Erasmus University Rotterdam

Erasmus School of Economics

Seminar Case Studies in Applied Econometrics

Nowcasting unemployment rate, private consumption and GDP one period ahead for Eurozone and US.

Supervisor:	Alberto Quaini
Date final version:	July 13, 2025

Abstract: In an era of heightened economic uncertainty and frequent financial shocks, accurately predicting macroeconomic indicators such as gross domestic product, private consumption, and unemployment, is of paramount importance to institutions of all types. This study explores whether ARIMAX, MIDAS, LSTM, ARIMAX-LSTM and MIDAS-LSTM models can effectively predict these indicators one period ahead for the Eurozone and US. The ARIMAX and LSTM rely on aggregated data, while MIDAS allows high frequency variables to be incorporated directly into the model estimation process. This enables further study of whether the inclusion of granual information improves model prediction power. Furthermore, hybrid models are implemented to leverage both linear and nonlinear patterns in series to enhance the predictive power by utilizing all available information in the data. The models are evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) accuracy metrics, while statistical significance of predictive differences across models is assessed with the Diebold-Mariano test. Upon evaluating accuracy metrics, we found that both the ARIMAX-LSTM and MIDAS-LSTM hybrid models outperform their respective benchmark models. However, the Diebold-Mariano test does not provide sufficient evidence to confirm a significant difference in predictive accuracy.

1 Introduction

In the aftermath of the COVID-19 pandemic, there has been an enormous interest in closely tracking the business cycle to assess the economic impact of measures imposed by the government. The unprecedented nature of the crisis revealed the importance of timely economic predictions, enabling policymakers and financial institutions to make informed decisions in response to the rapidly changing economic conditions.

Business cycle theory suggests that economies progress through four stages: expansion, peak, contraction, and trough. By accurately identifying these stages in real-time, large financial institutions are able to anticipate downturns and adjust their strategies accordingly. For instance, firms can restructure investment portfolios or provide clients with risk assessments for their investment strategies. Additionally, during times of crisis, the ability to predict economic recovery ahead of competitors can yield strategic advantages, for example by investing in high-growth sectors or regions at lower costs before broader market rebounds occur.

Macroeconomic indicators such as the unemployment rate, private consumption, and GDP serve as important signals for identifying the economy's current phase. However, forecasting these indicators remains a challenge. Furthermore, official macroeconomic data of these indicators often suffer from reporting lags, sometimes taking more than a quarter to become available. As a result, policymakers are frequently required to make decisions without the most recent data and are interested in the nowcasted values of these macroeconomic indicators.

Traditional economic forecasting models often rely on theory-driven methodologies and variables(Mullainathan and Spiess, 2017). These models require economic intuition and expert judgment to determine the most relevant data and methodologies. However, errors in these assumptions can lead to inaccurate predictions. These traditional models are primarily used to capture linear relationships between different macroeconomic variables and autoregressive relationships of the macroeconomics indicators themselves. An example of such a model is the Autoregressive Integrated Moving Average (ARIMA) model, which focuses on capturing the linear autoregressive structure. These models can be extended by incorporating exogenous variables, forming the ARIMAX model, which introduces additional explanatory power beyond purely endogenous time-series relationships.

In contrast, machine learning models primarily focus on pure prediction (Varian, 2014). Unlike traditional econometric models, machine learning approaches exhibit greater flexibility, allowing them to identify complex nonlinear relationships without relying on pre-specified assumptions or potentially subjective judgments. This characteristic has contributed to the growing adoption of machine learning techniques for forecasting macroeconomic indicators. One particularly promising approach within this domain is the use of artificial neural networks (ANNs), specifically Long Short-Term Memory (LSTM) networks, which are well capable of handling time-series data due to their ability to capture long-range dependencies and nonlinear patterns in sequential data.

Traditional macroeconomic forecasting heavily depends on low-frequency data, which aligns with the reporting intervals of macroeconomic indicators (such as quarterly GDP releases or monthly employment reports). However, the availability of high-frequency data has increased in recent years, presenting new opportunities for improving forecasting accuracy. One common approach to integrating high-frequency data into macroeconomic forecasting is through aggregation methods that convert high-frequency variables to match the lower-frequency data. The most straightforward techniques, such as mean aggregation and sum aggregation (hereafter referred to as "simple aggregation methods"), offer intuitive solutions but may overlook underlying structural relationships within the data. A more sophisticated alternative is the Mixed Data Sampling (MIDAS) approach, which tries to optimally combine high-frequency data with low-frequency data based on the relationships within the data into a regression. The overarching research goal of this paper is:

To develop the most accurate possible nowcasting predictions for changes in GDP, unemployment, and personal consumption for the US and Eurozone, using all available data at the end of each period. Importantly, this paper does not aim to address the interpretability or explainability of the models used; rather, its primary focus is predictive accuracy.

This global research goal is divided into the following key research questions:

- Does the MIDAS regression model outperform the ARIMAX model (with simple aggregation techniques)?
- Does a hybrid ARIMAX-LSTM model outperform its individual components (ARIMAX and LSTM) in forecasting macroeconomic indicators?
- Does a hybrid MIDAS-LSTM model outperform its individual components (MIDAS and LSTM) in forecasting macroeconomic indicators?
- Does a hybrid MIDAS-LSTM model outperform a hybrid ARIMAX-LSTM model?

These questions will be examined across all different scenarios within the paper. In other words, the paper investigates each scenario created by the different research questions. For example the study compares the predictive performance of the ARIMAX-LSTM hybrid model with its individual components for both the US and Eurozone economies and for each of the three target variables.

Our study requires a comprehensive set of macroeconomic and related variables that may correlate with the business cycle. To adequately test different (dis)aggregation methods, we require variables with varying time frequencies. Van Lanschot Kempen provided us with a dataset that meets these conditions. In order to utilize different frequencies of data, we employ simple (dis)aggregation methods (e.g. mean or sum) based on the type of variable.

The MIDAS regression is able to aggregate a higher frequency variable into a regression (e.g. incorporate monthly variables into a quarterly regression). However, the MIDAS model is not able to perform disaggregation or aggregate two different frequencies simultaneously (e.g. daily and monthly into a quarterly regression simultaneously). We therefore perform simple (dis)aggregation in scenarios that cannot be handled by our MIDAS model to make fair comparison between the ARIMAX model and the MIDAS model. Given that our variables have different starting points in time, we apply backward extrapolation using the ARIMA model for the ARIMAX case and MIDAS case and the LSTM model for the LSTM and hybrid models, ensuring that all data is temporally aligned.

As aforementioned, in our nowcasting approach, we employ 4 distinct models: ARIMAX Model, MIDAS model, LSTM Model, ARIMAX-LSTM Hybrid Model and MIDAS-LSTM Hybrid Model. The hybrid models are a two-step approach in which the ARIMAX/MIDAL model first nowcasts the target variables, and an LSTM model subsequently processes the residuals to capture additional nonlinear relationships. All three target variables are nowcasted separately.

To evaluate model performance, we compare accuracy using Mean Squared Error (MSE) and Mean Absolute Error (MAE). Additionally, we employ the Diebold-Mariano test to statistically assess whether differences between model predictions are significant. Residual analysis is also conducted to evaluate the reliability of each model.

Existing literature has extensively examined both ARIMAX and LSTM models for GDP forecasting. Additionally, prior research has explored ARIMA-LSTM hybrid models (only using lagged dependent variables). Our study contributes by enhancing the traditional ARIMA-LSTM hybrid model with the inclusion of exogenous variables, introducing a ARIMAX-LSTM hybrid approach.

The MIDAS regression model has been examined before in macroeconomic forecasting scenarios, however, not in direct comparison with an ARIMAX model. Moreover, in our research, we assess the performance of a MIDAS-LSTM hybrid approach and compare it with the individual models and an ARIMAX-LSTM model. To the best of our knowledge, this particular MIDAS-LSTM hybrid combination has not been investigated in before. Lastly, our paper extends the broader literature on macroeconomic nowcasting by integrating traditional econometric methodologies with advanced machine learning techniques. As economic forecasting becomes increasingly complex, our study provides valuable insights into improving predictive accuracy and leveraging both structured and high-frequency data for better economic decision-making.

Our results show that both our ARIMAX-LSTM hybrid and MIDAS-LSTM hybrid outperfrom their indvidual benchmark models. Nevertheless, when cannot yet conclude that there is a significant difference in predictive accuracy based on the results of our Diebold-Mariano test. On the contrary, when comparing the ARIMAX model and the MIDAS model, neither model is outperforming. Similarly, by comparing the ARIMAX-LSTM to the MIDAS-LSTM, the MIDAS-LSTM shows better predictive accuracy for five of the six target variables and no difference in predictive accuracy based on the Diebold-Mariano test.

The remainder of our paper is organized follows: Section 2 reviews existing literature on nowcasting/forecasting macroeconomic indicators. Subsequently, we describe our data and the corresponding data cleaning process in Section 3. In Section 4 we outline our estimation and evaluation methods required for the analysis. Section 5 presents and discusses our results, and Section 7 summarizes our main findings and draws a conclusion.

2 Literature Review

The forecasting and nowcasting of macroeconomic indicators (such as GDP, unemployment, and personal consumption) has long been an area of interest for various financial institutions and. Traditional time series prediction models, particularly ARIMA (Box George et al., 1976), have served as foundational tools for macroeconomic nowcasting due to their ability to capture autoregressive and moving average components. The ARIMAX model extends ARIMA by incorporating exogenous variables and therefore enhances its power in structured economic predictions (Peter and Silvia, 2012). However, these models are often limited to datasets where all variables share the same frequency, restricting their applicability when integrating high-frequency variables into a dataset.

To address this, researchers have introduced Mixed Data Sampling (MIDAS), a method proposed Ghysels et al. (2004) and Ghysels et al. (2007). MIDAS can be regarded as a time-series regression tool that allows the target variable and exogenous variable to be sampled at different frequencies, without doing any aggregation, preserving critical information. Several studies have applied MIDAS for macroeconomic time series, demonstrating its advantages over standard time-series models (Clements and Galvão, 2008), (Marcellino and Schumacher, 2007) (Wohlrabe, 2009).

Despite these advancements, both ARIMAX and MIDAS remain fundamentally linear models by construction, which limits their ability to capture the nonlinear relationships in macroeconomic indicators (Zhang, 2003). Recent studies have highlighted the increasing role of artificial neural networks in overcoming this limitation (Tan et al., 2022). Specifically, Long Short-Term Memory (LSTM) networks, a subclass of recurrent neural networks, have been found to excel at modeling long-term dependencies and nonlinear interactions (SANUSI et al., 2020). Unlike ARIMA-based approaches, LSTM models can integrate high-dimensional datasets and adapt to structural breaks, making them more adaptable to volatile economic conditions (Taslim and Murwantara, 2024).

Building on these developments, research has explored hybrid models that combine traditional techniques with machine learning. The ARIMA-LSTM hybrid has gained attention for combining ARIMA's structured forecasting with LSTM's ability to model complex, nonlinear dependencies (Hamiane et al., 2024). However, limited studies (and none for macroeconomic forecasting) have examined an ARIMAX-LSTM hybrid, which integrates exogenous variables into this hybrid approach. Similarly, while MIDAS has been successfully applied in macroeconomic forecasting, there is little research directly comparing MIDAS with ARIMAX or integrating MIDAS with LSTM. Our study aims to fill these gaps by assessing the MIDAS-LSTM hybrid model in direct comparison with ARIMAX-LSTM while simultaneously comparing their individual counterparts.

3 Data

This section starts by introducing the target variables and explanatory variable sets for the EU and US. Key problems with the datasets are then discussed, followed by the implemented data cleaning procedures. Data extrapolation methods used in this paper are discussed in detail, along with a custom framework for measurement frequency conversion, more specifically custom disaggregation and aggregation methods. Finally, a detailed overview is provided of the characteristics of the input datasets used to train each model, followed by a simple approach to remove the distortions of Covid from the data.

The two economic regions of interest are the Eurozone (EU) and the United States of America (US). The US data are sourced from the Federal Reserve Database and the Bureau of Transportation Statistics, while the EU data are sourced from EuroStat. The target variables are sourced from the Bureau of Economic Analysis: US Department of Commerce for the US, and the European Central Bank for the EU. EU data is available from 1 January 2000 to 31 September 2024, while US data is available from several earlier dates to 31 September 2024. In order to achieve cross-region comparability between the chosen models, a decision is made to also start the US dataset on 1 January 2000.

3.1 Target Variables, Raw Independent Variable Sets and Data Cleaning

Table 1 provides an overview of the frequencies, full names, and abbreviations of the target variables, per region. All targets are measured as percentage change per period.

Full Name	Abbreviation	Frequency	
		USA	EU
Gross Domestic Product	GDP	Quarterly	Quarterly
Personal Consumption Expenditure	PCE	Monthly	Monthly
Unemployment Rate	UNEMP	Monthly	Quarterly

Table 1: Overview of Target Variables

The EU and USA datasets were composed by selecting specific variables related to a set of macroeconomic headings most closely related to the GDP, PCE and UNEMP target variables. These headings are namely: monetary policy signals, fiscal policy signals, consumer behaviour, cross-border trade, population movements, foreign investment and balance of payments (Conrad, 2022). The raw data contained a number of flaws, namely: included weekend and public holiday values for some daily variables (1); quarterly variables expressed as monthly variables by introducing null entries (2); duplicate variables (3); null variables (4); discontinued variables (5); Composite variables included with their separate elements-linear combinations of other variables (6); levels and first-differenced forms of the same variables (7) and differing measurement starting times per variable.

Discontinued variables (variables that are no longer reported) were deleted from the initial dataset as they were deemed useless in the future despite any predictive power they might have had in the past. Daily variables were filtered to remove any weekend and public holiday dates included in the dataset. Incorrectly formatted monthly variables were converted to the correct quarterly format and null or duplicate variables were deleted. Composite or linearly combined variables were kept while their elements were deleted. Level forms of variables were deleted as first differenced forms require less manipulation to achieve stationarity and align with the 'change per period' format of the target variables. Table 2 provides an overview of variable counts for both regions and all frequencies after data cleaning. Differing variable measurement start times is addressed in the coming subsection.

	Daily		Monthly		Quarterly	
	USA	EU	USA	EU	USA	EU
Count	20	7	122	29	31	39

Table 2: Total Independent Variable Count by Frequency and Region

3.2 Extrapolation

Differing variable measurement start times (8), is solved with 2 methods of model-based data extrapolation. Incomplete variables with more than 50 percent missing entries are deleted while remaining incomplete variables are extrapolated. Table 3 provides an overview of the incomplete variables per region, per frequency. The 'Max %' column refers to the largest percentage of missing values encountered for any given variable within the incomplete set, per region, per frequency. A convention of 'doubling down' is adopted so as to match the choice of extrapolation model with the eventual model for which the dataset is used as input. As mentioned before, linear, non-linear and hybrid combination models are used in this paper. The linear ARIMAX and MIDAS models therefore use data extrapolated by a suitable linear model, while the non-linear LSTM, hybrid ARIMAX-LSTM and hybrid MIDAS-LSTM models use data extrapolated by a non-linear model.

Region	Daily		Monthly		Quarterly	
	Incomplete	Max %	Incomplete	Max %	Incomplete	Max %
USA	3	25	0	0	0	0
EU	1	20	6	9	5	20

Table 3: Total Number of Incomplete Variables and Maximum % Extrapolated for USA and EU by Frequency - from 1 January 2000 to 31 September 2024

A potential linear approach for extrapolation would be to use the incomplete variable as a dependent variable and a selection of other complete variables from the dataset as independent variables in an ARIMAX regression. However, as introducing linear relationships between explanatory variables in the dataset will likely lead to a multicollinearity problem, a standard linear ARIMA model is chosen for extrapolation instead, providing adequate coverage of autoregressive, moving average and integrated dynamics while avoiding the aforementioned multicollinearity problem. This choice will be further discussed in Section 4.2.

The nonlinear LSTM and hybrid model dataset is completed by using a LSTM neural network for extrapolation. In this approach a LSTM neural network is trained with a subset of complete variables as an independent input dataset and a single incomplete variable as dependent. The model is trained on the time interval for which the incomplete variable has values. The trained network and complete variables are then used to predict the missing values of the incomplete variable. This is further discussed in Section 4.4. See Appendix 7.1 for a detailed algebraic explanation of the extrapolation process.

3.3 Frequency Conversion Processes

In order to model relationships between independent and dependent variables of different measurement frequencies, accurate data aggregation or disaggregation methods are required. A commonly used aggregation approach is mean aggregation: simply taking the average over an interval, while the equivalent disaggregation procedure is to divide a data point into equal values depending on the desired output frequency. We instead inspect the data and derive a custom tailored approach to achieve realistic disaggregation and aggregation results depending on characteristics of specific variables.

The custom categorization criteria together with an illustrative example are as follows: Cumulative quantities or summed values- 'Government spending' (1); Percentage change per period - 'US wealth' (2); Up or down-trending - 'Total loans outstanding' (3); Constant for extended periods - 'Bank deposit interest rates' (4); Slow moving positive values - 'Index of any sort' (5); Changes in large values - 'Actual change in loans outstanding' (6); Uniquely measured per period actual value - 'Number of Rail Passengers' (7). Table 4 illustrates the chosen solution for aggregation and disaggregation in the aforementioned categories.

Frequency Conversion	(Q to M)	(D to M)/(D to Q)/(M to Q)
Category	Disaggregation	Aggregation
1	Division	Summation
2	Formula (D)	Formula (A)
3	Copy & Fill	Average
4	Copy & Fill	Newest Observation
5	Copy & Fill	Average
6	Division	Summation
7	Division	Summation

Table 4: Aggregation & Disaggregation Solutions by Category - The 'Frequency Conversion' row refers to the specific conversions necessary in this paper with D - Daily, M - Monthly and Q - Quarterly. Formula (A) and (D) are formulas treated at the end of Section 3.3.

The custom (percentage) category 2 formulas, (D) and (A) in Table 4 are defined as:

$$\text{D: } a(x_i) = \mathbf{1_3}[(\sqrt[3]{\frac{x_i}{100}+1})-1]\times 100, \text{ and A: } d(\boldsymbol{x_i}) = [[\prod_j(\frac{x_{ij}}{100}+1)]-1]\times 100, \tag{1}$$

where i refers to the observation number, a takes a scalar x_i as input and produces a 3-dimensional output vector, and d takes a vector $\boldsymbol{x_i}$ input and produces a scalar output. $\mathbf{1_3}$ is a vector of 1's. The specific use of formula (D) in this case is to convert frequencies from Quarterly to Monthly; the 3rd root and 3 dimensional output vector is therefore intuitively explained by the presence of 3 months in a quarter.

3.4 Dataset Preparation Overview

In Section 4 variable stationarity, normalization and selection is handled where applicable as part of the estimation processes of the models.

Model	Aggregation	Disaggregation	Extrapolation
ARIMAX	Custom	Custom	ARIMA
MIDAS	MIDAS	Custom	ARIMA
LSTM	Custom	Custom	LSTM
ARIMAX-LSTM	Custom	Custom	LSTM
MIDAS-LSTM	MIDAS & Custom	Custom	LSTM

Table 5: (Dis)Aggregation and Extrapolation Method Overview, by Model - Custom refers to the methods treated in Section 3.3, MIDAS-LSTM uses custom (dis)aggregation in the LSTM input dataset.

Finally, in order to prepare for estimation the cleaned EU and US datasets are processed according to the requirements of each model as provided by Table 5. The specifics of the (dis)aggregation method applied to each so called 'estimation ready' dataset is dependent on the frequency of the specific target in question. Additionally, it can also be seen in Table 5 that all models use the same custom disaggregation methods. MIDAS as applied in this paper is only being used for conversion from Monthly to Quarterly or Daily to Monthly - in cases where Daily to Quarterly is required, custom aggregation is first used to produce Monthly variables. This is discussed in more detail in 4.3. The influence of Covid on model performance is also treated by deleting the period of 1 Jan 2020 to 30 June 2022 from the 'estimation' datasets before repeating the estimation process.

4 Methodologies

This study follows a structured framework, beginning with variable selection, progressing through the implementation of linear, nonlinear and hybrid models, and concluding with performance evaluation. To identify the most relevant predictors while addressing multicollinearity, Elastic Net regression is applied (Section 4.1). This approach filters the set of exogenous variables used for our ARIMAX and MIDAS model. For our Long Short-Term Memory model we use our complete dataset. After fitting these three individual models, we additionally build our hybrid models. For the ARIMAX-LSTM hybrid model, the residuals from the ARIMAX model are used as inputs for the LSTM, attempting to capture remaining nonlinear relationship. Similarly, the residuals from our MIDAS model are used as inputs for the LSTM to obtain the MIDAS-LSTM hybrid. Note that for both hybrid models, we use the Elastic Net reduced dataset for the linear model and the full dataset for the LSTM. Model adequacy for the linear prediction models is validated through residual analysis (Section 4.7), while forecast accuracy is assessed using Mean

Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Diebold-Mariano (DM) test (Section 4.8). To provide a comprehensive understanding of each methodological component, the following sections delve into the details of the methods applied in this study.

4.1 Elastic Net

To ensure that only the most relevant exogenous variables were included in the linear models, we applied the Elastic Net regularization method (Zou and Hastie, 2005). Since both Elastic Net assumes a linear relationship, this method fits well with our linear modeling approach by maintaining interpretability while effectively handling multicollinearity and high-dimensional datasets by shrinking coefficients of less important variables toward zero.

The objective function for Elastic Net combines Lasso (L1) and Ridge (L2) regression penalties and is given by:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \left[\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right] \right\}. \tag{2}$$

In this formula, y_i is the target variable whereas X_is are exogenous variables and the first term, $\sum_{i=1}^{n}(y_i - X_i\beta)^2$, gives the residual sum of squares, making sure that the model minimizes prediction error. The second term is the regularization component controlled by λ, where α defines the weighting between Lasso and Ridge penalties.

The Lasso component, $\sum_{j=1}^{p} |\beta_j|$, applies L1-norm regularization, which shrinks some coefficients to exactly zero, effectively performing variable selection. The Ridge component, $\sum_{j=1}^{p} \beta_j^2$, applies L2-norm regularization, which prevents overfitting by penalizing large coefficient values. The combination of these two penalties allows Elastic Net to effectively handle multicollinearity and high-dimensional datasets, resulting in the selection of only the most relevant predictors while preserving the stability of the model. The optimal values for λ and α were selected using cross-validation, ensuring a balance between model complexity and predictive performance. Variables with non-zero coefficients were chosen to build the model.

4.2 ARIMAX

The AutoRegressive Integrated Moving Average with Exogenous Variables (ARIMAX) model is an extension of the ARIMA model which incorporates a linear relationship between the dependent and external explanatory variables, making it a suitable model for macroeconomic forecasting. To maintain consistency in frequency alignment, all variables are transformed to match the frequency of target variables using appropriate aggregation and disaggregation methods as mentioned in the section 3.3.

Note for extrapolating the variables, we opted for an ARIMA model instead of ARIMAX to maintain a linear relationship between variables and avoid the multicollinearity issue as is described in Section 3.2. Since ARIMAX introduces exogenous variables into the regression, ARIMAX extrapolated predictors could lead to redundancy and high correlation among predictors, making it difficult to separate their individual effects. By using ARIMA for extrapolation, we capture the autoregressive, moving average, and integrated components of the time series while making sure that it does not distort relationships between variables and keeps the integrity of our regression framework and allows for a more stable ARIMAX model for forecasting.

4.2.1 Model Compenents

The Autoregressive Integrated Moving Average (ARIMA) model is a widely used statistical technique for time series forecasting and consists of the following 3 key components:

Autoregressive (AR) Component: The Autoregressive (AR) model uses past values of the target variable to predict future values. The AR process of order p, denoted as AR(p), is mathematically expressed as:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t. \tag{3}$$

Moving Average (MA) Component: The Moving Average (MA) component captures the dependency on past forecasting errors. The MA process of order q, denoted as MA(q), is given by:

$$y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}. \tag{4}$$

Integrated (I) Component: If the data is non-stationary, it needs to be differenced to achieve stationarity. The number of differencing required is denoted as d. The differenced series is represented as:

$$\Delta^d y_t = y_t - y_{t-1}, \tag{5}$$

where Δ^d represents the differencing operation applied d times.

ARIMA Model and Exogenous Variables: The combination of AR and MA models forms the Autoregressive Moving Average (ARMA) model. If the data is stationary, ARMA(p, q) can be directly applied. However, if the data is non-stationary, the ARIMA(p, d, q) model is used by applying differencing. The AutoRegressive Integrated Moving Average with Exogenous Variables (ARIMAX) model is an extension of ARIMA that includes external explanatory variables. These exogenous variables influence the dependent variable, providing additional information that can improve forecasting accuracy. Note that these exogenous variables need to be stationary as well. The general equation is:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \sum_{k=1}^{m} \beta_k x_{k,t} + \epsilon_t, \tag{6}$$

where y_t represents the target variable, such as GDP, unemployment rate, or private consumption, while $x_{k,t}$ denotes the exogenous variables that influence y_t. The autoregressive coefficients are represented by ϕ_i, whereas moving average coefficients are given by θ_j. Furthermore, the term β_k shows the impact of each exogenous variable x_k on the dependent variable. Finally, ϵ_t is the error term, gives an unexplained variance in the model.

4.2.2 Stationarity

Stationarity is a crucial concept in time series modeling, it makes sure that statistical properties of a series such as mean, variance, and relationships between observations remain stable over time. A non-stationary series often exhibits trends, seasonality, or structural shifts, leading to spurious relationships and unreliable regression estimates. To make our models reliable ARIMAX model, we tested the stationarity of all exogenous variables using the Augmented Dickey-Fuller (ADF) test (Cheung and Lai, 1995), which checks for the presence of a unit root. If the test indicated non-stationarity, appropriate differencing is applied to remove trends. Furthermore, the lag structure of exogenous variables was adjusted if necessary to align with the stationary properties of the dependent variable. By checking the stationarity of all exogenous variables before including them in the model, we prevented spurious correlations and improved the reliability of the ARIMAX model.

4.2.3 Model Fitting

The ARIMAX model is fitted using a two-stage process, combining regression modeling with time series dynamics to enhance predictive accuracy. In the first stage, a multiple linear regression is performed to capture the relationship between the dependent variable and the selected exogenous predictors. This ensures that significant explanatory variables are identified before incorporating time-dependent structures. The residuals from this regression are then extracted to analyze any remaining autocorrelation, which is addressed in the second stage using the Box-Jenkins methodology (Box George et al. (1976).

The Box-Jenkins method involves three key steps: model identification, parameter estimation, and diagnostic checking. Initially, the appropriate orders of autoregressive (AR) and moving average (MA) components are determined based on autocorrelation and partial autocorrelation function (ACF and PACF) plots. Next, parameter estimation is carried out using maximum likelihood estimation or other optimization techniques to refine the model fit and finally optimal (p,d,q) are chosen based on the AIC of the models. Once the model is built, residual analysis is performed to check whether the model adequately consumes the underlying data structure.

By integrating regression with ARIMA-based modeling, ARIMAX effectively handles both exogenous influences and time dependence structure, making it one of the powerful tool for forecasting macroeconomic indicators.

4.3 MIDAS

In traditional time series forecasting, aligning variables of different frequencies often requires aggregation, which can result in the loss of valuable high-frequency information. In the ARIMAX model, all variables are converted to the same frequency before modeling, potentially discarding meaningful variations present in the original high-frequency data. To overcome this limitation, the Mixed Data Sampling (MIDAS) regression model (Ghysels et al., 2020) is employed, allowing high-frequency variables (e.g., monthly indicators) to be directly incorporated into forecasting low-frequency targets (e.g., quarterly GDP). By avoiding unnecessary aggregation, MIDAS preserves the granularity of information, enhancing predictive accuracy, and capturing short-term dynamics that traditional models might overlook.

4.3.1 Model Specification

MIDAS regression can be expressed as follows:

$$y_t = \alpha + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=0}^{k} \sum_{j=0}^{l_i} \beta_j^{(i)} X_{tm_i-j}^{(i)} + \epsilon_t, \tag{7}$$

where y_t is the low-frequency target variable which is influenced by both its own past values having autoregressive coefficient ϕ_i and high-frequency explanatory variables which are represented as $X_{tm_i-j}^{(i)}$, where m_i is the frequency ratio, which represents the number of times the high-frequency variable $X^{(i)}$ is observed within the lower-frequency period of y_t, l_i is the maximum lag length applied to the high-frequency variable $X^{(i)}$ and $tm_i - j$ represents the specific lagged observation of the high-frequency variable.

Since including multiple high-frequency lags can result in a large number of parameters, MIDAS applies parametric constraints (e.g., Nealmon polynomial, Beta lag function) to control model complexity and improve efficiency.

4.3.2 Frequency Alignment in MIDAS

To effectively integrate high-frequency variables into a low-frequency model, a process known as frequency alignment is required. This transformation ensures that high-frequency predictors are structured into a format compatible with low-frequency observations.

For example, when using monthly data to explain quarterly variables, the frequency ratio is 3 (since each quarter contains three months). If we assume that both the current and previous quarter's monthly data affect GDP, the model incorporates the following terms:

$$y_t = \phi_1 y_{t-1} + \sum_{i=0}^{k} \sum_{j=0}^{5} \beta_j^{(i)} X_{tm_i-j}^{(i)} + \epsilon_t. \tag{8}$$

Similarly, a daily variable (Z_t) is added to the model with the monthly target, each month consists of 21 business days, hence considering the frequency ratio of 21 days per month

$$y_t = \phi_1 y_{t-1} + \sum_{j=0}^{20} \alpha_j Z_{tm-j} + \epsilon_t, \tag{9}$$

where α_j is coefficient which defines the impact of lag Z_{tm-j}.

Note that in this study, while fitting MIDAS regression, for quarterly target variable, the high-frequency data used is monthly. If daily data is available, it is first aggregated to a monthly frequency using simple aggregation before adding it to the model.

4.3.3 Weighting Constraints in MIDAS

To reduce the number of parameters and avoid overfitting, MIDAS applies parametric constraints on the coefficients $\beta_j^{(i)}$. Instead of estimating each coefficient independently, a weighting function is imposed:

$$\beta_j^{(i)} = f(\gamma^{(i)}, j), \quad j = 0, 1, ..., l_i \tag{10}$$

where $\gamma^{(i)}$ is a set of estimated parameters controlling the weight structure.

In this study, Nealmon polynomial function has been implemented, which ensures that lag weights decrease smoothly while maintaining numerical stability by using an exponential structure. The function is given by:

$$\beta_j = \frac{\exp(\gamma_1 j + \gamma_2 j^2 + ... + \gamma_p j^p)}{\sum_{j=0}^{l} \exp(\gamma_1 j + \gamma_2 j^2 + ... + \gamma_p j^p)}, \tag{11}$$

where The parameters $\gamma_1, \gamma_2, \ldots, \gamma_p$ are estimated to determine the optimal lag structure and denominator normalizes the weights, preventing instability.

In implementing this weighting structure, it is crucial to specify appropriate starting values for the optimization process. Poorly chosen starting values can lead to divergence issues, resulting in non-finite intercept estimates and failed optimization. Hence, to achieve reliable parameter estimation and improve stability and convergence, suitable initial values need to be selected carefully for high-frequency variables in each model.

4.3.4 Estimation of MIDAS Model

The MIDAS model is estimated using Nonlinear Least Squares (NLS), which minimizes the sum of squared residuals while handling the nonlinear weighting constraints efficiently:

$$\hat{\gamma} = \arg\min_{\gamma} \sum_{t=1}^{n} \left(y_t - \alpha - \sum_{i=1}^{k} \sum_{j=0}^{l_i} f(\gamma^{(i)}, j) X_{t,m_i-j}^{(i)} \right)^2 . \tag{12}$$

Once the MIDAS model parameters are estimated using Nonlinear Least Squares (NLS), the model's performance is validated through various metrics and statistical tests. The residual analysis is conducted to assess whether the model adequately captures the data patterns. By integrating high-frequency data through MIDAS, this study allows to include more granular information without losing valuable data to aggregation techniques while maintaining a parsimonious parameterization.

4.4 Long Short-Term Memory Network (LSTM)

Long Short-Term Memory Networks (LSTM) are neural networks that excel in capturing both long-term and short-term dependencies in time-series data and therefore excellent for capturing the non-linear relationships between the target variables and the exogenous variables. For both the LSTM prediction model and the ARIMAX-LSTM hybrid prediction, we assume nonlinear relations between the variables (as the LSTM is able to capture all types of non-linear relationships). We therefore decide to 'double down' on this regard and use LSTM to extrapolate missing values in the data, as mentioned in Section 3.2.

Regarding the extrapolation, we only use exogenous variables for the LSTM which did not have missing values in the original data set. In other words, we do not add already extrapolated variables to the exogenous data set for the LSTM. This recursive extrapolation approach introduces a cumulative error, which means that as the recursion proceeds, the prediction error of the previous time step will propagate over the subsequent variables (Xu et al., 2022).

Note the difference between the extrapolation of the data variables and the nowcasting prediction of the target variables. The extrapolation concerns missing datapoints; the LSTM

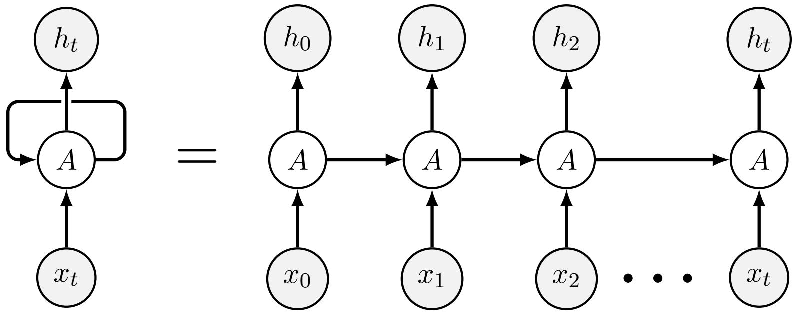

Figure 1: A Recurrent neural work unrolled.

model will be trained once on all available data before extrapolating all missing entries. In our nowcasting scenario for the macroeconomic indicators, our LSTM is retrained after each increment, as the actual values of the target variables become known.

4.5 Theory

Long Short-Term Memory networks (LSTMs) are a specialized type of recurrent neural network (RNN).

The major difference between artificial neural networks (ANN's) and RNN's is that RNN's have the ability to learn and predict based on previous time points by introducing a so called feedback loop. This makes them especially powerful for processing time-series data, where past values have an influence future predictions. RNN's can be thought of as multiple copies of the same neural network, each passing a message to its successor. Consider what happens if we unroll the following loop: The LSTM model was introduced by Hochreiter and Schmidhuber in 1997 and has proven to be particularly effective in incorporating information from the past without under- or overestimating it (a common drawback associated with traditional RNN's). Therefore, it is capable of successfully capturing long term relationships. The major key to the architecture of LSTM are memory units, which retain information over long time periods. Each memory unit consists of mechanisms called "gates" that regulate the flow of information in and out of the unit, making sure that relevant information is retained while irrelevant data is discarded (Gers et al., 2000). Each LSTM unit consists of four interconnected neural networks, all with an input layer and an output layer. These networks perform different functions to regulate the flow of information:

- Forget Gate: Determines which information should be discarded from memory.
- Input Gate: Regulates which new information should be stored.
- Output Gate: Regulates the use of information present in memory
- Candidate Memory: Generates new candidate information to be stored in memory.

These gates allow LSTMs to perform a more efficient management of memory, making sure that critical long-term dependencies are preserved while preventing the model from being overwhelmed by irrelevant information.

An LSTM unit receives three vectors as input at each time step: Cell State (C_{t-1}) represents long-term memory. Hidden State (H_{t-1}) captures short-term dependencies. Input Vector (X_t) contains new external information.

In practice, the LSTM unit uses recent past information (the short-term memory, H) and new information coming from the outside (the input vector, X) to update the long-term memory (cell state, C). Finally, it uses the long-term memory (the cell state, C) to update the short-term memory (the hidden state, H). The hidden state determined at instant t is also the output of the LSTM unit in instant t.

All three gates are neural networks and use an activation function (often the sigmoid function) in the output layer to generate an output vector composed of values between zero and one, allowing for precise control over information flow (where a 1 represents keeping information and a 0 represents deleting information).

A figure representing the workings of a LSTM unit is presented in Figure 2. In more detail we have:

Forget Gate: The forget gate decides, based on X_t and H_{t-1}, what information to remove from the cell state vector coming from time t1. The outcome of this decision is a selector vector, which is applied to the previous cell state (C_{t-1}) to filter out irrelevant information. The candidate memory, which often uses the hyperbolic tangent function to generate a vector with values between -1 and 1, normalizing potential new information. The input gate, which determines how much of this new candidate information should be incorporated into the cell state. The final updated cell state is computed by multiplying the candidate vector with the selector vector from the input gate and adding the result to the cell state vector. Output Gate: the cell state, once updated, is used by the output gate to determine the new hidden state (H_t). The output gate selects relevant information from the updated cell state and transforms it into the hidden state using another sigmoid function. This hidden state is then passed to the next LSTM unit at time t+1.

The core training mechanisms in RNNs, including LSTMs, is Backpropagation Through Time (BPTT). Unlike standard backpropagation in feedforward neural networks, BPTT considers dependencies across multiple time steps, which introduces unique challenges. The weight updates are based on a chain of dependencies over time. However, this chain of dependencies leads to two issues: The Vanishing Gradient Problem: when gradients become very small due to repeated multiplications with small values, leading to negligible weight updates. This hinders the model's ability to learn the existing long-term dependencies.

Exploding Gradient Problem: when gradients grow exponentially due to repeated multiplications with very large values, it results in numerical instability and makes training difficult. These problems limit the effectiveness of basic RNNs for learning long-range dependencies. The vanishing gradient problem particularly restricts standard RNNs to only short-term memory behaviour, as long-range dependencies become almost impossible to learn.

LSTM's, however, mitigate the vanishing gradient problem by utilizing their gating mechanisms to control the flow of information. By carefully regulating which information is retained and which is discarded, LSTMs ensure that critical long-term dependencies are preserved during training.

While LSTMs significantly reduce the vanishing gradient problem, they are not completely immune to the exploding gradient issue. In some cases, extremely large gradients can still propagate through alternative pathways in the network. For mathematical details on this we refer to Hochreiter (1997).

The training for the LSTM is done differently for extrapolation than for the actual prediction of the target variables. For extrapolation the model is trained once on the available data before extrapolation. For the prediction of the target variables, the LSTM model is retrained for each nowcasting prediction, as the available data also increases with each of those increments. The hyperparameter tuning is done the same for both cases, for which we refer to Appendix B.

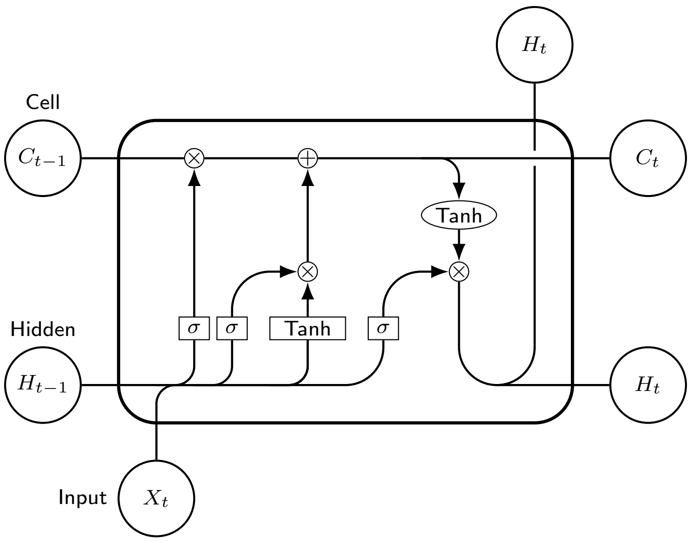

Figure 2: Illustration of an LSTM Cell.

4.6 Hybrid Models

Both Linear Models and Neural Networks have demonstrated success within their respective domains. ARIMAX and MIDAS for linear patterns and LSTM for nonlinear relationships.

However, neither serves as a universal solution suitable for all scenarios. ARIMAX and MIDAS may struggle to approximate complex nonlinear structures, while neural networks (such as LSTM) have shown mixed effectiveness when applied to linear problems and have a risk of overfitting or instability (Zhang, 2003). Moreover Markham and Rakes (1998) found that the performance of neural networks in linear regression tasks depends on factors such as sample size and noise levels.

Since the true nature of real-world data is often unknown, a hybrid methodology that integrates both linear and nonlinear modeling approaches can be a practical strategy. Thus, time series can be expressed as:

$$y_t = L_t + N_t + \epsilon_t \tag{13}$$

where L_t represents the linear structure of the data at time t, N_t captures the nonlinear component, and ϵ_t denotes the error term. Since the linear models identify the structured linear trend, the residuals are assumed to contain the remaining nonlinear features:

$$y_t - L_t = N_t + \epsilon_t \tag{14}$$

This decomposition forms the basis for the hybrid models, where linear models capture L_t and LSTM models the nonlinear component N_t.

4.6.1 ARIMAX-LSTM Hybrid

The ARIMAX-LSTM hybrid approach integrates ARIMAX for modeling linear dependencies and LSTM for capturing nonlinear residual patterns. The ARIMAX model is first fitted to the data, and its residuals are extracted. These residuals represent the unexplained variance, which is then modeled using an LSTM network. While ARIMAX focuses on a subset of exogenous variables due to data limitations, the LSTM model incorporates all available predictors to enhance forecasting performance. This two-stage process allows the model to address both linear and nonlinear dependencies.

4.6.2 MIDAS-LSTM Hybrid

Similar to ARIMAX-LSTM, the MIDAS-LSTM hybrid model leverages MIDAS regression to integrate mixed frequency data while accounting for lag structures through parametric weighting constraints. MIDAS incorporates the relationship between macroeconomic indicators at different frequencies, reducing information loss that arises from aggregation. However, since MIDAS is still a linear model, it does not capture nonlinear dependencies within high-frequency predictors.

To address this, an LSTM model is trained on the residuals of the MIDAS regression, identifying hidden nonlinear dynamics that MIDAS may overlook. By feeding these residuals into the LSTM model along with the complete set of predictors, the hybrid approach improves the accuracy of the prediction.

4.7 Residual Analysis

Residual analysis is a critical step in evaluating the performance of linear forecasting models. It helps determine whether the residuals follows key statistical assumptions such as normality and lack of autocorrelation, ensuring the model's validity. In this study, Shapiro-Wilk test was performed for normality assessment and Ljung-Box test to detect autocorrelation in the residuals. These statistical checks provided insights into whether the residuals are well-behaved and whether the model captures the underlying data structure effectively.

4.7.1 Shapiro-Wilk Test for Normality

The Shapiro-Wilk test (Shapiro and Wilk, 1965) is a widely used test for checking if a dataset follows a normal distribution. The test statistic W is computed as:

$$W = \frac{\left(\sum_{i=1}^{n} a_i Z_{(i)}\right)^2}{\sum_{i=1}^{n}(Z_i - \bar{Z})^2}, \tag{15}$$

where: $Z_{(i)}$ are the ordered residuals (smallest to largest); a_i are weights derived from the expected values of order statistics of a normal distribution; $\bar{Z}$ is the sample mean; n is the sample size. A small p-value corresponds with a rejection of the normality of the residuals, indicating possible skewness or heavy tails in the error terms.

4.7.2 Ljung-Box Test for Autocorrelation

The Ljung-Box test examines whether the residuals exhibit autocorrelation by testing the null hypothesis that the residuals are independent over time (Ljung and Box, 1978). The test statistic is given by:

$$Q = n(n+2)\sum_{k=1}^{m} \frac{\hat{r}_k^2}{n-k}, \tag{16}$$

where: n is the sample size; $\hat{r}_k$ is the sample autocorrelation at lag k; m is the number of lags being tested.

The test follows a chi-square distribution with m degrees of freedom. A small p-value corresponds with a rejection of the independence of the residuals, implying serial correlation and possibly needing model modifcations.

4.8 Model Performance/Accuracy

In time series forecasting, evaluating model performance requires a comprehensive set of metrics to measure accuracy and prediction power. For our fitted models, two of the most popular evaluation metrics Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are used. Furthermore, to compare the predictive power of two competing models, the Diebold-Mariano (DM) test is applied, which statistically determines whether one model provides significantly better forecasts than another.

4.8.1 Root Mean Squared Error (RMSE)

RMSE measures the average magnitude of forecast errors, giving higher weight to larger errors due to squaring. This makes it useful for models where large deviations need to be penalized. RMSE is given by:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}. \tag{17}$$

RMSE is measured in the same units as the dependent variable, making it interpretable and suitable for comparing different models.

4.8.2 Mean Absolute Error (MAE)

MAE measures the average absolute difference between actual and predicted values:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|. \tag{18}$$

Unlike RMSE, MAE treats all errors equally and does not penalize large errors more than small ones and is therefore more robust against outliers. It is often preferred for interpreting real-world forecasting errors in business and finance.

4.8.3 Diebold-Mariano Test for Model Comparison

In addition to evaluating individual model performance using the above metrics, it is essential to determine whether one model significantly outperforms another. The Diebold-Mariano (DM) test is a statistical test that compares the forecast accuracy of two models based on their prediction errors (Diebold and Mariano, 2002). The DM test evaluates the difference between the forecast errors of two models:

$$d_t = L(e_{1,t}) - L(e_{2,t}), \tag{19}$$

where $e_{1,t}$ and $e_{2,t}$ are the forecast errors from Model 1 and Model 2, respectively and $L(e)$ is the loss function, typically squared error or absolute error.
The test statistic is given by:

$$DM = \frac{\bar{d}}{\sqrt{\frac{2\pi \hat{f}_d(0)}{n}}}, \tag{20}$$

where n denotes the sample size, $\bar{d}$ gives mean difference in loss functions and $\hat{f}_d(0)$ is an estimate of the spectral density at frequency zero
The null hypothesis of the DM test states that both models have equal forecasting accuracy

5 Results

In this section, the results of the model predictions are presented and discussed. First, we compare the MAE and RMSE of each model together with the p-values of the Diebold-Mariano test and relate this to the research questions. Additionally, the normality and independence of residuals of the linear benchmark models (ARIMAX and MIDAS) are assessed using the Shapiro-Wilk test and the Ljung-Box test. Subsequently we point out the substantial impact of COVID on predictive quality by means of visual illustrations. Finally, we present and discuss the prediction performance of the models without the COVID period.

5.1 Predictive Performance

In Table 6 we observe that the MIDAS regression model outperforms the ARIMAX model for three of the six target variables (EU-UNEMP, US-GDP, and US-PCE). Notably, for US-GDP and US-PCE, the MIDAS model significantly outperforms the ARIMAX model as the MAE decreases from 8.88 to 4.80 (approximately 46% decrease). However, the Diebold-Mariano test does not indicate significant differences in predictions between the two models with all p-values being above 0.05. The results are therefore inconclusive for the linear model specifications. The estimation datasets for the monthly target variables contain only daily variables as high frequency data, as is explained in Section 3. This may limit the effect of the MIDAS aggregation method and hamper the model's performance. On the contrary, for quarterly target variables both daily and monthly variables are considered as high frequency, increasing the influence of MIDAS aggregation on model performance. However, in this scenario daily variables are first aggregated into monthly variables by means of the custom aggregation as described in Section 3.4. This may once again influence the effect of MIDAS aggregation. Neither of the linear models significantly outperform the other for both monthly and quarterly target variables.

Apart from the difference in aggregation, the MIDAS method does not contain the moving average component of the ARIMAX model (see Section 4.2 and Section 4.3), which may also influence its predictive performance.

The ARIMAX-LSTM hybrid model demonstrates superior predictive performance compared to both benchmark models (ARIMAX and LSTM). Nevertheless, the Diebold-Mariano test yields only one p-value below 0.005 (US-GDP), while for all six target variables the ARIMAX-LSTM hybrid achieves lower MAE and RMSE values. As described in Section 4.4, the LSTM requires a large number of observations to perform well, leading to the expectation that the LSTM will be better suited for predicting monthly target variables. This is however not what is observed in the predictions. For example, for US-GDP, a quarterly target, we see the LSTM model outperforming the ARIMAX model, while for EU-UNEMP, a monthly variable, the ARIMAX model outperforms LSTM. Moreover, we do not necessarily see a clear relation between the overall performance of the LSTM and the relative improvement of the ARIMAX-LSTM hybrid over the standard ARIMAX model. An example is the EU-PCE target, for which the LSTM demonstrates relatively weak predicting power by having a higher MAE and RMSE than the ARIMAX model, yet the hybrid decreases the MAE from 2.89 to 1.55. On the contrary, for US-GDP the LSTM model does outperform the ARIMAX model, but the hybrid MAE 'only' decreases from 8.76 to 8.17.

Similarly, the MIDAS-LSTM hybrid model outperforms its benchmark models (MIDAS and LSTM). However, this is again not supported by the Diebold-Mariano test with p-values below 0.05 for only EU-UNEMP (MIDAS) and US-UNEMP(LSTM).

When comparing the MIDAS-LSTM hybrid and the ARIMAX-LSTM hybrid, neither model completely dominates and the Diebold-Mariano test does not suggest a significant difference in predictive accuracy for each of the six target variables. However, The MIDAS-LSTM hybrid does outperform the ARIMAX-LSTM hybrid for for all target variables except EU-PCE. Notably, for EU-GDP and US-UNEMP, the MIDAS-LSTM hybrid outperforms the ARIMAX-LSTM hybrid, despite the standard ARIMAX model outperforming the standard MIDAS model. It is therefore within reason to state that the performance gain introduced by the LSTM component in the hybrid specification is much larger for MIDAS than for ARIMAX. The largest performance gap is observed for US-GDP, where the MIDAS-LSTM hybrid achieves significantly higher prediction accuracy than the ARIMAX-LSTM hybrid with MAE values of 3.00 and 8.17 respectively.

Target	Error	ARIMAX	MIDAS	LSTM	A-LSTM	M-LSTM
EU-GDP	MAE	1.916	2.213	3.195	1.553	0.764
	RMSE	3.493	4.337	4.279	1.899	0.930
EU-PCE	MAE	2.893	4.538	4.257	1.544	2.011
	RMSE	5.379	7.890	5.797	2.028	2.815
EU-UNEMP	MAE	7.970	6.972	9.668	6.755	5.160
	RMSE	11.110	9.928	12.385	9.217	6.964
US-GDP	MAE	8.876	4.804	7.587	8.176	3.009
	RMSE	14.626	9.768	12.217	10.422	3.690
US-PCE	MAE	0.311	0.162	0.252	0.145	0.139
	RMSE	1.091	0.216	0.326	0.190	0.184
US-UNEMP	MAE	18.109	18.665	9.134	13.760	11.827
	RMSE	47.726	52.139	31.691	24.941	24.243

Table 6: Model MAE and RMSE Measures for All Target Variables. The table presents MAE and RMSE for different forecasting models applied to GDP, private consumption (PCE), and unemployment (UNEMP) for the Eurozone and US. It compares ARIMAX, MIDAS, LSTM, A-LSTM (ARIMAX-LSTM hybrid), and M-LSTM (MIDAS-LSTM hybrid), with lower values indicating better predictive performance.

Comparison	EU-GDP	EU-PCE	EU-UNEMP	US-GDP	US-PCE	US-UNEMP
A vs M	0.092	0.188	0.092	0.271	0.311	0.582
A vs L	0.392	0.874	0.392	0.452	0.338	0.010
A vs A-L	0.049	0.953	0.049	0.268	0.435	0.218
A vs M-L	0.489	0.226	0.489	0.285	0.316	0.510
M vs L	0.281	0.300	0.281	0.009	0.025	0.001
M vs A-L	0.113	0.188	0.113	0.174	0.306	0.824
M vs M-L	0.379	0.947	0.379	0.255	0.099	0.364
L vs A-L	0.571	0.878	0.571	0.213	0.334	0.012
L vs M-L	0.874	0.273	0.874	0.164	0.098	0.001
A-L vs M-L	0.630	0.229	0.630	0.191	0.312	0.748

Table 7: DM Test p-values for All Model Comparisons. The table presents p-values from the Diebold-Mariano test. The models compared include ARIMAX (A), MIDAS (M), LSTM (L), ARIMAX-LSTM hybrid (A-L), and MIDAS-LSTM hybrid (M-L) across GDP, private consumption (PCE), and unemployment (UNEMP) for the EU and US. Lower p-values (< 0.05) indicate a significant difference in predictive performance between the compared models, while higher values suggest similar forecasting accuracy.

Target	Statistical Test	ARIMAX	MIDAS
EU-GDP	Shapiro-Wilk	< 0.001	0.653
	Ljung-Box	0.988	0.022
EU-PCE	Shapiro-Wilk	< 0.001	< 0.001
	Ljung-Box	0.999	0.350
EU-UNEMP	Shapiro-Wilk	< 0.001	0.016
	Ljung-Box	0.997	0.001
US-GDP	Shapiro-Wilk	< 0.001	0.496
	Ljung-Box	0.983	0.721
US-PCE	Shapiro-Wilk	< 0.001	< 0.001
	Ljung-Box	0.323	0.732
US-UNEMP	Shapiro-Wilk	< 0.001	< 0.001
	Ljung-Box	0.999	0.009

Table 8: p-values for Shapiro Wilk and Ljung box test. Table shows p-values for Shapiro wilk and Ljung box test performed on residuals generated for Linear models ARIMAX and MIDAS

5.2 Residual Analysis

Table 8 presents the p-values for the Shapiro-Wilk and Ljung-Box tests, which assess normality and serial correlation of the residuals, as discussed in Section 4.7 . The results of the Shapiro-Wilk test indicate a rejection of normality for the error terms of all six target variables. This contradicts the assumptions underlying the ARIMAX model, as discussed in Section 4.2. Therefore, we have some reservations for statistical inference on these results. Moreover, it suggests that the model fails to capture all relevant patterns in the data. Consequently, these findings support the implementation of an LSTM model on the residuals, as utilized in the ARIMAX-LSTM hybrid. However, the Ljung-Box test results do not indicate significant autocorrelation among the residuals, implying that the ARIMAX model might have effectively captured the autoregressive patterns of the target variables.

For the MIDAS regression model, the results indicate that, for both GDP predictions, the Shapiro-Wilk test does not reject normality. However, for three of the six target variables (EU-UNEMP, EU-GDP, and US-UNEMP), the Ljung-Box test suggests significant autocorrelation (p ¡ 0.05), implying residual dependence. In conclusion, only for US-GDP, the residuals exhibit white noise characteristics, suggesting model adequacy. This finding aligns with the superior performance of the MIDAS model compared to ARIMAX for US-GDP, as previously discussed. However, despite the white noise characteristics for the MIDAS residuals, the MIDAS-LSTM hybrid still substantially increase prediction accuracy, decreasing the RMSE from 9.76 to 3.69. The remaining target variables appear to support the implementation of the LSTM model on the residuals to capture possible remaining patterns in the data. Moreover, we would have reservations for statistical inferencing on these variables.

5.3 Deletion of COVID Period - Special Investigation

In this section, we provide visual comparisons of each model's predictions against the actual values from our test set for each target variable (see Figure 3 - 5). Notably, the COVID-19 period, accounts for a substantial portion of the prediction loss, significantly influencing the results related to our research questions but also significantly influencing the model predictions after this period, as our prediction models are retrained after every increment. Figure 9 presents the MAE and RMSE values for each prediction model when trained and tested on data for which the COVID peak period (1 Jan 2020 - 30 June 2022) is deleted. Figure 6 - 8 (Appendix 7.2) are the new visualizations of the predictions for each target variable without the COVID peak period.

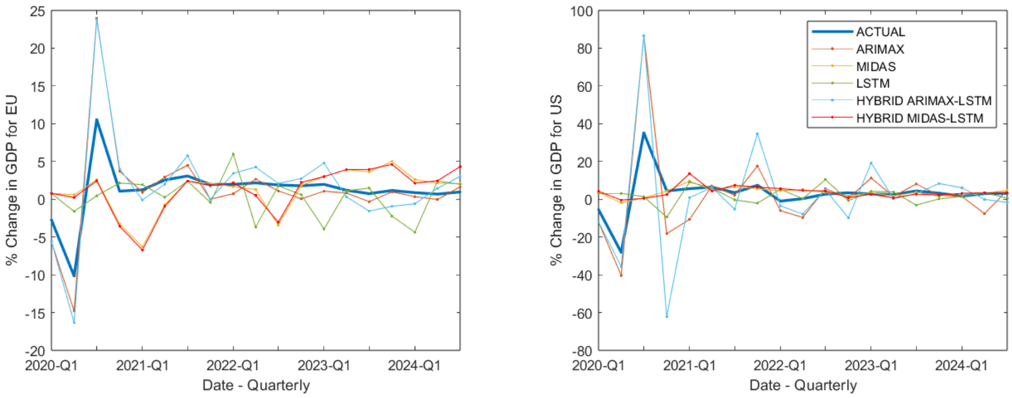

Figure 3: GDP % Change for EU(left panel) and US(right panel). The graph illustrates the quarterly percentage change in GDP for the Eurozone and the US across different forecasting models.

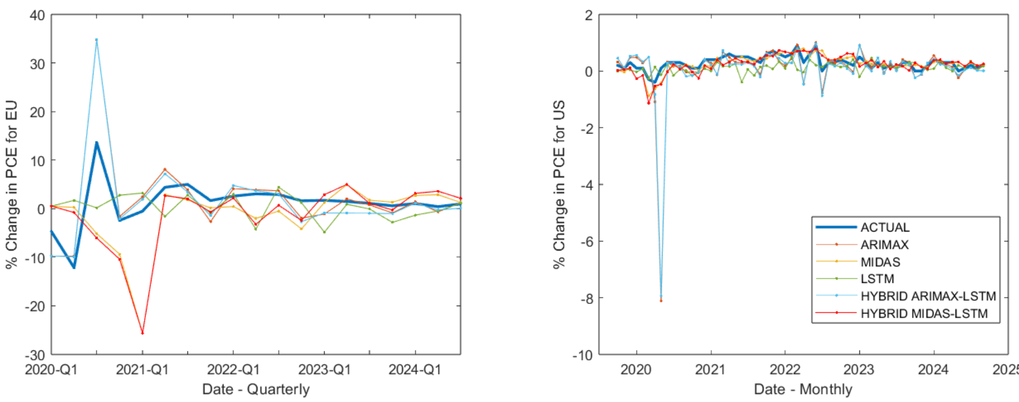

Figure 4: Private Consumption % Change for EU(left panel) and US(right panel). The graph presents the percentage change in Private Consumption Expenditures (PCE) for the Eurozone (quarterly frequency) and US (monthly frequency) across various fitted models.

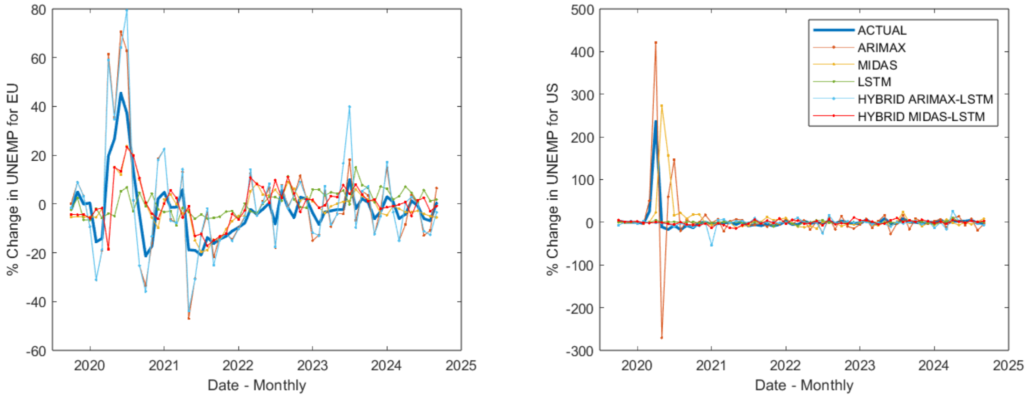

Figure 5: Unemployment % Change for EU(left panel) and US(right panel). The graph for the Eurozone and the US illustrates the percentage change forecast in unemployment as predicted by different forecasting models at monthly frequency.

The new 'without COVID' plots clearly show that the total prediction error is no longer extremely dominated by a small subset of observations. Although the average prediction accuracy has substantially increased for each scenario compared to the original 'with COVID' data, relationships in performance between the different models appear to be similar. Again, neither ARIMAX nor MIDAS is outperforming the other, with MIDAS now outperforming in three of the six target variables. Similarly, for the comparison of the ARIMAX-LSTM with the MIDAS-LSTM hybrid there remains to be no definitive winner. Lastly, it is again observed that both hybrids outperform their individual counterparts. In other words, COVID does not change our conclusions on the prediction models.

Target	Error	ARIMAX	MIDAS	LSTM	A-LSTM	M-LSTM
EU-GDP	MAE	0.456	0.628	0.687	0.296	0.612
	RMSE	0.578	0.905	1.005	0.363	0.702
EU-PCE	MAE	0.661	0.699	0.770	0.276	0.332
	RMSE	1.396	1.346	1.070	0.352	0.403
EU-UNEMP	MAE	5.292	4.343	6.495	4.099	4.332
	RMSE	6.494	5.409	8.453	4.795	5.380
US-GDP	MAE	3.066	1.822	1.929	1.840	1.083
	RMSE	6.702	1.472	2.223	2.542	1.432
US-PCE	MAE	0.359	0.129	0.194	0.160	0.129
	RMSE	1.622	0.172	0.262	0.206	0.171
US-UNEMP	MAE	2.790	2.809	2.657	2.594	2.592
	RMSE	3.286	3.317	3.275	3.102	3.185

Table 9: Model MAE and RMSE Measures for All Target Variables: The table presents MAE and RMSE for different forecasting models applied to GDP, private consumption (PCE), and unemployment (UNEMP) for the EU and US, with COVID-19 observations removed. It compares ARIMAX, MIDAS, LSTM, A-LSTM (ARIMAX-LSTM hybrid), and M-LSTM (MIDAS-LSTM hybrid), showing model performance without pandemic-induced volatility

6 Practical Assessment of Model Results

This research is conducted in partnership with the model validation department of Van Lanschot Kempen. Valuable theoretical insight is provided by assessing and comparing the performance of different models on the basis of pure mathematical measures such as MAE, RMSE and the DM test. These methods may however fall short in providing the partner organization with practical insight into model performance. In agreement with Van Lanschot Kempen, two practical model insight are addressed: 1. Does a model correctly predict the sign of a change in a target variable; and 2. How often does the model produce a prediction that can be classed as 'excessively' incorrect? Intuitively, an imprecise model in this particular case is not practically useless if it can correctly predict the change direction of a target variable with an acceptable degree of accuracy. If the specific values of model predictions are of importance in their respective use, then it is likely valuable to know how often the model may produce excessively bad predictions with damaging consequences for the organization. The remainder of this section is divided into 2 subsections, addressing each of the aforementioned practical considerations. Each section provides a brief overview of how predictions are classified, followed by results and interpretation.

6.1 Correct Prediction of Target Variable Change Directions

Correctly predicted change directions of a specific target variable are counted by constructing a binary vector $\boldsymbol{b}$ according to the following classification rule:

$$b_i = \left\{ \begin{array}{ll} 1, & \text{if } (y_i < 0 \ \& \ \hat{y}_i < 0) \text{ OR } (y_i \geq 0 \ \& \ \hat{y}_i \geq 0) \\ 0, & \text{if } (y_i \geq 0 \ \& \ \hat{y}_i < 0) \text{ OR } (y_i < 0 \ \& \ \hat{y}_i \geq 0) \end{array} \right\}, \text{ for i} = 1...n_{test}, \tag{21}$$

where n_{test} is the number of test set observations, y_i is the actual target variable percent change and $\hat{y}_i$ is the predicted percent change. The percentage of 1's in $\boldsymbol{b}$ is the percentage of correctly predicted direction changes in the target variable of interest. Table 10 contains such calculated percentages for all target variables and models, with the addition of average model accuracy across all targets. As seen in 5.1, the hybrid models outperform the standard LSTM and linear models when considering RMSE and MAE measures. However, referring to Table 10 it is evident that the linear models predict sign changes correctly in a larger percentage of predictions than the LSTM or hybrid models.

6.2 Frequency of Extreme Predictions

We adopt a convention to class predictions as extreme if their error exceeds the standard deviation of the corresponding target variable. Percentages are calculated similarly to those in 6.1,

Target	ARIMAX	MIDAS	LSTM	A-LSTM	M-LSTM
EU-GDP	89.5	68.4	68.4	73.7	68.4
US-GDP	78.9	89.4	57.9	68.4	89.4
EU-PCE	68.4	68.4	36.8	52.6	68.4
US-PCE	76.7	91.2	83.3	75.0	90.0
EU-UNEMP	86.7	60.0	46.7	85.0	50.0
US-UNEMP	76.7	40.0	46.7	56.7	58.3
Average:	79.5	70.0	56.6	68.6	70.8

Table 10: % of Correctly Predicted Change Directions for All Models and Targets

using a binary vector $\boldsymbol{b}$. Table 11 contains the calculated percentages of the number of extreme predictions made per model, for all models and target variables. Standard deviation values are included to provide context about the 'extreme prediction' threshold used.

Target (STDev)	ARIMAX	MIDAS	LSTM	A-LSTM	M-LSTM
EU-GDP (3.56)	10.5	36.8	36.8	10.5	31.5
US-GDP (10.62)	31.5	10.5	15.8	26.3	10.5
EU-PCE (4.68)	10.52	36.8	36.8	10.5	31.6
US-PCE (0.24)	21.7	20.0	43.3	30.0	31.7
EU-UNEMP (12.03)	18.3	15.0	23.3	23.3	21.6
US-UNEMP (31.08)	6.7	6.7	1.7	3.3	1.7

Table 11: Target Variable Standard Deviations and Percentages of Extreme Predictions

In Table 11 the best performing models are observed to be: ARIMAX and A-LSTM for EU-GDP; MIDAS and M-LSTM for US-GDP; ARIMAX and A-LSTM for EU-PCE; MIDAS for US-PCE; MIDAS for EU-UNEMP; LSTM and M-LSTM for US-UNEMP. These results provide insufficient evidence to conclude that any of the 5 compared models more frequently generates extreme predictions than the others.

7 Conclusion

This study applies several nowcasting models to predict the percentage change of three macroeconomic indicators, Gross Domestic Product (GDP), Private Consumption (PCE) and the Unemployment Rate (UNEMP), for both the Eurozone and the US. We evaluate five prediction models: ARIMAX, MIDAS, LSTM, ARIMAX-LSTM and MIDAS-LSTM.

By comparing ARIMAX and MIDAS we attempt to answer our first research question. We highlight two distinct approaches to aggregating high-frequency data for our nowcasting predictions: an intuitive simple aggregation method for ARIMAX and the more sophisticated aggregation properties of MIDAS. Our results do not provide conclusive evidence that one approach consistently outperforms the other across all our target variables.

Additionally, we introduce two novel hybrid models, ARIMAX-LSTM and MIDAS-LSTM, and evaluate their predictive performance relative to their individual components, revisiting our second and third research question. These hybrid models first employ their linear component (ARIMAX or MIDAS) to predict the target variable and then subsequently use an LSTM model to predict the residuals, capturing potential remaining non-linear relationships. We observe that the ARIMAX-LSTM hybrid outperforms both ARIMAX and LSTM across all our target variables. Similarly, the MIDAS-LSTM hybrid consistently performs better than both MIDAS and LSTM. However, for both hybrid models, the Diebold-Mariano test does not confirm statistically significant differences in predictive accuracy between the hybrids and their corresponding individual components.

Furthermore, regarding our fourth research question, no clear superiority is observed between ARIMAX-LSTM and MIDAS-LSTM, mirroring the inconclusive comparison between ARIMAX and MIDAS. Consequently, our results do not identify a single superior model for our main research goal nowcasting changes in macroeconomic indicators. Nevertheless, despite the lack of statistically significant differences, both hybrid models do consistently outperform their individual counterparts across all our target variables, suggesting their practical relevance.

This research contributes to the existing literature on nowcasting macroeconomic indicators by evaluating the predictive accuracy of various models. In particular we compare a simple intuitive aggregation approach in ARIMAX with the more sophisticated MIDAS regression, adding to the growing body of research on utilizing high-frequency data for predictive modeling. Moreover, to the best of our knowledge, this study is the first to propose and evaluate the ARIMAX-LSTM and MIDAS-LSTM hybrid models for nowcasting macroeconomic indicators.

Future research could explore alternative hybrid models by combining different prediction techniques. For instance, integrating a dynamic factor model with a neural network could provide further insights. Additionally, our current MIDAS approach is limited to aggregating a single frequency into another. Investigating methods for simultaneously integrating multiple high-frequency data sources may enhance predictive accuracy.

References

Box George, E., Jenkins Gwilym, M., Reinsel Gregory, C., and Ljung Greta, M. (1976). Time series analysis: forecasting and control. *San Francisco: Holden Bay.*

Cheung, Y.-W. and Lai, K. S. (1995). Lag order and critical values of the augmented dickey–fuller test. *Journal of Business & Economic Statistics*, 13(3):277–280.

Clements, M. P. and Galvão, A. B. (2008). Macroeconomic forecasting with mixed-frequency data: Forecasting output growth in the united states. *Journal of Business & Economic Statistics*, 26(4):546–554.

Conrad, C. A. (2022). *Applied Macroeconomics*. Springer Nature.

Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 20(1):134–144.

Gers, F. A., Schmidhuber, J., and Cummins, F. (2000). Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451–2471.

Ghysels, E., Kvedaras, V., and Zemlys-Balevičius, V. (2020). Chapter 4 - mixed data sampling (midas) regression modelsthe opinions expressed are those of the authors only and should not be considered as representative of the european commission's official position. In Vinod, H. D. and Rao, C., editors, *Financial, Macro and Micro Econometrics Using R*, volume 42 of *Handbook of Statistics*, pages 117–153. Elsevier.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2004). The midas touch: Mixed data sampling regression models.

Ghysels, E., Sinko, A., and Valkanov, R. (2007). Midas regressions: Further results and new directions. *Econometric reviews*, 26(1):53–90.

Hamiane, S., Ghanou, Y., Khalifi, H., and Telmem, M. (2024). Comparative analysis of lstm, arima, and hybrid models for forecasting future gdp. *Journal homepage: http://iieta. org/journals/isi*, 29(3):853–861.

Hochreiter, S. (1997). Long short-term memory. *Neural Computation MIT-Press.*

Ljung, G. and Box, G. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65.

Longo, L., Riccaboni, M., and Rungi, A. (2022). A neural network ensemble approach for gdp forecasting. *Journal of Economic Dynamics and Control*, 134:104278.

Marcellino, M. G. and Schumacher, C. (2007). Factor-midas for now-and forecasting with ragged-edge data: a model comparison for german gdp.

Markham, I. S. and Rakes, T. R. (1998). The effect of sample size and variability of data on the comparative performance of artificial neural networks and regression. *Computers & operations research*, 25(4):251–263.

Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Peter, Ď. and Silvia, P. (2012). Arima vs. arimax–which approach is better to analyze and forecast macroeconomic time series. In *Proceedings of 30th international conference mathematical methods in economics*, volume 2, pages 136–140.

SANUSI, N. A., MOOSIN, A. F., and KUSAIRI, S. (2020). Neural network analysis in forecasting the malaysian gdp. *The Journal of Asian Finance, Economics and Business*, 7(12):109–114.

Tan, M., Hu, C., Chen, J., Wang, L., and Li, Z. (2022). Multi-node load forecasting based on multi-task learning with modal feature extraction. *Engineering applications of artificial intelligence*, 112:104856.

Taslim, D. and Murwantara, I. (2024). Comparative analysis of arima and lstm for predicting fluctuating time series data. *Bulletin of Electrical Engineering and Informatics*, 13:1943–1951.

Varian, H. (2014). Machine learning and econometrics. *Slides package from talk at University of Washington.*

Wohlrabe, K. (2009). *Forecasting with mixed-frequency time series models.* PhD thesis, lmu.

Xu, Z., Chen, J., Shen, J., and Xiang, M. (2022). Recursive long short-term memory network for predicting nonlinear structural seismic response. *Engineering Structures*, 250:113406.

Zhang, G. P. (2003). Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.

Appendix: Omitted Visual Illustrations and Tables

7.1 Algebraic Explanation of Extrapolation

Let T represent the total number of observations and t be a point in time beyond which there are no more *nan* values. Lower values of t represent older dates. Denote by a_j the jth column of A with an arbitrary number of *nan* values. a_{j+} is the complete data vector that is used to estimate the *nan* values in a_{j-}. ARIMA extrapolation handles each problem variable independently and reverses the indexing of a_{j+} such that 'future' prediction actually equates to predicting values in the past. In the case of LSTM extrapolation the inter-variable relationships are modeled and used for prediction of missing values. Suppose once again that a_j is the problem variable and that a_1 and a_2 are complete variables. a_{1+} and a_{2+} will be used as independent input variables while a_{j+} will be used as dependent output variable in the training process. a_{1-} and a_{2-} will be used to predict the *nan* values in a_{j-} with the trained LSTM Neural Network. The Matrix A, and vectors a_j, a_{j+}, a_{j-} referred to in the text are respectively:

$$A_{T\times k} = \begin{bmatrix} a_{11} & a_{12} & \cdots & nan & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & \vdots & \cdots & a_{2k} \\ \vdots & \vdots & \cdots & nan & \cdots & \vdots \\ a_{t1} & a_{t2} & \cdots & a_{tj} & \cdots & a_{tk} \\ a_{(t+1)1} & a_{(t+1)2} & \cdots & a_{(t+1)j} & \cdots & a_{(t+1)k} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ a_{T1} & a_{T2} & \cdots & a_{Tj} & \cdots & a_{Tk} \end{bmatrix}, \; a_j = \begin{bmatrix} nan \\ \vdots \\ nan \\ a_{tj} \\ a_{(t+1)j} \\ \vdots \\ a_{Tj} \end{bmatrix},$$

$$a_{j+} = \begin{bmatrix} a_{tj} \\ a_{(t+1)j} \\ \vdots \\ a_{Tj} \end{bmatrix}, \text{ and } a_{j-} = \begin{bmatrix} nan \\ nan \\ \vdots \\ nan \end{bmatrix}.$$

7.2 Plots of Model Predictions After Covid Deletion

This section provides graphs of the model predictions after deleting the Covid period of 1 Jan 2020 to 30 June 2022 and is referenced in 5.

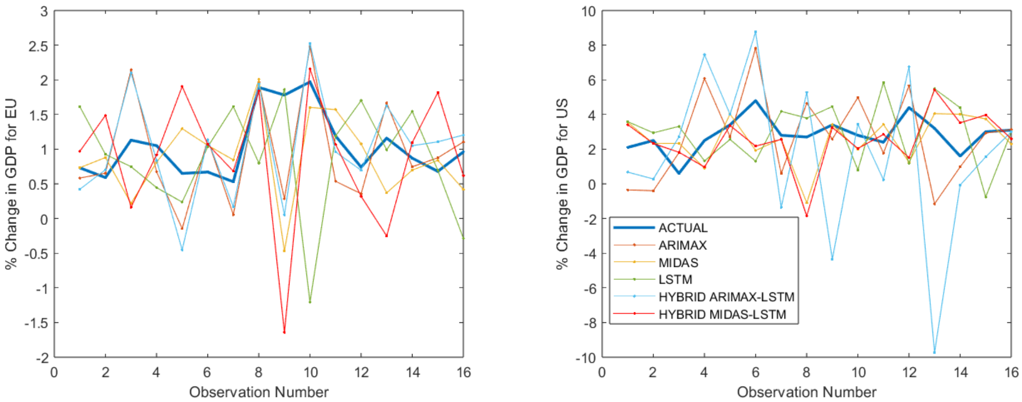

Figure 6: GDP % Change for EU(left panel) and US(right panel). This figure presents the percentage change in GDP for the Eurozone and US, using various forecasting models, excluding the COVID period.

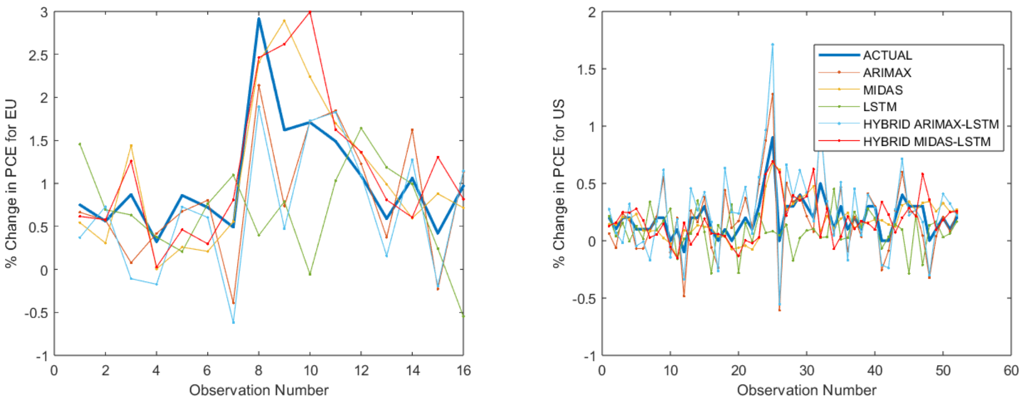

Figure 7: Private Consumption % Change for EU(left panel) and US(right panel). The graph illustrates the predicted percentage change in private consumption expenditure (PCE) for the Eurozone and US, without COVID observations.

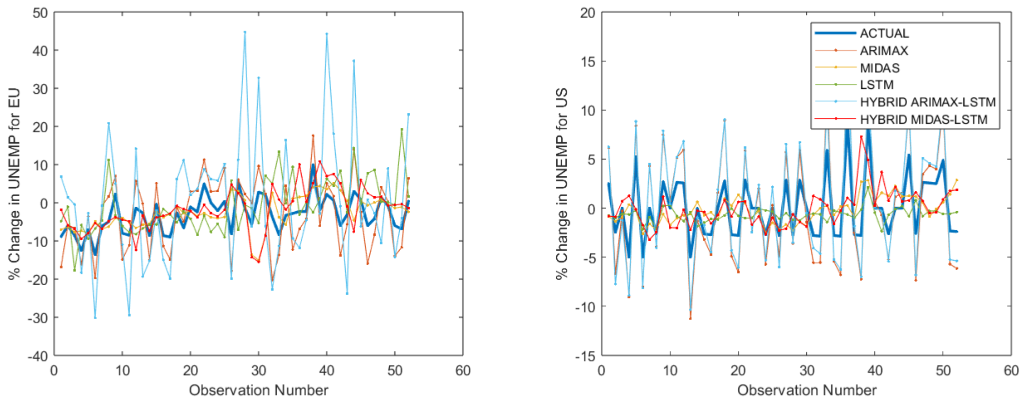

Figure 8: Unemployment % Change for EU(left panel) and US(right panel). This figure shows the percentage change in unemployment for the Eurozone and US, based on different forecasting models, with COVID data removed.

7.3 Hyperparameter Tuning

Hyperparameter tuning was performed using Bayesian Optimization via Keras Tuner. The search space for the hyperparameters is detailed in Table 12, in line with the tuning in Longo et al. (2022). Note that before any tuning or training, we normalize our data.

Hyperparameter	Values Considered
Number of LSTM Layers	{1, 2}
Units in First LSTM Layer	{10, 20, 30, 40, 50}
Activation Function (LSTM Layer 1)	{ReLU}
Dropout Rate (LSTM Layer 1)	{0.1, 0.15, 0.2, 0.25, 0.3}
Units in Second LSTM Layer (if present)	{10, 20, 30}
Activation Function (LSTM Layer 2)	{ReLU, SELU}
Dropout Rate (LSTM Layer 2)	{0.1, 0.15, 0.2, 0.25, 0.3}
Units in Dense Layer	{5, 10, 15, 20}
Learning Rate (Adam Optimizer)	{0.001, 0.005, 0.01}
Beta 1 (Adam Optimizer)	{0.9, 0.95, 0.99}
Beta 2 (Adam Optimizer)	{0.999, 0.9999}
Epsilon (Adam Optimizer)	{1e-8, 1e-7}

Table 12: Hyperparameter Search Space for LSTM Model Tuning