tilde{y}_{t+1:t+H,\ell}
\]

\[
    y_{t-L:t+1} = y_{t-L:t,\ell} - \tilde{y}_{t-L:t,\ell}
\]

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 \( \mathcal{Y}(\cdot \mid y_{t-L:t}) : [0, 1]^L \to \mathcal{F} \), 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 \( V_w = \{\phi_{w,h}(\tau) = \phi(2^w(\tau - h)) \mid w \in \mathbb{Z}, h \in 2^{-w} \times [0, \ldots, 2^{w}]\} \) can arbitrarily approximate \( \mathcal{L}^2([0, 1]) \). And the projection \( \text{Proj}_{V_w}(\mathcal{Y}(\tau)) \) varies smoothly on \( y_{t-L:t} \). Then the forecast mapping \( \mathcal{Y}(\cdot \mid y_{t-L:t}) \) can be arbitrarily approximated by a neural basis expansion learning a finite number of multi-resolution coefficients \( \theta_{w,h} \).

That is \( \forall \epsilon > 0,
\]

\[
    \int |\mathcal{Y}(\tau) - y_{t-L:t} - \sum_{w,h} \theta_{w,h}(y_{t-L:t})\phi_{w,h}(\tau)|d\tau \leq \epsilon
\]

Examples of multi-resolution functions \( V_w = \{\phi_{w,h}(\tau) = \phi(2^w(\tau - h)) \mid w \in \mathbb{Z}, h \in 2^{-w} \times [0, \ldots, 2^{w}]\} \) 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 ETTm 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 $H$:

$$\text{MSE} = \frac{1}{H} \sum_{\tau=t}^{t+H} (y_\tau - \hat{y}_\tau)^2, \quad \text{MAE} = \frac{1}{H} \sum_{\tau=t}^{t+H} |y_\tau - \hat{y}_\tau|$$

(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, $y_{t-L:t}$, 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 ETTm2 that uses twenty percent as validation.

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 ETTm2 dataset. For the Transformer-based models we used hyperparameters reported in (Wu et al. 2021). Compared to the Transformer-based methods, NHITS is $45 \times$ 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 $1.26 \times$ 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.
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.

| H     | NHITS | N-BEATS | FEDformer | Autoformer | Informer | LogTrans | DilRNN | ARIMA |
|-------|-------|---------|-----------|------------|----------|----------|--------|-------|
|       | MSE   | MAE     | MSE       | MSE        | MSE      | MSE      | MSE    | MSE   |
| 96    | 0.176 | 0.202   | 0.184     | 0.203      | 0.207    | 0.208    | 0.207  | 0.207 |
| 192   | 0.245 | 0.235   | 0.231     | 0.256      | 0.257    | 0.258    | 0.258  | 0.258 |
| 336   | 0.295 | 0.270   | 0.293     | 0.311      | 0.313    | 0.315    | 0.315  | 0.315 |
| 720   | 0.401 | 0.341   | 0.366     | 0.399      | 0.402    | 0.405    | 0.405  | 0.405 |

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.
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.

|    | NHITS | NHITS₂ | NHITS₃ | NHITS₄ | 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    |

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 NHITS₄ (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 Multi-step-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.
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.

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.