Individual Report

COMP5530 Group Project: A Comparison of Inflation Forecasting Models

AYESHA RAHMAN, University of Leeds, United Kingdom

1 INTRODUCTION

This report outlines my individual contribution to our group project on forecasting U.S. inflation (PCEPI) using statistical, machine learning, and deep learning models. I implemented four models (ARDL, N-BEATS, N-BEATSx, and N-HiTS), curated the initial exogenous variable list and co-developed the group's standardized preprocessing pipeline with George Edward Bignall. This pipeline ensured consistent feature engineering, scaling, and fair evaluation across models. Although I contributed to multiple forecasting methods, this report focuses exclusively on my implementation and analysis of the N-HiTS (Neural Hierarchical Interpolation for Time Series) model. Given its architectural strengths and superior empirical performance over N-BEATS and N-BEATSx in our inflation forecasting task, N-HiTS was selected as the focus of this report. The complete project codebase is available on GitHub [Rahman et al. 2025].

1.1 Background research

N-HiTS (Neural Hierarchical Interpolation for Time Series) is a deep learning model for multi-step forecasting that extends N-BEATS via hierarchical interpolation blocks, enabling the model to learn and reconstruct signals at multiple temporal resolutions [Challu et al. 2023]. This makes it especially suited for long-range forecasts where capturing both coarse and fine-grained patterns is critical.

N-BEATS and N-BEATSx, by contrast, use fully connected residual blocks for univariate and exogenous forecasting respectively [Oreshkin et al. 2020]. While effective for short-term tasks, my experiments showed N-BEATSx struggled with generalisation over longer horizons. Even with historical and future exogenous inputs, it produced unstable forecasts and poor R^2 scores. N-HiTS consistently showed better alignment with actual inflation trends across all horizons. Its hierarchical design allowed more stable representation learning at different scales, capturing both seasonal and lagged effects essential in macroeconomic forecasting.

I implemented N-HiTS using NeuralForecast [Montero-Manso et al. 2023]—unlike the rest of the group who used Darts [Unit8 nd]—due to three key advantages found after experimentation with both: explicit separation of historical and future exogenous inputs to avoid leakage in multi-horizon tasks; built-in horizon-specific walk-forward evaluation with aligned train/validation splits, offering better suitability for economic forecasting than Darts' generic `backtest()`; and full control over architectural parameters (e.g., stacks, blocks, dropout, interpolation), enabling detailed experimentation, which Darts abstracts away.

2 METHOD / DESCRIPTION

All models, including N-HiTS, used a shared, group-standardised preprocessing pipeline based on a cleaned economic dataset spanning 1990s–2023. I co-developed and adapted this pipeline for deep learning and multi-horizon forecasting for models like N-BEATSx and N-HiTS. The implementation and evaluation of the N-HiTS model for forecasting U.S. monthly inflation (PCEPI) across 1-, 3-, 6-, and 12-month horizons followed a walk-forward forecasting strategy. This approach was designed to align with the group's agreed-upon project specifications for fair and reproducible model comparison. N-HiTS was selected after comparative experimentation due to its superior stability and generalisation in long-range forecasts. For N-HiTS, I refined the standardised pipeline to align with NeuralForecast's input structure with engineered features (Time-based variables (month, year), Fourier terms (seasonality), Lagged inflation values, Momentum and rolling statistics (mean, std)). Transformations such as log-scaling of the target and MinMax scaling were applied to stabilise variance and improve convergence.

After thorough testing and iteration with different architectural variants—including deeper stacks, more blocks, and alternative dropout rates—I selected the following configuration based on validation performance and generalisation stability: **Structure:** 3 stacks, each with 2 blocks; each block used MLPs with [128, 128] units; **Inputs:** Historical PCEPI + selected exogenous variables (Fourier, lags, momentum); **Scaling:** RobustScaler for inputs, global MinMax for target variable; **Loss and Optimiser:** SMAPE loss with Adam optimiser (lr = 3×10^{-4}), dropout 0.3; **Validation:** Early stopping based on rolling window.

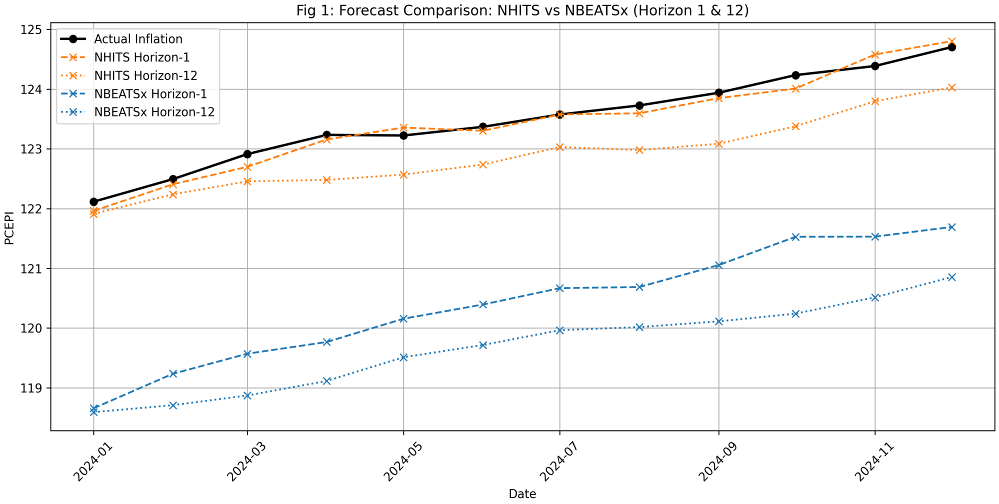
The model was trained independently for each forecast horizon (1, 3, 6, and 12 months) using a sliding window over the training period. Final predictions were made for each month of the 2024 test period. The forecasts were inverse-transformed to the original PCEPI scale and stored by model-horizon combination. Evaluation focused on three main metrics: RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) and R^2 (Coefficient of Determination). Model predictions were visualised alongside actual inflation values for qualitative comparison. Results were benchmarked against internal group baselines, demonstrating N-HiTS's superiority in capturing temporal dynamics and long-range structure compared to all the other models done by the group.

2.1 Results

I iteratively refined both N-HiTS and N-BEATSx throughout this project. My early experiments with N-BEATSx faced several issues—date misalignment, dropout instability, and overfitting due to high learning rates. These were mitigated through lower learning rates, validated feature scaling, and restructured exogenous handling. However, even after tuning, N-BEATSx yielded poor R^2 scores across all horizons and proved unstable for long-range forecasting.

In contrast, building N-HiTS involved extensive trial and error. Though theoretically well-suited for multi-horizon forecasting, initial runs produced flat or erratic outputs due to poor scaling and weak exogenous structure. I tested various input sizes, loss functions,

Author's address: Ayesha Rahman, sc21ar@leeds.ac.uk, University of Leeds, Woodhouse, Leeds, West Yorkshire, United Kingdom, LS2 9BW.



and model depths. Early setups (e.g., 4 stacks × 3 blocks) overfit, while insufficient features led to underfitting. Eventually, I settled on 3 stacks of 2 blocks for better generalisation.

To stabilise training, I refined the preprocessing pipeline by experimenting with Fourier terms ($k = 1$ to 6), log-transformed targets, lag combinations, and rolling features. Multiple runs failed due to shape mismatches between `hist_exog_list` and `futr_exog_list`, or due to inconsistent scaling across time windows. These were resolved by ensuring aligned timestamps, removing NaNs post-scaling, and debugging the walk-forward forecasting logic to maintain sequence continuity. I also compared different scaling and loss options: `scaler_type="robust"` outperformed `"identity"`, and SMAPE loss proved more robust than MAE during inflation spikes. Dropout was fixed at 0.3, as higher values led to underfitting.

Table 1. Evaluation Metrics by Model and Forecast Horizon

Horizon	**N-HiTS**			**N-BEATSx**		
	RMSE	MAE	R^2	RMSE	MAE	R^2
1	0.138	0.122	0.965	3.090	3.081	-16.717
3	0.242	0.197	0.891	3.229	3.203	-18.346
6	0.409	0.347	0.690	3.354	3.333	-19.873
12	0.635	0.602	0.253	3.812	3.808	-25.957

I retrained N-HiTS for each forecast horizon (1, 3, 6, 12 months) using a sliding window: at each step, I concatenated historical and partial test data, ran H-step-ahead predictions, and inverse-transformed the scaled and log values. Inputs were scaled with `StandardScaler`, and the log-transformed target with `MinMaxScaler`. After scaling the inputs with ‘StandardScaler‘ and the log-transformed target with ‘MinMaxScaler‘, I trained N-HiTS independently for each horizon getting optimal results. The final model configuration was: `input_size=96` (8-year lookback window), 3 stacks of 2 blocks, MLP units of [128, 128], dropout prob = 0.3, and `SMAPE()` loss.

As seen in Table 1, N-HiTS outperformed N-BEATSx across all horizons, delivering higher R^2 and significantly lower RMSE/MAE. Horizon 12 results still retained signal integrity, validating the model's generalisation capacity for long-term inflation forecasting. All forecasts were saved per horizon and visually compared to actual inflation for consistency checks. These results were shared with the team and confirmed N-HiTS as the most resilient and interpretable deep learning model for our inflation use case.

3 SELF REFLECTION

This project has been one of the most rewarding experiences, allowing me to apply deep learning and statistical forecasting to a real-world macroeconomic task—predicting US inflation (PCEPI)—in a collaborative setting. A key challenge was balancing model complexity with limited data. I learned to design robust walk-forward validation strategies, use log-transformed targets, and incorporate standardised Fourier features to boost model performance. Working with George on the preprocessing pipeline strengthened my skills in reproducibility, modular design, and collaborative debugging.

Choosing NeuralForecast over Darts provided greater architectural control and deepened my understanding of PyTorch-based model development. It also highlighted the importance of tooling and model customisation in multi-horizon forecasting. Overall, the experience boosted my confidence in leading technical components and prepared me to tackle forecasting problems in both academic and applied settings.

REFERENCES

Alejandro Challu, Jens Lagergren, Anthony Bagnall, Slawek Smyl, and Pablo Montero-Manso. 2023. NHITS: Neural Hierarchical Interpolation for Time Series Forecasting. *International Conference on Learning Representations (ICLR)* (2023).

Pablo Montero-Manso, Alejandro Challu, and Slawek Smyl. 2023. NeuralForecast: Deep Learning for Time Series Forecasting. https://github.com/Nixtla/neuralforecast. Available at: https://nixtla.github.io/neuralforecast/ (Accessed: 9 January 2025).

Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. 2020. N-BEATS: Neural Basis Expansion Analysis for Time Series Forecasting. *International Conference on Learning Representations (ICLR)* (2020).

Ayesha Rahman, George Edward Bignall, Muhammed Murat Kurmaz, James Zhangly, Kevin Raffaelli, Sandra Guran, and Natalie Leung. 2025. COMP5530M Group Project: Inflation Forecasting. https://github.com/sc20geb/COMP5530M-Group-Project-Inflation-Forecasting. Accessed: 9 May 2025.

Unit8. n.d.. NHITS — Darts documentation. https://unit8co.github.io/darts/generated_api/darts.models.forecasting.nhits.html. Accessed: 9 January 2025.

A HOW ETHICAL ISSUES ARE ADDRESSED

Due to our consistent usage of publicly available economic data and no sensitive information, no ethical concerns were identified.