Erasmus University Rotterdam
Erasmus School of Economics
Master Thesis Econometrics and Management Science

An inflation forecasting contest: Classical models, survey forecasts or machine learning methods?

Author: Rohan Adjodha
Student ID number: 511443

Supervisor:	dr. M. Khismatullina
Second assessor:	T. van der Zwan, MSc
Date final version:	30th April 2024

Abstract

Accurately anticipating inflation rates is highly relevant in economic-decision making. The goal of this study is to compare the accuracy of judgemental forecasts, classical statistical models and (non-linear) machine learning models in forecasting inflation out-of-sample. The datasets that I use contain annualized US CPI, GDP deflator and PCE deflator inflation rates, as well as real-time macroeconomic predictors, at the quarterly frequency. In particular, I examine survey forecasts, autoregressive models, Phillips curve models, a stochastic volatility specification, term structure models, the random forest and the long short-term memory recurrent neural network. My results exhibit that the survey data generally provide more accurate inflation forecasts than the remaining methods, due to their advantage in measuring the current inflation rate.

Keywords: Forecasting; Machine learning; Phillips curve; SPF; Term structure models; Real-time data

Contents

1 Introduction

The relevance of accurately anticipating inflation rates can hardly be overstated. As long-term nominal obligations regarding labor, sales, leases, mortgages, and other debts are prevalent in today's economies, the private sector has an inherent stake in making accurate inflation forecasts. Moreover, central banks establish their monetary policy on inflation forecasts, and implement inflation expectations to improve potency of policies. Precise inflation forecasts are not only beneficial in economic decision-making, but also in predicting other related (macro)economic indicators, such as consumer spending and interest rates. Finally, inflation forecasts are also important for traders who aim to optimize portfolios that consist of inflation-related instruments, such as indexed-linked bonds and inflation derivatives.

The academic literature comprises various inflation forecasting methods. A traditional approach is to implement survey data (Grant and Thomas, 1999; Thomas Jr, 1999; Mehra, 2002; Ang, Bekaert and Wei, 2007), for example the Survey of Professional Forecasters, which is the oldest quarterly survey of macroeconomic forecasts in the United States. Classical statistical inflation forecasting models consist of many distinct approaches, such as models based on the economically motivated Phillips curve as in Fuhrer (1995), Brayton, Roberts and Williams (1999), Stock and Watson (1999) and Stock and Watson (2008) and stochastic volatility models as in Stock and Watson (2007). In addition, vector autoregressive models are also known to provide accurate inflation predictions (Sims, 1993; Stock and Watson, 1996; Cogley and Sargent, 2001; Athanasopoulos and Vahid, 2008). Recently, machine learning methods have gained more popularity within the inflation forecasting problem field as a result of the current availability of large amounts of data and improvement in computational power. Specifically, non-linear machine learning models including the random forest and the long short-term memory recurrent neural network are accurate in forecasting inflation, as exhibited by Medeiros et al. (2021) and Almosova and Andresen (2023), respectively.

As there are many distinct methods, it is useful to gain more insight in the performance of the different methods and to investigate their respective benefits and drawbacks. Not many researches have yet conducted an extensive comparison of the performance of the existing methods within the inflation forecasting problem field. Faust and Wright (2013) give a comprehensive review consisting of both recently developed inflation forecasting methods and traditional approaches. However, their comparison does not involve machine learning methods, such as random forest and neural network specifications, which have exhibited accuracy gains when compared to autoregressive and random walk benchmark models (Medeiros et al., 2021; Goulet Coulombe et al., 2022; Almosova and Andresen, 2023).

Hence, I build upon the existing studies through comparing a variety of inflation forecasting approaches that have been successful in previous research, including machine learning methods and other recently developed models, in a so-called "horse race" setting and investigate which methods result in the most accurate inflation forecasts. In particular, I distinguish three streams of inflation forecasting methods that have shown effectiveness in previous research, namely judgemental forecasts derived from surveys, classical statistical models and machine learning methods. The judgemental forecasts in this study include projections derived from the Survey of Professional Forecasters as well as forecasts obtained from the contemporary term structure

model introduced by Borağan Aruoba (2020), which combines the data of several major surveys. Moreover, the statistical models in my research consist of a random walk, AR models, a stochastic volatility model, Phillips curve models and a VAR model. Lastly, the machine learning models that I examine involve both a random forest and a long short-term memory recurrent neural network specification. I aim to answer which of these inflation forecasting models generally performs best in terms of predictive accuracy. By investigating these particular models, I intend to not only discover their advantages and disadvantages, but also to assess the robustness of the findings from past research on these methods.

For my research, I implement real-time US Consumer Price Index, Price Index for GNP/GDP and Price Index for Personal Consumption Expenditures inflation data at the quarterly frequency for the period from 1948 until 2022, which are provided by the Federal Reserve Bank of Philadelphia. In my forecasting experiment, I investigate both short- and longer-term predictions based on an expanding window, with the first prediction made for the first quarter of 1997. Furthermore, I assess the forecasting performance by evaluating both the accuracy and bias.

This paper contributes to the academic literature in various ways. That is, I compare a wide-ranging set of inflation forecasting methods including classical statistical models, survey forecasts and non-linear machine learning models, whereas prior comparison studies solely focus on either one or two of these streams in the inflation forecasting problem field. For instance, the forecast comparisons from Ang, Bekaert and Wei (2007) and Faust and Wright (2013) only fixate on the former two streams. Moreover, the random forest and long short-term memory neural network have not yet been compared with other successful inflation forecasting approaches (e.g., survey forecasts), as previous research exclusively compares these methods with rather simple benchmarks or other machine learning models. Thus, my research provides more insight on the dominance of certain inflation forecasting methods.

Another contribution is that I investigate the performance of state-of-the-art inflation forecasting models. Specifically, I examine the Aruoba Term Structure of Inflation Expectations from Borağan Aruoba (2020) and the long short-term memory recurrent neural network. Although the latter model was introduced more than two decades ago by Hochreiter and Schmidhuber (1997) and has been implemented in several time series forecasting exercises, only recently a few papers adopted the long short-term memory recurrent neural network in predicting inflation rates, i.e., Rodríguez-Vargas (2020), Peirano, Kristjanpoller and Minutolo (2021) and Almosova and Andresen (2023). As both of these models are contemporary within the inflation forecasting problem field, there are fewer implications on their forecasting abilities with respect to other inflation forecasting methods. Therefore, the findings of my study bring better understanding of the performance of these models in predicting inflation. In addition, I propose a dynamic feature selection approach for the long short-term memory neural network, which allows the set of predictors to change over time.

The key findings from my study are as follows. The judgemental forecasts are the most accurate inflation forecasting methods in this comparison, consistently outperforming the classical statistical models and machine learning models. The survey projections have the greatest benefit over the remaining methods in predicting the current inflation rate and their dominance over longer

horizons is partially caused by this advantage. Moreover, the superiority of the judgemental forecasts can be ascribed to the expertise of professional macroeconomic forecasters, which allows them to more rapidly recognize and respond to certain shifts in the dynamics of inflation rates. Even though the performance of the machine learning methods differs across inflation measures, these methods are more accurate in forecasting over the long (two-year) horizon than over shorter intervals. The rationale behind this is that, over the long-run, the machine learning models are less affected by temporary shifts in inflation and the benefit of their non-linear structure is the highest. Furthermore, non-stationary specifications generally outperform the stationary models and the disparities in predictive accuracy between both increase over longer horizons, since the trend in inflation is evidently varying over time.

This paper proceeds as follows. In Section 2, I discuss previous research on inflation forecasting. Next, the data is described in Section 3. Subsequently, the methods and models that I use to predict inflation are explained in Section 4. I present the findings on the different inflation forecasting methods in Section 5. Finally, I draw conclusions in Section 6.

2 Literature

Predicting inflation rates is a frequently discussed topic in the academic literature. As the literature on the prediction of inflation rates is vast, it is not feasible to review all the past research. However, I aim to give a general overview of the most commonly used inflation forecasting methods and discuss the key developments in the academic research.

Classical inflation forecasting approaches consist of models based on the economically motivated Phillips curve, which relates inflation to the unemployment rate or other real activity measures (Fuhrer, 1995; Brayton, Roberts and Williams, 1999; Stock and Watson, 1999; Stock and Watson, 2008). Stock and Watson (1999) find that inflation predictions derived from the Phillips curve are, in general, more accurate than predictions constructed by other macroeconomic variables, such as interest rates, money and commodity prices.

However, follow-up research questions the validity of the findings from Stock and Watson (1999). Namely, many papers including Atkeson, Ohanian et al. (2001), Sims (2002), Fisher, Liu and Zhou (2002), and Clark and McCracken (2006) exhibit that the predictive accuracy of the Phillips curve-based models heavily relies on the sample period and that in many cases these models are outperformed by simple benchmarks, implying that the former results are not robust. Given that more than a decade has gone by since the latter research was published, it is relevant to study how the Phillips curve models perform in forecasting inflation over a more recent sample period. Hence, I incorporate the Phillips curve specification in my comparative analysis.

The scrutiny against the traditional Phillips curve led to the investigation of possible enhancements on the standard Phillips curve approach. Various studies find that employing the Phillips to the so-called inflation "gap" rather than directly applying it to inflation rates, leads to more accurate inflation forecasts. The inflation gap is defined as the difference between the inflation rate and its slowly time-varying local mean. Clark and McCracken (2006), Cogley, Primiceri and Sargent (2010) and Faust and Wright (2013) find that Phillips curve forecasts employing the inflation gap are more accurate than inflation expectations derived from stationary models, due to their ability to capture the low-frequency trend component prevailing in the past dynamics

of inflation rates. Since the majority of past research only compares the Phillips curve in gap form to stationary specifications, the performance of the inflation gap specification relative to other non-stationary specifications is rather uncharted. Accordingly, I also consider the Phillips curve gap model in my extensive forecast comparison, which includes additional non-stationary inflation forecasting approaches.

Another class of inflation forecasting models that allow for time-varying volatility is that of the stochastic volatility models. Stock and Watson (2007) show that an univariate unobserved component stochastic volatility model is able to accurately capture a variety of previous inflation rate shocks, which autoregressive equivalents were not able to apprehend. Similar results are found by Cogley, Primiceri and Sargent (2010), who apply a stochastic volatility model to the inflation gap. Moreover, Kim, Manopimoke and Nelson (2014), Cecchetti et al. (2017) and Mertens and Nason (2020) all motivate the implementation of unobserved component stochastic volatility model variants in forecasting inflation.

Chan (2013) builds upon the stochastic volatility model by introducing a new class of models that has stochastic volatility as well as moving average errors, using a state space representation for the conditional mean. Even though the resulting moving average stochastic volatility models have better in-sample fitness and out-of-sample forecasting ability than standard stochastic volatility models, the estimation of moving average stochastic volatility models is more challenging, as the errors in the measurement equation are no longer serially independent due to the moving average component. Considering there are numerous adaptations of the unobserved component stochastic volatility specification and to avoid additional computational complexity relative to the other inflation forecasting approaches reviewed in this paper, I concentrate on the fundamental model from Stock and Watson (2007).

Vector autoregressive (VAR) models are also known to provide accurate forecasts for macroeconomic variables, including inflation measures (Sims, 1993; Stock and Watson, 1996; Cogley and Sargent, 2001; Athanasopoulos and Vahid, 2008). In particular, such models often aim to explain the relation between Treasury yields and macroeconomic variables. For example, Faust and Wright (2013) fit a first-order VAR model to dynamic yield curve factors, the inflation gap, and the unemployment rate to construct a term structure based inflation forecast. Comparable approaches are taken by Primiceri (2005) and Joslin, Priebsch and Singleton (2014), from which the former even imposes yield factor coefficients to drift slowly over time by incorporating stochastic volatility components.

Motivated by the prior findings on the term structure VAR forecasts, I cover the specification in my comparative research. However, I reckon the approach from Faust and Wright (2013) rather than the VAR specification with stochastic volatility components, as I already investigate the stochastic volatility framework by itself in this study.

It is prevailing knowledge in the academic literature that survey forecasts are among the most precise inflation forecasting methods (Grant and Thomas, 1999; Thomas Jr, 1999; Mehra, 2002; Ang, Bekaert and Wei, 2007). Survey datasets comprise a large amount of subjective information about the inflation expectations from agents. Popular inflation expectation surveys are the Survey of Professional Forecasters, the Blue Chip Economic Indicators and Blue Chip Financial Forecasts, the Livingston Survey, and the Michigan Survey. Ang, Bekaert and Wei (2007) find

that survey forecasts outperform ARIMA models, regressions derived from the Phillips curve and (non-)linear term structure models. Similar results are obtained by Faust and Wright (2013), who find that the surveys provide the most accurate predictions in a horse-race among a large set of inflation forecasting methods including autoregressive models, Phillips curve models, stochastic volatility models and models in gap form.

Since many papers exhibit the superiority in terms of accuracy of survey forecasts over other inflation forecasting methods, it is essential to include such judgemental forecasts in my comparison study. Therefore, I examine The Survey of Professional Forecasters, which is the oldest quarterly survey of macroeconomic projections in the United States. The American Statistical Association and the National Bureau of Economic Research started conducting the survey in 1968 and the Federal Reserve Bank of Philadelphia assumed control of the survey in 1990. The panelists of this survey are experienced industry professionals who construct macroeconomic forecasts as part of their job responsibilities.

Even though survey forecasts are generally found to be more accurate than other commonly used inflation forecasting approaches, a major drawback from survey forecasts is that they are only available for a discrete set of forecast horizons, as respondents are asked about their expectations over predetermined time intervals. Clearly, it is beneficial to have judgemental expectations over arbitrary horizons, because it not only provides more insight on how the survey forecasts shift over the forecast horizon, but it also construes how economists react to changes in monetary policy. Moreover, in my forecasting exercise it is particularly useful to include such continuous judgemental forecasts, since the Survey of Professional Forecasters only records expectations over horizons up to four quarters. For this reason, I also examine the Aruoba Term Structure of Inflation Expectations from Borağan Aruoba (2020).

The Aruoba Term Structure of Inflation Expectations is a statistical term structure model of inflation expectations, which combines major surveys including the Survey of Professional Forecasters and the Blue Chip Economic Indicators and Blue Chip Financial Forecasts published by Wolters Kluwer Law and Business. Borağan Aruoba (2020) finds that the Aruoba Term Structure of Inflation Expectations, similar to other judgemental inflation forecasts, outperforms commonly used alternatives in terms of predictive accuracy. However, in their research, they not compare the predictive ability of the term structure model relative to the individual surveys that are being merged by the model. As I inspect both the Survey of Professional forecasters and the Aruoba Term Structure of Inflation Expectations in my forecast comparison, I can investigate whether combining multiple surveys leads to more accurate inflation predictions than considering the surveys on their own.

As mentioned before, recent developments in computational power and the current availability of large datasets led to the popularity of high-dimensional machine learning methods. This trend is also evident within the inflation forecasting research field. Goulet Coulombe et al. (2022) establish that the main advantages of machine learning models with respect to classical models is that they are data-driven and account for non-linearity. Earlier research on machine learning methods for the prediction of inflation primarily fixated on linear methods, e.g. Inoue and Kilian (2008) discover that incorporating bagging and linear shrinkage methods with a set of real economic activity measures leads to mean-squared prediction error reductions in forecasting

US CPI, when compared to univariate benchmarks. In addition, Medeiros and Mendes (2016) find that Least Absolute Shrinkage and Selection Operator (LASSO) based models produce superior US CPI forecasts than factor and AR benchmarks.

Currently, non-linear machine learning approaches are becoming more popular for the prediction of inflation rates. That is, Garcia, Medeiros and Vasconcelos (2017) exhibit that not only shrinkage and complete subset regression models, but also random forests perform well in the real-time forecasting of Brazilian inflation rates in a data-rich environment. Medeiros et al. (2021) show that the finding on the inflation forecasting performance of random forests are not particular for Brazilian inflation rates as they replicate this result for US inflation, and additionally, they obtain that the random forest model dominates among a wide-ranging set of machine learning methods. Furthermore, Goulet Coulombe et al. (2022) also find that random forests dominate autoregressive benchmarks in forecasting inflation.

Even though random forests were initially not constructed to fit time series, Medeiros et al. (2021) attribute the superior forecasting performance of the random forest to its particular variable selection method and ability to derive non-linear relations between inflation and lagged macroeconomic variables. Despite the existing papers provide evidence that the random forest is more accurate in forecasting inflation than simple benchmarks as well as other machine learning methods, the predictive ability of the random forest has not yet been compared to that of survey forecasts, VAR models nor models in gap form. Thus, I include the random forest in my comparative review in order to gain more insight on the forecast accuracy of the random forest relative to other effective inflation forecasting methods.

Another successful non-linear machine learning approach is that of Nakamura (2005), who implements neural networks to construct inflation projections over short horizons, which are more accurate than predictions derived from univariate autoregressive models. The accurate performance of neural networks in forecasting inflation rates has also been exhibited by McAdam and McNelis (2005) and Chen, Racine and Swanson (2001). Although these papers find significant accuracy gains over univariate autoregressive models, they only consider rather simple neural network specifications, which are not able to recognize the sequential patterns that time series display[1].

This suggests that there is more to gain in terms of forecasting accuracy by considering neural networks that are able to identify dependency over time. Almosova and Andresen (2023) propose an alike neural network, i.e., the long short-term memory recurrent neural network (LSTM), for the prediction of inflation rates and find that the model outperforms a simple fully connected neural network. Therefore, I also examine the LSTM framework in my research.

From the existing literature, one can conclude that many distinct methods have shown effectiveness for the prediction of inflation rates. Furthermore, proper inflation forecasting methods should be able to outperform random walk models and simple AR models, which are reasonable and commonly used benchmarks for the prediction of inflation. Finally, it is useful to conduct an extensive comparison among effective methods, rather than only comparing approaches separately to naive benchmarks, which often is the case in previous research.

[1]For an extensive review on the ability of capturing temporal dependency among the distinct neural network classes, I refer to Längkvist, Karlsson and Loutfi (2014).

3 Data

For this research I examine three measures of US inflation, namely inflation according to the Consumer Price Index (CPI), the Gross Domestic Product (GDP) deflator[2] and the Personal Consumption Expenditures (PCE) deflator. The CPI assesses the average change in prices of goods and services consumed by households, whereas the GNP and PCE deflator evaluate price changes in goods and services purchased by consumers, businesses and governmental institutions. I focus on inflation rates at the quarterly frequency, which are derived using the price level index data at the end of the quarters, over the period from 1948Q1 until 2022Q4.

To be specific, I consider real-time annualized quarterly inflation rates, which are computed as $\pi_t = 400\,(log(P_t) - log(P_{t-1}))$, where P_t indicates the underlying price level index at the end of quarter t. The real-time datasets on the price level indices are from the Federal Reserve Bank of Philadelphia and can be found on their website[3]. The quarterly vintages are seasonally adjusted and already consist of quarterly observations (except for the CPI vintages, which have monthly observations[4]). The vintages are collected in the middle of the quarter, i.e., on February 15, May 15, August 15, and November 15, such that each vintage consists of data up to one quarter prior to the quarter in which the vintage is collected.

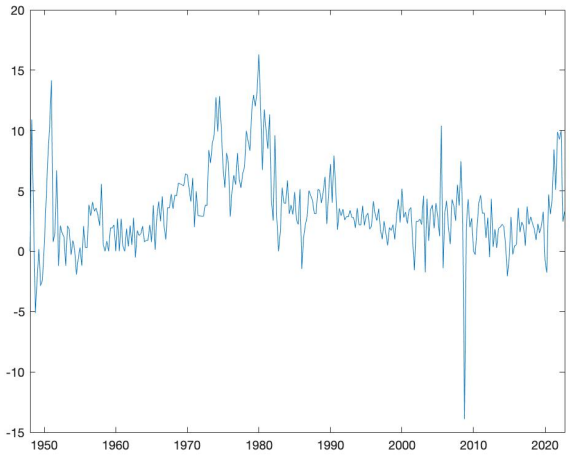
Figure 1 exhibits the quarterly inflation measures for the period 1948Q1 - 2022Q4. One should note that even though the inflation measures display similar patterns and tend to move in tandem, their dynamics still have differences due to their distinct compositions. For instance, the CPI inflation rate is consistently higher than the GDP and PCE deflator inflation rate, as the CPI index suffers from the renowned upward substitution bias (Noe and Furstenberg, 1972; Hamilton, 2001; Boskin, 2005). Thus, the performance of the forecasting models that I study in this research might differ across the inflation measures.

Nonetheless, for all three inflation measures displayed in Figure 1, I observe that the trend is evolving over time. In particular, the inflation rates are rising in the early 50s due to the aftermath of World War II and the Korean War, and are declining rates in the succeeding years. During the period between 1970 and the early 1980s, also known as the Great Inflation, inflation rates are increasing again, after which rates are declining once more. Moreover, negative spikes are observed in the period between 2008 and 2009, corresponding with the Great Recession. At the end of the sample, more upward spikes continue to appear, which are related to post-COVID issues in global supply chains and the Russian invasion of Ukraine. As discussed in Section 2, numerous papers including Clark and McCracken (2006) and Cogley, Primiceri and Sargent (2010) have shown that the potential to follow this slowly-varying trend in inflation rates strongly influences the performance of inflation forecasting methods. Since the slowly-varying trend feature is clearly evident in the sample that I investigate, it is likely that the most accurate forecasting approaches in my comparison study will be methods that explicitly model the inflation trend.

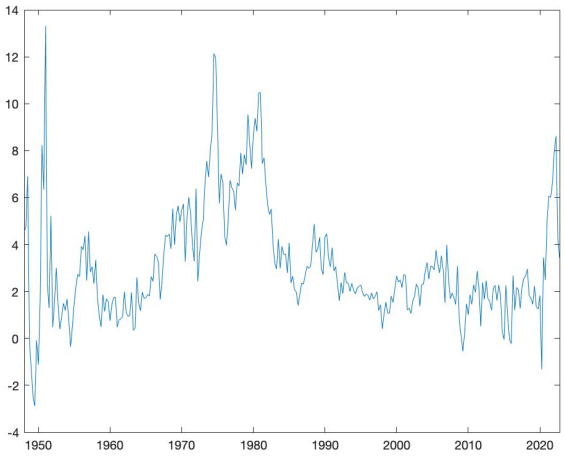
[2]The Gross National Product (GNP) before 1991Q4.

[3]`https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/cpi`, accessed on October 2023.

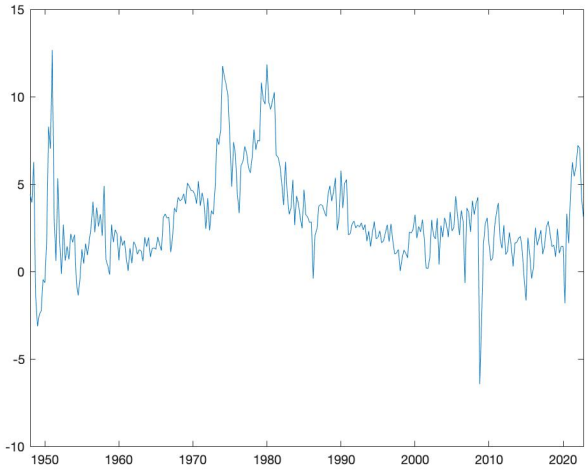
[4]I use the CPI values corresponding to the end months of the quarters. That is, the observations in March, June, September and December.



(a) CPI



(b) GDP deflator



(c) PCE deflator

Figure 1: Annualized inflation rates from 1948Q1 until 2022Q4 according to different price level indices (using 2023Q1 vintage data).

In addition, I note that the inflation rates appear to be more volatile in the period between 1965 and 1980 than in the subsequent period, which is in accordance with important implications in the academic literature (Stock and Watson, 2002; Sims and Zha, 2006; Stock and Watson, 2007). During the period after the Great Recession, the volatility of the inflation rates also seems to be increasing, which is in line with findings of Chan (2013). Hence, the inflation rates are more difficult to predict over these spells. As my forecast sample starts in 1997, only the latter interval will impact the results of my forecast exercise. Moreover, among the inflation forecasting models that I examine in this study, the stochastic volatility model is expected to provide the most accurate projections over these periods.

As some of the inflation forecasting models that I investigate incorporate data on alternative macroeconomic indicators, I also use other datasets. Firstly, I implement the FRED-QD macroeconomic database specifically designed for the empirical analysis of "big data" (again at the

quarterly frequency), which is provided by the Federal Reserve Bank of St. Louis and can be found on their website[5]. The database consists of 246 macroeconomic series (modified to be stationary) and the observations are updated in real-time through the FRED-QD database, already incorporating data changes and revisions, as this dataset is prepared by the data desk at the Federal Reserve Bank of St. Louis. Furthermore, I use data on economic activity (i.e., the unemployment rate), which are also collected from the Federal Reserve Bank of Philadelphia. Finally, I apply panel data of annualized continuously-compounded US government bond yields from Liu and Wu (2021).

4 Methodology

In this section, I describe the methods of my research. The general set-up for my forecasting experiment is described first. Subsequently, I discuss the various inflation forecasting methods that I examine in this study. Finally, I elaborate on the metrics being used to evaluate and compare the accuracy of the inflation forecasts.

My forecasting experiment has the following setting. The forecasts of the annualized quarterly inflation rates are made real-time using the vintage datasets from the Federal Reserve Bank of Philadelphia and the Federal Reserve Bank of St. Louis. Moreover, I predict from the middle month of each quarter, since the vintages from the Federal Reserve Bank of Philadelphia are collected in the middle of the quarter (i.e. on February 15, May 15, August 15, and November 15). Thus, solely data that are available at the middle of the quarter, videlicet the data up to the previous quarter, are incorporated in the inflation predictions. Predictions are made over multiple horizons. That is, I assess predictions for the current quarter (corresponding to forecast horizon $h = 0$) and over 1, 4 and 8 quarters, such that the forecasting performance over both short- and longer-term horizons can be evaluated. The resulting out-of-sample projections of the quarterly inflation rates are constructed for the period from 1997Q1 until 2022Q4 and are based upon an expanding window of observations starting from the first quarter of 1948.

I now move on to the inflation forecasting approaches I study in this research. I make a distinction between three streams of inflation forecasting approaches existing in the current academic literature, that is, judgmental forecasts, classical statistical models, and machine learning methods.

4.1 Judgemental forecasts

As already mentioned in Section 2, previous research shows that subjective forecasts outperform many other inflation forecasting methods (Grant and Thomas, 1999; Thomas Jr, 1999; Mehra, 2002; Ang, Bekaert and Wei, 2007; Faust and Wright, 2013). Therefore, I firstly examine real-time subjective forecasts, namely survey forecasts. For this purpose, I implement the Survey of Professional Forecasters from the Federal Reserve Bank of Philadelphia and the Aruoba Term Structure of Inflation Expectations from Borağan Aruoba (2020).

[5] `https://research.stlouisfed.org/econ/mccracken/fred-databases/`, accessed on October 2023.

4.1.1 Survey of Professional Forecasters

Each quarter, the panelists of the Survey of Professional Forecasters (SPF) are asked about their expectations on macroeconomic indicators over horizons of 0,1,2,3 and 4 quarters. Among the macroeconomic variables that the panelists are asked to predict in the questionnaire are the CPI inflation rate and the GNP/GDP deflator[6]. The survey is conducted in the middle month of each quarter, thus corresponding to the timing at which I construct the forecasts in my study. I apply the SPF by taking the mean forecasts[7] of the variables of interest, which are published on the website of the Federal Reserve Bank of Philadelphia[8]. Note that the (mean) SPF forecasts of the GNP/GDP deflator have to be transformed to inflation rates in accordance with the same formula used to construct the inflation measures in this research (i.e. by taking 400 times the first differences of the logs).

4.1.2 Aruoba Term Structure of Inflation Expectations

Since the SPF only records expectations over horizons up to 4 quarters and I also aim to assess the performance of judgemental forecasts over a horizon of 8 quarters, I examine the Aruoba Term Structure of Inflation Expectations (ATSIX) from Borağan Aruoba (2020). The ATSIX is a smooth and continuous curve representing 1 to 40 quarters ahead (CPI) inflation forecasts, similar to the way a yield curve describes the term structure of interest rates. Borağan Aruoba (2020) constructs the ATSIX by implementing a statistical factor model that merges major surveys, i.e., the Blue Chip Economic Indicators and Blue Chip Financial Forecasts published by Wolters Kluwer Law and Business together with the SPF. From this factor model a term structure of inflation expectations is derived. That is, inflation forecasts over any arbitrary horizon can be derived from the model.
In particular, the Borağan Aruoba (2020) specifies the ATSIX as a state-space model with the following measurement equation

$$\hat{\pi}_t(h) = L_t + \Big(\frac{1-e^{-\lambda h}}{\lambda h}\Big) S_t + \Big(\frac{1-e^{-\lambda h}}{\lambda h} - e^{-\lambda h}\Big) C_t + \epsilon_t, \tag{1}$$

where $\hat{\pi}_t(h)$ represents the subjective inflation expectation over horizon h, λ is a parameter controlling the factor loadings for all horizons and ϵ_t indicates the measurement error. Moreover, L_t represents long-term inflation expectations, S_t denotes the differential between long- and short-term expectations, and C_t designates medium-term expectations that are higher or lower than long- and short-term expectations. Subsequently, the three latent factors are imposed to follow independent AR(3) processes in the state equations:

[6]Projections of the PCE inflation rate were only included from the survey in 2007Q1 onward and are therefore not applicable to my forecasting exercise.

[7]I have also considered the median forecasts, but these predictions performed slightly worse in terms of accuracy than the mean forecasts.

[8]`https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/survey-of-professional-forecasters`, accessed on October 2023.

$$L_t = \mu_1 + \rho_{11}\left(L_{t-1} - \mu_1\right) + \rho_{12}\left(L_{t-2} - \mu_1\right) + \rho_{13}\left(L_{t-3} - \mu_1\right) + \eta_{1,t}$$
$$S_t = \mu_2 + \rho_{21}\left(S_{t-1} - \mu_2\right) + \rho_{22}\left(S_{t-2} - \mu_2\right) + \rho_{23}\left(S_{t-3} - \mu_2\right) + \eta_{2,t}$$
$$C_t = \mu_3 + \rho_{31}\left(C_{t-1} - \mu_3\right) + \rho_{32}\left(C_{t-2} - \mu_3\right) + \rho_{33}\left(C_{t-3} - \mu_3\right) + \eta_{3,t},$$

where $\eta_{i,t}$ is normally distributed with mean zero and variance σ_i^2, and independent of $\eta_{j,t}$ for $i \neq j$. Since the model is a state-space representation, its estimation and inference is attained by applying the Kalman filter and smoother (Durbin and Koopman, 2012).
The implied inflation forecasts of the ATSIX can be derived by the equations above using the most recent factor estimates. Nonetheless, the ATSIX forecasts are directly accessible on the website of the Federal Reserve Bank of Philadelphia[9].

4.2 Classical statistical models

Next, I examine commonly used statistical specifications for the prediction of inflation rates. This set of models consists of a random walk, AR models, Phillips curve models, a stochastic volatility model and a VAR specification. Below, I describe each of these models separately, starting with the models that serve as benchmarks in my forecasting exercise.

4.2.1 Benchmark models

The first benchmark in my forecasting experiment is the random walk (RW) model. The random walk model is a reasonable and generally employed benchmark model within the inflation forecasting research field (Medeiros et al., 2021; Groen, Paap and Ravazzolo, 2013; Canova, 2007; Atkeson, Ohanian et al., 2001). The random walk model has the following specification for the inflation forecast made at quarter T over horizon $h = 0, 1, 4, 8$

$$\hat{\pi}_{T+h}^{RW} = \pi_{T-1},$$

where π_{T-1} indicates the inflation rate in quarter $T-1$.
In addition, I implement the AR model of order p as a benchmark in my inflation forecasting study, following existing research on the prediction of inflation rates encompassing, but not limited to Medeiros et al. (2021), Choudhary and Haider (2012), Ang, Bekaert and Wei (2007) and Bos, Franses and Ooms (2002). The AR(p) forecasts are obtained as follows. Beyond each horizon h, I estimate the regression

$$\pi_{t+h} = \phi_0 + \phi_1\pi_{t-1} + ... + \phi_p\pi_{t-p} + \epsilon_{t+h} \quad for\ \ t = 1, ..., T-1-h.$$

Subsequently, I use the OLS parameter estimates to directly compute the h-quarters ahead forecast at current quarter T given by

$$\hat{\pi}_{T+h}^{AR} = \hat{\phi}_0 + \hat{\phi}_1\pi_{T-1} + ... + \hat{\phi}_p\pi_{T-p}.$$

[9] `https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/atsix`, accessed on October 2023.

For each inflation measure and horizon h separately, I determine the order p according to the Bayesian information criterion (BIC), as Marcellino, Stock and Watson (2006) exhibit that this criterion is most appropriate for the order selection in direct forecasting exercises. Furthermore, I also use the BIC to determine p in the following specifications that incorporate lags, again for each inflation measure and horizon individually.

4.2.2 AR-gap model

As exhibited in Section 3, the trend in inflation rates tends to vary over time. To account for this feature, one can decompose inflation into a stochastic trend assumed to follow a random walk, denoted by $\bar{\pi}_t$, and a transient component, which describes the transitory deviations of realized inflation with respect to trend inflation (Cogley, Primiceri and Sargent, 2010; Faust and Wright, 2013; Morley, Piger and Rasche, 2015). This latter component is called the inflation gap, with formal definition

$$g_t = \pi_t - \bar{\pi}_t.$$

There are multiple instruments suitable to measure the trend in inflation. Some papers along with Cogley, Primiceri and Sargent (2010) and Morley, Piger and Rasche (2015) use a central bank's (e.g., the Federal Reserve's) long-run inflation target for this purpose, whereas others including Clark (2011) and Faust and Wright (2013) attach the inflation trend to long-run survey expectations. I reckon the latter approach, since I also investigate survey forecasts separately and thus can discover in which framework the use of survey forecasts is more beneficial. Specifically, I use the (mean) 10-year-ahead inflation forecasts from the SPF[10], which are again accessible on the website of the Philadelphia Fed's website[11]. As there are no long-run SPF inflation forecasts available before the third quarter of 1991, I follow Faust and Wright (2013) and Clark (2011) by using exponential smoothing of real-time inflation rates as a proxy for $\bar{\pi}_t$ over these periods. That is, for this period I set $\bar{\pi}_t$ equal to π_t^{ES} which is computed as

$$\pi_t^{ES} = \alpha \pi_{t-1}^{ES} + (1-\alpha)\pi_t,$$

where α indicates the smoothing parameter, which I set equal to 0.95 implying a slowly-varying trend, once more adhering to Faust and Wright (2013). Moreover, I set π_0^{ES} equal to the first observation of the sample (i.e., the observation for 1948Q1).
Various researches have illustrated the benefits of forecasting inflation by modeling the inflation gap (Cogley, Primiceri and Sargent, 2010; Clark, 2011; Morley, Piger and Rasche, 2015). A straightforward approach to implement the inflation gap in constructing inflation forecasts is to fit an AR-model to the inflation gap. That is, for each horizon h, I run the regression

$$g_{t+h} = \zeta_0 + \zeta_1 g_{t-1} + ... + \zeta_p g_{t-p} + \epsilon_{t+h} \quad for \;\; t = 1, ..., T-1-h.$$

Next, I use the parameter estimates to iterate the h-quarters ahead projection of the inflation

[10]The long-run SPF predictions are only made for CPI and PCE deflator inflation. Hence, I implement the PCE deflator inflation forecasts as trend measure for GDP/GNP deflator inflation.
[11]`https://www.philadelphiafed.org/surveys-and-data/data-files`, accessed on October 2023.

gap from the current quarter T:

$$\hat{g}^{AR}_{T+h} = \hat{\zeta}_0 + \hat{\zeta}_1 g_{T-1} + ... + \hat{\zeta}_p g_{T-p},$$

Lastly, I add the current trend observation to the inflation gap projection, such that the resulting inflation forecast is given by

$$\hat{\pi}^{AR-gap}_{T+h} = \hat{g}^{AR}_{T+h} + \bar{\pi}_{T-1}. \tag{2}$$

I label these predictions as the AR-gap forecasts.

4.2.3 Phillips curve models

Aforementioned, the Phillips curve is a commonly used tool in forecasting inflation (Stock and Watson, 1999; Brayton, Roberts and Williams, 1999; Atkeson, Ohanian et al., 2001; Ang, Bekaert and Wei, 2007; Stock and Watson, 2008; Groen, Paap and Ravazzolo, 2013). Since there are many different variants and extensions on Phillips curve based models, I consider a selection of two distinct methods within the Phillips curve framework.
As a starting point, I generate forecasts by fitting the generalized Phillips curve directly to the inflation rates. Thus, for each h, I estimate the Phillips curve

$$\pi_{t+h} = \delta_0 + \delta_1 \pi_{t-1} + ... + \delta_p \pi_{t-p} + \theta a_{t-1} + \epsilon_{t+h} \quad for \;\; t = 1, ..., T-1-h,$$

where a_{t-1} is an economic activity measure at quarter $t-1$, for which I consider real-time unemployment rate data[12]. Then the Phillips curve (PC) forecasts are computed as

$$\hat{\pi}^{PC}_{T+h} = \hat{\delta}_0 + \hat{\delta}_1 \pi_{T-1} + ... + \hat{\delta}_p \pi_{T-p} + \hat{\theta} a_{T-1}.$$

Additionally, I investigate the approach taken by Faust and Wright (2013) and Stock and Watson (2010), where the Phillips curve is fitted to the inflation gap discussed in the previous segment, instead of inflation itself. Accordingly, I estimate

$$g_{t+h} = \xi_0 + \xi_1 g_{t-1} + ... + \xi_p g_{t-p} + \nu a_{t-1} + \epsilon_{t+h} \quad for \;\; t = 1, ..., T-1-h,$$

for each h and subsequently calculate the inflation gap forecast

$$\hat{g}^{PC}_{T+h} = \hat{\xi}_0 + \hat{\xi}_1 g_{T-1} + ... + \hat{\xi}_p g_{T-p} + \hat{\nu} a_{T-1}.$$

Again, the implied inflation forecast is obtained by adding the current trend inflation, $\bar{\pi}_{T-1}$, to the inflation gap projection, similar as in equation 2. I label the resulting projections as PC-gap forecasts.

[12] I have also examined implementing real-time output and industrial production growth data as economic activity measures, but these approaches performed rather worse in terms of predictive accuracy.

4.2.4 Unobserved component stochastic volatility model

The unobserved component stochastic volatility (UCSV) model has gained traction in the inflation forecasting literature since its establishment by Stock and Watson (2007), who find that the specification performs well in terms of prediction accuracy. Namely, Cogley, Primiceri and Sargent (2010), Chan (2013), Kim, Manopimoke and Nelson (2014), Cecchetti et al. (2017) and Mertens and Nason (2020) all implement variations of the UCSV model. As there are many different adaptations of the UCSV model and to avoid increased computational complexity relative to the other inflation forecasting methods that I consider in this study, I focus on the primary UCSV model.

The univariate UCSV model has the following specification

$$\pi_t = \bar{\pi}_t + \eta_t, \tag{3}$$

$$\bar{\pi}_t = \bar{\pi}_{t-1} + \chi_t, \tag{4}$$

where $\eta_t \overset{\text{iid}}{\sim} \text{N}(0,\sigma^2_{\eta,t})$ and $\chi_t \overset{\text{iid}}{\sim} \text{N}(0,\sigma^2_{\chi,t})$. Furthermore, the error terms are assumed to follow stochastic volatility processes

$$ln(\sigma^2_{\eta,t}) = ln(\sigma^2_{\eta,t-1}) + \psi_{\eta,t}, \tag{5}$$

$$ln(\sigma^2_{\chi,t}) = ln(\sigma^2_{\chi,t-1}) + \psi_{\chi,t}, \tag{6}$$

where $(\psi_{\eta,t},\ \psi_{\chi,t})' \overset{\text{iid}}{\sim} \text{N}(0,\ \gamma I_2)$ and γ is a parameter that determines the smoothness of the stochastic volatility process, which I set equal to 0.2 following Stock and Watson (2007). Thus, the only parameter of the UCSV model is fixed, whereas the remaining components of the model are all observed.

The UCSV forecasts are derived as follows. I simulate estimates of $\psi_{\eta,T}$ and $\psi_{\chi,T}$, which I substitute in equation 5 and 6, respectively, in order to obtain estimates of $\sigma^2_{\eta,T}$ and $\sigma^2_{\chi,T}$. Subsequently, using the estimates of $\sigma^2_{\eta,T}$ and $\sigma^2_{\chi,T}$, I generate simulation estimates of η_T and χ_T. Finally, the UCSV forecast $\hat{\pi}^{UCSV}_{T+h}$ is acquired through equation 3 and 4. The resulting projection is the filtered projection of the current inflation trend $\bar{\pi}_{T-1}$.

4.2.5 Term structure VAR model

As previously stated, vector autoregressive specifications in which inflation is linked to the term structure of government bond yields are also known to provide accurate inflation projections (Sims, 1993; Cogley and Sargent, 2001; Diebold and Li, 2006; Primiceri, 2005; Joslin, Priebsch and Singleton, 2014). Therefore, I also examine the term structure VAR forecasts from Faust and Wright (2013), in which the dynamic Nelson-Siegel yield curve from Diebold and Li (2006) is extended to construct inflation forecasts.

The term structure VAR predictions are constructed in the following manner. Firstly, I consider the dynamic Nelson-Siegel yield curve formulated as

$$y_t(\tau) = \beta_{1,t} + \beta_{2,t}\Big(\frac{1-e^{-\lambda\tau}}{\lambda\tau}\Big) + \beta_{3,t}\Big(\frac{1-e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}\Big) + \epsilon_t(\tau),$$

where $y_t(\tau)$ represents the yield to maturity on a government bond with maturity τ at time t and the dynamic factors $\beta_{1,t}$, $\beta_{2,t}$ and $\beta_{3,t}$ can be interpret as the level, (minus) slope and curvature, respectively. Furthermore, $\epsilon_t(\tau)$ denotes the error term assumed to have mean zero and a variance that is independent over time and across maturities. Similar as in the ATSIX specification from equation 1, λ controls the factor loadings by determining both the rate at which the loading on β_{2t} converges to zero and the optimal maturity for the loading on β_{3t} (Van Dijk et al., 2014), and is fixed at 0.0609 following Diebold and Li (2006).
I fit the dynamic Nelson-Siegel yield curve to annualized continuously-compounded US government bond yields, which are obtained from the dataset of Liu and Wu (2021). Thereupon, I fit a VAR(1) to the dynamic Nelson-Siegel factors together with the inflation gap and the unemployment rate:

$$\begin{bmatrix} g_t \\ \beta_{1,t} \\ \beta_{2,t} \\ \beta_{3,t} \\ u_t \end{bmatrix} = c + A \begin{bmatrix} g_{t-1} \\ \beta_{1,t-1} \\ \beta_{2,t-1} \\ \beta_{3,t-1} \\ u_{t-1} \end{bmatrix} + e_t \quad for \;\; t = 1, ..., T-1-h,$$

where u_t indicates the unemployment rate at quarter t, c is a vector of length 5 representing the model intercept, A is $5\times$ 5 coefficient matrix and e_t is a vector of length 5 containing the error terms. With the VAR(1) specification from above, I derive the inflation gap forecast $\hat{g}^{VAR}_{T+h}$ by iterating the equation forward over the desired horizon h and subsequently obtain the implicit h-quarters-ahead inflation forecast analogously as in equation 2.

4.3 Machine learning methods

Currently, non-linear machine learning methods are emerging in the inflation forecasting research field, as such models can provide precise predictions, which is attributed to their ability to handle high-dimensional datasets (Almosova and Andresen, 2023; Medeiros et al., 2021; Garcia, Medeiros and Vasconcelos, 2017; Nakamura, 2005). Accordingly, I also include inflation forecasting methods based non-linear machine learning models in my forecast comparison. In particular, I study a random forest model and a long short-term memory recurrent neural network.

4.3.1 Random forest

Random forests, which were first designed by Breiman (2001), diminish the variance of regression trees through bootstrap aggregation (bagging) of random regression trees. Regression trees are non-parametric supervised learning models, which estimate non-linear relationships between target and explanatory variables by recursive binary partitioning of the covariate space. That is, a single regression tree has the structure of a binary decision tree composed by split nodes and terminal nodes (which are also referred to as leaf nodes). To further explain how regression trees operate, I consider the example provided by Medeiros et al. (2021), which is illustrated in Figure 2.

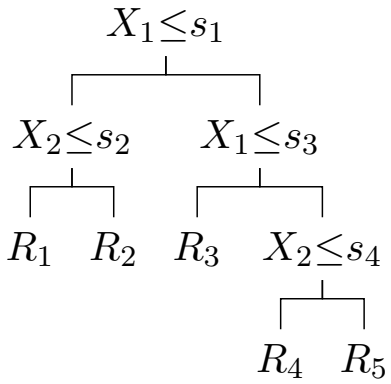


Figure 2: The structure of a regression tree. Reconstruction of Figure 1 from Medeiros et al. (2021).

In this case, there are two explanatory variables X_1 and X_2. As shown, the covariate space is partitioned into discrete regions R_i with i= 1,2,...,5, by a set of rectangular hyperplanes, where each hyperplane splits one of the explanatory variables. In each of the regions, the relationship between the explanatory variables and the dependent variable is described by a separate model, such that each region can be interpreted as a different regime.

Thus, random forests aggregate an extensive set of regression trees that are based on random samples of observations and consider random subsets of regressors for each possible split node, in order to lower variance and proneness to overfitting with respect to individual regression trees. The forecasts of all individual trees are averaged to derive the final forecast.

I construct inflation forecasts by implementing the random forest (RF) methodology in the following manner. Firstly, I train the random forest regressor by fitting the model to all data available at the current quarter T. Particularly, the training of the random forest is given by the equation

$$\pi_{t+h} = f(x_{t-1}; \kappa) \quad for \;\; t = 1, ..., T - 1 - h,$$

where $f(\cdot)$ denotes the random forest regressor and x_{t-1} is a vector consisting of a large set of macroeconomic indicators including various inflation measures, which are obtained from the real-time FRED-QD database of the Federal Reserve Bank of St. Louis[13], specifically designed for the empirical analysis of "big data". I do not apply a variable selection on the FRED-QD dataset prior to fitting the random regressor to the data, as the random forest aggregates regression trees that already select the most informative variables to split the data and accordingly less important features are not chosen. Furthermore, κ represents the hyperparameter vector[14].

After obtaining the fitted random forest model $\hat{f}(\cdot)$, I derive h-quarters ahead inflation forecasts using the latest observations on the predictors:

[13]The database is publicly accessible at `https://research.stlouisfed.org/econ/mccracken/fred-databases/`, accessed on October 2023.

[14]The hyperparameters of the random forest model consist of the maximum tree depth, maximum number of features, maximum number of leaf nodes, maximum sample size, minimum number of samples necessary for a split, minimum number of samples per leaf, number of trees, bootstrap indicator and the criterion to assess the quality of a split.

$$\hat{\pi}_{T+h}^{RF} = \hat{f}(x_{T-1}; \kappa).$$

The performance of the random forest model heavily relies on the selection of the hyperparameters stored in κ. Since I fixate on constructing out-of-sample projections, I implement cross-validation to tune the hyperparameters of the random forest model. Specifically, I use blocked cross-validation proposed by Snijders (1988), which is the conventional procedure when dealing with time series (Bergmeir and Benítez, 2012). This implies that the validation set only consists of observations after the final observation of the training set, such that the chronological order of the time series is maintained. The tuning process is initialized with a randomized grid, which is followed by a second search on a grid that is closer around the optimal hyperparameters of the first search. To reduce computational time, I only update the hyperparameters after making four consecutive forecasts, which corresponds to an annual frequency.

4.3.2 Long short-term memory recurrent neural network

The LSTM is a specific kind of recurrent neural network (RNN) introduced by Hochreiter and Schmidhuber (1997). A RNN is a particular type of artificial neural network that implements data sequentially through its recurrent structure. That is, the RNN estimates forecasts by taking in lags of data serially while updating the prediction, such that the intermediary output, the so-called state, is employed as source for the next updating step.
Figure 3 illustrates the structure of the RNN. The RNN admits a single lag of explanatory data stored in vector x_{t-2} together with the state denoted by the horizontal arrows, after which the same RNN receives the next lag of data along with the state resulting from the previous step. This process is repeated until the final prediction is obtained. Since the lags of explanatory variables are considered in chronological order, more recent observations tend to have more influence on the final projection.

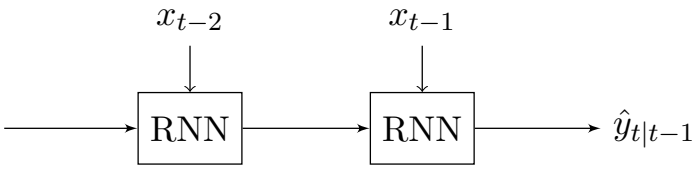


Figure 3: The structure of a recurrent neural network. Reproduction of Figure 4 from Almosova and Andresen (2023).

The LSTM has a similar recursive structure as the general RNN, but is distinguishable due to its embodiment of so-called "input", "output" and "forget gates". The gates facilitate the network's ability to filter both the state and lagged data input for the next recurrent propagation, such that the network itself determines which part of the input should be memorized and which part can be forgotten. Figure 4 depicts the internal structure of a single LSTM cell at time step t. The LSTM output at time t is denoted by h_t and c_t indicates the state at t, which carries the memory on the past. Moreover, σ and $tanh$ denote the gates, which on their own are smaller neural networks with sigmoid or hyperbolic tangent activation functions. As can be seen in the left bottom corner of the figure, the gates filter the output of the previous time step h_{t-1} and

the lagged explanatory variables stored in the vector x_{t-1} , in order to derive the updated state c_t. That is, the gates regulate how c_{t-1} should be adjusted. Finally, based on c_t the LSTM cell produces output h_t, from which a projection is derived.

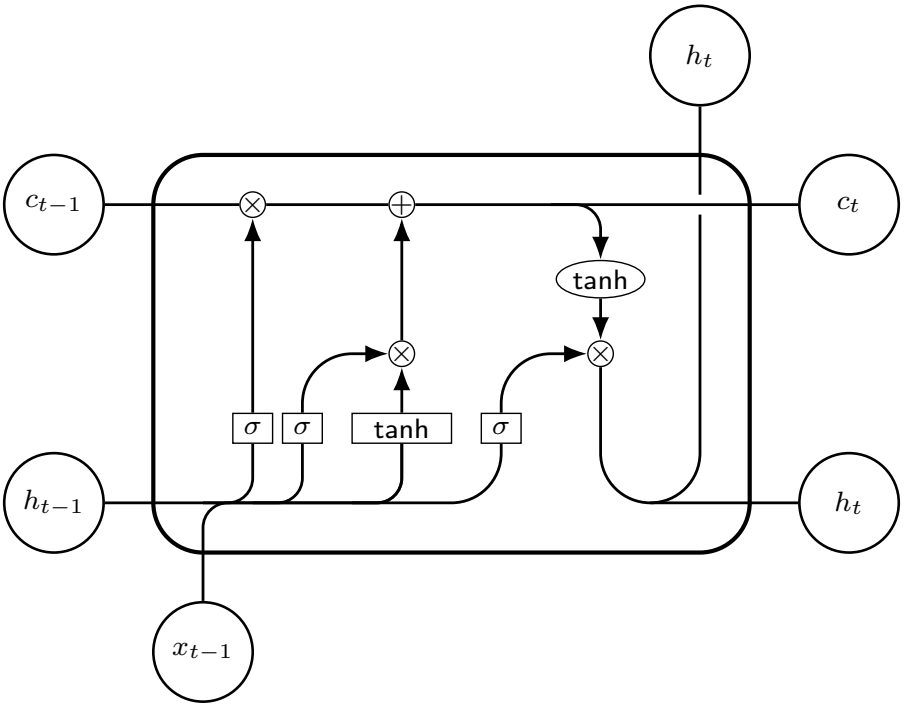


Figure 4: The structure of a single LSTM cell. Reproduction of Figure 6 from Almosova and Andresen (2023).

Similar as for random forests, the performance of the LSTM relies on the selected hyperparameters of the model. However, due to the complex structure of the LSTM an extensive grid search is not feasible, as this would dramatically increase computational time. Hence, I initialize hyperparameters in accordance with previous work and manually adjust these in a process of trial and error while optimizing the validation loss in the training sets. From this process, I attain the following settings for the LSTM specification. I use one LSTM layer with 100 neurons, followed by a dropout layer that randomly filters away 30% of the input units to prevent overfitting on the training set. The dropout layer is followed by a fully connected dense layer consisting of 32 neurons, which summarizes the information passed through the previous layers. Then, one more dropout layer is added, again filtering away 30%, after which the output layer follows.

Furthermore, I implement the Adam optimizer and both the *ReLU* and *tanh* activation functions for the LSTM layer and dense layer, respectively. To determine the number of epochs (i.e., the number of times the training data is passed around the network), I use an early stopping tweak, where the maximum number of epochs is set equal to 50 and the process terminates when the validation loss has increased over 14 consecutive epochs. The model derived in the epoch with the lowest validation loss is selected to construct the final inflation prediction.

The resulting h-quarters-ahead LSTM forecast can be represented by the following equation

$$\hat{\pi}_{T+h}^{LSTM} = \hat{l}(\pi_{T-1}, x_{T-1}; \omega),$$

where $\hat{l}(\cdot)$ portrays the LSTM model and ω indicates the hyperparameter vector. Again, the explanatory variables stored in the vector x_{T-1} are obtained from the FRED-QD dataset. However, for the LSTM model it is necessary to incorporate a variable selection method prior to fitting the neural network, as there is no implicit feature selection within the LSTM framework. Therefore, I implement feature selection approach based on the Information Gain from Kent (1983), which is a commonly used variable selection method in the machine learning research field (Jadhav, He and Jenkins, 2018; Azhagusundari, Thanamani et al., 2013; Lei, 2012). Specifically, I select features of the FRED-QD dataset that have an Information Gain greater than 0.10 and subsequently feed the resulting selection of variables to the LSTM neural network. Moreover, I execute the feature selection for each forecast individually, such that the set of input variables evolves over time, taking account of time-varying explanatory power of features on inflation.

4.4 Forecast evaluation metrics

I assess and compare the predictive ability of the different inflation forecasting approaches by various metrics. Firstly, I examine the root mean squared prediction errors (RMSPE) of the inflation forecasts to assess their predictive accuracy, which is defined as follows

$$RMSPE = \sqrt{\frac{1}{T_N - T_0 + 1} \sum_{t=T_0}^{T_N} \hat{e}_t^2},$$

where T_0 and T_N respectively denote the index of the starting and end quarter of the forecast sample, and the forecast error $\hat{e}_t$ is quantified as actual minus forecast value (viz. $\pi_t - \hat{\pi}_t$).
Moreover, I take into account another accuracy measure, namely the mean absolute prediction error (MAPE), as the RMSPE metric could be influenced by extreme values, whereas the MAPE is less sensitive to extreme values. Thus, the MAPE metric confirms whether found results on forecast accuracy are robust to the accuracy measure and not prompted by a few substantial forecast errors. The MAPE is given by

$$MAPE = \frac{1}{T_N - T_0 + 1} \sum_{t=T_0}^{T_N} |\hat{e}_t|.$$

I report the RMSPEs and MAPEs of the different inflation forecasting methods relative to the benchmark models. That is, I present both the RMSPE and MAPE as fraction of their equivalent found for a benchmark model. Consequently, values less than one correspond with an forecast accuracy gain relative to the benchmark.
To test whether found accuracy disparities among models are statistically significant, it is customary to implement the Diebold-Mariano statistic from Diebold and Mariano (1995). However, since several forecasting models in this comparison study are nested and all models are estimated over an expanding window, the asymptotic distribution of the Diebold-Mariano statistic is affected, which is discussed in more detail in Clark and McCracken (2015), Clark and McCracken (2013), Gneiting and Ranjan (2011), and Amisano and Giacomini (2007). To overcome the econometric complexities caused by the presence of nested models and the use of expanding

windows, I follow the approach from Clark and McCracken (2013), which is also adopted by Faust and Wright (2013) and Groen, Paap and Ravazzolo (2013), who based on a Monte Carlo simulation find that comparing the small sample correction of the Diebold-Mariano statistic from Harvey, Leybourne and Newbold (1997) with standard normal critical values provides a proper sized test of the null of equal finite-sample prediction accuracy in both cases where models are nested and non-nested, as well as being estimated on an expanded window.
The augmented Diebold-Mariano (DM) test with the null hypothesis of equal finite-sample prediction accuracy is given by

$$DM = \sqrt{T_N - T_0 + 1}\,\frac{\bar{d}}{\sqrt{\gamma_0 + 2\sum_{i=1}^{h-1}\gamma_i}},$$

where $\bar{d}$ denotes the average loss-differential between two forecasting models and γ_i is the auto-covariance of the loss-differential at lag i. I implement the quadratic loss function, such that the loss-differential for the forecasts at quarter t is defined as

$$d_t = \hat{e}_{1,t}^2 + \hat{e}_{2,t}^2,$$

where $\hat{e}_{1,t}$ and $\hat{e}_{2,t}$ are the forecast errors belonging to the two models that are being compared[15]. Under the null hypothesis $H_0 : \mathbb{E}[d_t] = 0$, the DM test statistic is standard normally distributed. Since I perform a separate DM test for the comparison of each forecasting method with one of the two benchmark models, the number of tests rises and hence the probability of the occurrence of Type-I errors (i.e., having false positives) increases. To correct for the higher probability of Type-I errors when making multiple comparisons, researchers often adjust the critical values of statistical tests, e.g., by implementing the Bonferroni correction. However, such corrections are known to be overly conservative, especially when operating a large number of tests, and thus regarded as unnecessary (Rothman, 1990; Perneger, 1998; Rutter, 2008). Moreover, as argued by Barnett et al. (2022), decreasing the probability of a Type-I error, causes the risk of a Type-II error (i.e., having a false negative) to rise, which is just as severe as the former error. Therefore, I do not apply an adjustment method that accounts for the multiple comparisons in my study, following the vast majority of prior research comparing multiple inflation forecasting methods[16].
Apart from the forecast accuracy, I also inspect the bias of the distinct forecasting approaches. Following the notation from above, the bias of the forecasts over a selected horizon is defined as

$$bias = \frac{1}{T_N - T_0 + 1}\sum_{t=T_0}^{T_N}\hat{e}_t.$$

One has to note that inflation measures are repeatedly revised, such that it is unclear which value should be taken as the actual. Whereas revisions to CPI are rather trivial compared to other inflation measures, the PCE and GDP deflator inflation measures are more frequently revised. Therefore, I follow Tulip (2009), Faust and Wright (2009), and Faust and Wright (2013) by measuring actual realized inflation using data from the Federal Reserve Bank of Philadelphia's

[15] To obtain the DM test statistics, I implement the ready-to-use MATLAB function from Ibisevic (2024).

[16] Instead, I shed light on past findings confirming the results of my comparison and challenge forthcoming research to investigate newfound insights of my research.

real-time dataset two quarters following the quarter in question.

5 Results

In this section, I examine the results of my forecasting exercise. Table 1 provides an overview of the inflation forecasting methods that I consider in this study, together with their respective labels.

Table 1: Overview of the alternative inflation forecasting methods.

Method	**Label**
Classical statistical methods:	
Random walk model	RW
Autoregressive model	AR
Autoregressive model fitted to inflation gap	AR-gap
Phillips curve model	PC
Phillips curve model fitted to inflation gap	PC-gap
Unobserved component stochastic volatility model	UCSV
Term structure VAR model fitted to inflation gap	Term structure VAR
Judgemental forecasts:	
Survey of Professional Forecasters	SPF
Aruoba Term Structure of Inflation Expectations	ATSIX
Machine learning methods:	
Random forest	RF
Long short-term memory recurrent neural network	LSTM

This section proceeds as follows. First, I discuss the main findings on the accuracy of the alternative forecasting methods. Next, I assess the robustness of the found results on the accuracy of the different forecasting approaches. Thereafter, I examine the bias of the forecasts. Ultimately, I consider the implications on the explanatory power of the predictors of inflation.

5.1 Forecast accuracy

Table 2 shows the RMSPEs in percentages for the alternative forecasts of the three inflation measures relative to the random walk benchmark. Since the RMSPEs are reported as fractions of their equivalent found for the RW model, values less than one imply a gain in forecast accuracy with respect to the RW model and thus smaller values signify more accurate forecasts. Moreover, the asterisks indicate whether found accuracy gains relative to the RW model are significant according to the augmented Diebold-Mariano test discussed in Section 4.4.

Table 2: Relative root mean square prediction errors of the alternative inflation forecasts with respect to the random walk forecast.

Panel A: CPI inflation

	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
AR	0.81**	0.79**	0.80**	0.79*
AR-gap	0.81**	0.78**	0.78**	0.75*
PC	0.86**	0.82**	0.80**	0.80*
PC-gap	0.85**	0.81**	0.79**	0.75*
UCSV	0.76**	0.74**	0.78**	0.73*
Term structure VAR	0.80**	0.76**	0.83*	0.77*
SPF	**0.63****	0.74**	0.78**	
ATSIX		**0.73****	**0.77****	**0.72****
RF	0.80**	0.78**	0.83**	0.78*
LSTM	0.77**	0.76**	0.81**	0.75*

Panel B: GDP deflator inflation

	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
AR	0.87***	0.90**	0.96	0.90
AR-gap	0.88***	0.91*	0.91***	**0.85****
PC	0.97	0.98	0.96	0.92
PC-gap	0.97	0.98	0.93**	**0.85****
UCSV	1.15	1.11	0.95	**0.85***
Term structure VAR	1.02	1.05	1.14	0.98
SPF	**0.82***	**0.86***	**0.82****	
RF	0.98	1.00	0.95	**0.85****
LSTM	1.18	1.13	1.03	0.90*

Panel C: PCE deflator inflation

	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
AR	**0.88*****	0.85***	0.88*	0.85
AR-gap	0.90**	0.86***	**0.83****	**0.79***
PC	0.96	0.90*	0.88*	0.87
PC-gap	0.96	0.91*	0.84**	**0.79***
UCSV	0.95	**0.84****	0.85*	**0.79***
Term structure VAR	0.93*	0.86**	0.92	0.87
RF	1.03	0.88*	0.88*	0.84*
LSTM	1.01	0.87*	0.88*	0.84*

This table discloses the RMSPE of the different forecasts on three measures of inflation, across forecasting horizons of 0, 1, 4 and 8 quarters. The forecasts are constructed for the period from 1997Q1 until 2022Q4 and are based upon an expanding window of observations with data running back to 1948Q1. The RMSPEs are reported as fractions of the RMSPEs of the random walk model. The asterisks indicate the significance level of found accuracy gains with respect to the random walk model; one, two and three asterisks correspond to an accuracy improvement at the 10%, 5% and 1% significance level, respectively. The bold values denote the lowest RMSPEs across each forecast horizon.

Focusing on the CPI inflation forecasts first, I find that the alternative forecasts are significantly more accurate than the RW forecast over all horizons. However, it appears to be more difficult to beat the AR benchmark[17]. That is, the PC model fails to outperform the AR specification in terms of predictive accuracy over all horizons. The PC-gap model also does not provide more accurate CPI inflation forecasts than the AR model over short-term horizons. The rather poor performance of the Phillips curve models when being compared with the AR model confirm the findings of Atkeson, Ohanian et al. (2001), Sims (2002), Fisher, Liu and Zhou (2002), and Clark and McCracken (2006) that the predictive performance of the Phillips curve-based models is not robust and such models fail to outperform simple benchmarks, as opposed to the implications of Stock and Watson (1999).

Yet, I note that the PC-gap forecasts are more precise than the PC forecasts. Analogously, the AR-gap projections are more accurate than the AR predictions. These findings suggest that the stationary specifications for inflation perform less accurate than their non-stationary counterparts in gap form. Specifically, accounting for the time-varying trend in inflation through modeling the inflation gap is even more beneficial over longer horizons.

The UCSV and term structure VAR model, which also impose non-stationarity, both outperform the AR benchmark in terms of predictive accuracy excepting the term structure VAR forecasts over one year. Similar results are found for the machine learning methods, where the LSTM neural network has consistently lower relative RMSPEs than the RF. The accurate performance of the machine learning methods can be ascribed to both the incorporation of non-stationarity and their ability to derive non-linear relations between inflation and the set of predictors.

The best performing CPI inflation forecasts are the judgemental forecasts, i.e., the SPF and ATSIX projections, with substantial RMSPE reductions relative to the RW benchmark up to 37%. The judgemental forecasts have the lowest relative RMSPEs over all horizons, demonstrating their superiority over the other forecasting approaches. Certainly, the professional field experience of the panelists of macroeconomic surveys highly contributes to the dominance of the judgemental inflation forecasting methods.

Moving on to the RMSPEs of the GDP deflator inflation forecasts, I note that the forecasting performance of the alternative methods is generally worse than for the CPI inflation measure. That is, not all methods are able to significantly beat the RW benchmark, not to mention the AR benchmark. The PC model still has rather poor predictive accuracy, being unable to significantly outperform the RW benchmark over any horizon, once more confirming that the predictive ability of the Phillips curve is defective. Comparing the AR-gap and PC-gap models to their stationary counterparts, I find that the former models are more accurate over one- and two-year horizons, again affirming that imposing a time-varying trend is particularly advantageous in forecasting inflation beyond long-term horizons.

The UCSV model performs far worse for GDP deflator inflation and is only able to outperform both benchmarks over the two-years forecast horizon, implying that filtering the current GDP deflator inflation trend not amounts to precise GDP deflator inflation forecasts. One cause of the poor performance of the UCSV model in forecasting GDP deflator inflation could be that

[17]The RMSPEs for the alternative forecasts relative to the AR benchmark together with the outcomes of the augmented DM test on the significance of found accuracy gains with respect to the AR benchmark are reported in Appendix A. Additionally, the absolute RMSPEs for the benchmarks are reported in Appendix B.

the long-run survey expectations on this inflation measure not provide an accurate measurement of the trend in GDP deflator inflation. However, the gap models also incorporate the subjective long-run GDP deflator inflation expectations as trend measures and their performance is similar for the prediction of CPI inflation, making this claim less plausible. Another potential explanation for the worse performance of the UCSV model is that GDP deflator inflation is less volatile than CPI inflation, such that the edge of the stochastic volatility specification is less pronounced.

Moreover, it appears that the dynamic Nelson-Siegel yield curve factors have less predictive power on GDP deflator inflation, since the term structure VAR is, in general, less accurate than both benchmarks. Both the RF and LSTM provide significant accuracy gains with respect to the RW benchmark over the two-year horizon, where the RF also outperforms the AR benchmark, notwithstanding the two-year-ahead LSTM forecasts provide no improvements when compared with the AR forecasts. Over the shorter horizons, the machine learning methods perform less well, failing to outperform the benchmarks. Thus, similar as in the previous panel, the incorporation of non-stationarity and non-linear relations between inflation and predictors is most useful in forecasting over longer intervals.

One should note that the RF is consistently more accurate than the LSTM, as opposed to the results in the previous panel. The shift in upper hand among the machine learning methods corresponds with the differences in the dynamics between the CPI and GDP deflator inflation rates. The best GDP deflator inflation forecasts remain the SPF forecasts with relative RMSPE improvements between 14% and 18%, once more exhibiting their dominance over the other forecasting methods.

Lastly, the performance of the alternative approaches in predicting PCE deflator inflation shows more resemblance with the results that are found for the prediction of CPI inflation. Whereas none of the models are able to outperform the AR benchmark and only a few methods significantly improve the RW benchmark when considering the nowcast (i.e., the forecast over horizon zero), the majority of the models significantly improve the RW yardstick over the remaining horizons. In particular, the models in gap form and the UCSV model provide the most accurate PCE deflator inflation predictions over horizons of one and two years, with RMSPEs around 6% smaller than the AR benchmark.

Thus, imposing non-stationarity is the most important catalyst in accurately forecasting PCE deflator inflation over longer horizons. The coherence of this assertion follows from the fact that the trend in PCE deflator inflation evidently evolves over time, which also holds for the other inflation measures. When the forecast horizon increases, the trend tends to deviate more from the sample mean, such that the impact of accounting for non-stationarity is higher.

Both the RF and LSTM generally fail to improve the predictive accuracy in comparison with the AR model, only marginally outperforming this benchmark in predicting PCE deflator inflation beyond a horizon of two years, anew providing evidence that allowing for non-linearity is most profitable over long-term horizons. Over shorter horizons, the gain of the non-linear structure of the machine learning methods appears to be overshadowed by other factors driving the accuracy of the PCE deflator inflation forecasts, such as the ability to rapidly reckon patterns in the dynamics of the inflation rates. The term structure VAR projections of PCE deflator inflation

consistently fail to surpass the AR benchmark, which implies that the dynamic Nelson-Siegel yield curve factors also have less predictive power on GDP deflator inflation.

Overall, the alternative forecasting methods generally outperform the RW benchmark for all three inflation measures and across all horizons. However, comparing the forecasting methods with the AR benchmark, less accuracy gains are found. In addition, the performance of the models differs for each inflation measure, which is not unexpected considering the differences in dynamics of the three inflation measures discussed in Section 3. Yet there are some remarks on the forecast performance of the alternative methods that hold for all three measures of inflation. Firstly, the non-stationary specifications mostly outperform the stationary models, especially when considering forecasts over longer horizons. That is, models that account for the time-varying trend in inflation rates provide more accurate forecasts than specifications that impose stationarity. This advantage of non-stationary models over stationary specifications increases with the forecast horizon, as the stationary models construct forecasts based on the sample mean, resulting in projections that are consistently off when the forecast horizon expands.

Secondly, the judgemental forecasts are the most accurate methods in this comparison study. This finding extends the findings of past research exhibiting the superiority of subjective inflation forecasts (Ang, Bekaert and Wei, 2007; Faust and Wright, 2013; Borağan Aruoba, 2020), since the SPF and ATSIX consistently outperform the machine learning methods, implying that the subjective forecasts are also superior over the latter stream of methods. It appears that the SPF forecasts have the greatest advantage over the other models when making projections for the current quarter and their superiority over longer horizons is partially caused by their ability to precisely assess the current inflation rate, as the disparities in RMSPE between the judgemental forecasts and the remaining projections decrease over longer horizons. This implication is in line with the suggestions of Sims (2002) and Faust and Wright (2009).

The latter research motivates that the benefits of judgemental forecasts in nowcasting stem from the empirical fact that the relevant data for the nowcast is generally dissimilar to the set of predictors for the prospective quarters. Another reason for the superiority of the judgemental forecasts is that the subjective forecasts integrate the expertise of the financial industry professionals. The know-how of professional macroeconomic forecasters allows them to more rapidly recognize and respond to certain shifts in the dynamics of inflation rates.

Although the performance of the machine learning methods is different for each inflation measure, the lowest RMSPEs for both the RF and LSTM are observed over the two-year horizon, implying that the accuracy of these methods is higher over long-horizons, which is also suggested by Behrens, Pierdzioch and Risse (2018) and Almosova and Andresen (2023). In the long-run, these methods are less affected by temporary deviations and have the most advantage of their non-linear structure.

5.2 Robustness to forecast accuracy measure

Aforementioned, I also examine the MAPE metric in order to confirm whether the found results on the forecast accuracy discussed in the previous segment are robust to the accuracy measure and not caused by a few outliers in forecast errors. Table 3 shows the MAPEs in percentages for the alternative forecasts of the three inflation measures relative to the random walk benchmark.

Table 3: Relative mean absolute prediction errors of the alternative inflation forecasts with respect to the random walk forecast.

Panel A: CPI inflation

	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
AR	0.82**	0.83**	0.80**	0.83*
AR-gap	0.83**	0.83**	0.74**	0.74*
PC	0.87**	0.86**	0.79**	0.84*
PC-gap	0.87**	0.84**	0.74**	0.74*
UCSV	**0.70****	**0.75****	**0.71****	0.72*
Term structure VAR	0.83**	0.83**	0.82*	0.79*
SPF	0.64***	0.76**	0.72**	
ATSIX		0.76**	0.72**	**0.71****
RF	0.77**	0.83**	0.77**	0.79*
LSTM	0.71**	0.77**	0.77**	0.75*

Panel B: GDP deflator inflation

	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
AR	0.91***	0.92**	0.99	0.95
AR-gap	0.91***	0.94*	0.85***	**0.81****
PC	0.95	0.92	0.99	0.97
PC-gap	0.93	0.93	0.86**	**0.81****
UCSV	1.12	1.02	0.94	0.86*
Term structure VAR	1.03	1.02	1.14	1.08
SPF	**0.80***	**0.81***	**0.81****	
RF	1.01	0.93	0.96	0.82**
LSTM	1.13	1.02	1.01	0.91*

Panel C: PCE deflator inflation

	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
AR	**0.88*****	**0.85*****	0.90*	0.89
AR-gap	0.89**	0.86***	**0.83****	**0.77***
PC	0.94	0.89*	0.90*	0.91
PC-gap	0.95	0.90*	**0.83****	**0.77***
UCSV	1.01	0.87**	0.88*	0.80*
Term structure VAR	0.93*	0.89**	0.95	0.91
RF	1.04	0.87*	0.89*	0.85*
LSTM	1.05	0.87*	0.89*	0.84*

This table discloses the MAPE of the different forecasts on three measures of inflation, across forecasting horizons of 0, 1, 4 and 8 quarters. The forecasts are constructed for the period from 1997Q1 until 2022Q4 and are based upon an expanding window of observations with data running back to 1948Q1. The MAPEs are reported as fractions of the MAPEs of the random walk model. The asterisks indicate the significance level of found accuracy gains with respect to the random walk model; one, two and three asterisks correspond to an accuracy improvement at the 10%, 5% and 1% significance level, respectively. The bold values denote the lowest MAPEs across each forecast horizon.

In addition, the asterisks denote whether found accuracy gains relative to the RW model are significant according to the augmented Diebold-Mariano test.

Looking at the CPI inflation panel, similar conclusions as those for the RMSPE results can be drawn. Again, all methods provide significantly more accurate CPI inflation forecasts than the RW model. Indeed, the performance of the Phillips curve models with respect to the AR benchmark[18] is rather poor, excluding the PC-gap forecasts over one and two years. The models in gap form still seem to outperform their stationary equivalents, in particular over the longer horizons.

The performance of the term structure VAR in terms of MAPE is worse than for the former metric, now only improving the AR benchmark in forecasting two years ahead. Another discrepancy with respect to the results on the RMSPE is that according to the MAPE metric, the judgemental CPI inflation forecasts no longer have the highest accuracy over all horizons. That is, the UCSV forecasts, which also are among the best performing methods according to the RMSPE, have the lowest MAPE for both the horizons of one quarter and one year, thus outperforming the judgemental forecasts over these horizons. The found MAPEs for the machine learning methods are in correspondence with the findings from the previous table, with the only noteworthy difference being that the benefit of the SPF nowcast over the RF and LSTM nowcasts is less substantial.

Investigating the GDP deflator inflation panel, the relative MAPEs for the PC forecasts are of comparable magnitude as their relative RMSPEs reported in the previous table, once more affirming that the predictive ability of the PC model is defective. According to the MAPE metric, it also holds that the models in gap form are more accurate than the stationary specifications over the one- and two-year horizon, while one should note even greater disparities in accuracy among these models. In addition, the conclusions on the performance of the UCSV model remain identical, only leading to accuracy improvements relative to both benchmarks over the two-year horizon.

Furthermore, the high MAPEs found for the term structure VAR model back the inference that the model is less accurate in forecasting GDP deflator inflation than the RW model as well as the AR model. The achieved MAPE reductions for the RF and LSTM over the long-horizon in this panel also correspond with the previous results, though the one-year-ahead RF forecasts appear to have more edge over the AR benchmark than implied by the RMSPE. The SPF forecasts lead to the highest MAPE reductions, varying between 19% and 20%, thus of similar size as the found RMSPE improvements for the judgemental forecasts.

Finally, examining the MAPEs in the PCE deflator inflation panel, I do not find results that truly contradict the findings based on the RMSPE. Namely, the AR nowcast is more accurate than all alternative models and over the longer horizons the models in gap form and the UCSV model provide the most accurate predictions. In terms of MAPE, the term structure VAR model is still outperformed by the AR benchmark across all horizons. Moreover, the machine learning methods beat both benchmarks in predicting PCE deflator inflation over two years by a even higher margin than inferred through the RMSPE metric.

[18]The MAPEs for the alternative forecasts relative to the AR benchmark together with the outcomes of the augmented DM test on the significance of found accuracy gains with respect to the AR benchmark are reported in Appendix A. Additionally, the absolute MAPEs for the benchmarks are reported in Appendix B.

In sum, the results on the MAPE metric confirm the robustness of the findings based on the RMSPE metric, excepting marginal differences in the accuracy disparities among the alternative inflation forecasting approaches. The key findings based on the MAPE remain the same as for the prior accuracy measure. That is, the non-stationary methods continue to outperform the stationary specifications over longer horizons and out of all methods, the judgemental forecasts mostly have the highest accuracy. Again, the SPF forecasts have the greatest advantage over the remaining methods in predicting the current inflation rate. Thus, conforming to the MAPE, the findings on the accuracy of the distinct forecasting methods are robust to the accuracy measure.

5.3 Robustness to sample period

As mentioned in Section 3, inflation rates were more volatile during the period after the Great Recession, making the inflation rates more difficult to predict over this interval. In addition, I noted upward spikes in inflation at the end of the forecast sample, which are related to COVID issues in global supply chains and the Russia-Ukraine war. To discover to what extent these spells influence the performance of the alternative forecasting approaches, I examine the predictive accuracy of the different forecasts over the sample before the Great Recession.

Table 4 shows the RMSPEs in percentages for the different forecasts of the three inflation measures relative to the random walk benchmark, over the period before the Great Recession (thus prior to the final quarter of 2007). Once more, the asterisks indicate whether found accuracy gains relative to the RW model are significant according to the augmented Diebold-Mariano test. I note that excluding the period in which the volatility of the inflation rates is higher and the levels of inflation are increasing, does not alter the key implications drawn above, since the obtained RMSPEs are of similar magnitude as the RMSPEs derived for the full sample period (being exhibited in Table 2). Thus, there is no evidence that the results are affected by the sample on which the forecasts are constructed.

Nevertheless, there are some deviations worth mentioning. That is, the PC and the PC-gap forecasts beyond the shorter horizons are significantly more accurate than the RW benchmark over the smaller forecast sample, for all three inflation measures, which is not the case over the full forecast sample. This suggests that the models based on the Phillips curve perform better in periods of low volatility, which is another indication that the Phillips curve is misspecified. Moreover, the advantage of the SPF forecasts over the remaining models in nowcasting CPI inflation somewhat decreases. The logical cause for this finding is that, in times of low volatility, the current inflation rate is more easily gauged. Especially, with respect to the machine learning methods, the nowcasting benefit of the SPF forecasts appears to vanish completely. It seems that the machine learning models are able to accurately track the short-term patterns in CPI inflation, as long as there are no sudden shifts in levels.

Finally, for the prediction of GDP and PCE deflator inflation, the LSTM model no longer significantly outperforms the benchmarks[19] over the one- and two-year horizon. The reason behind this could be that the advantage of the dynamic feature selection feature is decreased, when volatility is lower.

[19]The RMSPEs for the alternative forecasts relative to the AR benchmark over the period before the Great Recession, together with the outcomes of the augmented DM test on the significance of found accuracy gains with respect to the AR benchmark are reported in Appendix C.

Table 4: Relative root mean square prediction errors of the alternative inflation forecasts with respect to the random walk forecast, over the period before the Great Recession.

Panel A: CPI inflation

	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
AR	0.77***	0.76***	0.80**	0.80
AR-gap	0.77***	0.76***	0.76**	0.76*
PC	0.79**	**0.73*****	0.80*	0.82
PC-gap	0.77***	**0.73*****	0.76**	0.77*
UCSV	0.66***	**0.73*****	0.76**	0.73*
Term structure VAR	0.78**	0.79**	0.83*	0.79
SPF	**0.63*****	0.74***	0.76**	
ATSIX		**0.73*****	**0.75****	**0.72***
RF	0.65***	0.78**	0.76**	0.78*
LSTM	0.64***	0.74***	0.79**	0.74*

Panel B: GDP deflator inflation

	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
AR	0.91**	0.88	1.10	0.99
AR-gap	0.92**	0.87**	0.86*	**0.80***
PC	0.90**	0.87**	1.09	1.01
PC-gap	0.92**	0.84***	0.86*	**0.80***
UCSV	1.08	0.98	1.04	0.93
Term structure VAR	0.96	0.96	1.18	1.13
SPF	**0.74*****	**0.74*****	**0.80***	
RF	0.92**	0.84*	0.92	0.83*
LSTM	1.12	0.94	1.10	0.99

Panel C: PCE deflator inflation

	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
AR	0.87**	0.83**	1.00	1.08
AR-gap	0.90*	0.84***	0.83**	**0.84**
PC	0.86**	0.82**	1.01	1.11
PC-gap	0.89***	0.82***	0.81**	0.86
UCSV	0.92	0.83	0.93	1.00
Term structure VAR	0.93	0.90	1.03	1.16
RF	**0.84****	**0.76****	**0.81****	0.97
LSTM	1.00	0.85	0.95	1.12

This table discloses the RMSPE of the different forecasts on three measures of inflation, across forecasting horizons of 0, 1, 4 and 8 quarters. The forecasts are constructed for the period from 1997Q1 until 2007Q3 and are based upon an expanding window of observations with data running back to 1948Q1. The RMSPEs are reported as fractions of the RMSPEs of the random walk model. The asterisks indicate the significance level of found accuracy gains with respect to the random walk model; one, two and three asterisks correspond to an accuracy improvement at the 10%, 5% and 1% significance level, respectively. The bold values denote the lowest RMSPEs across each forecast horizon.

5.4 Bias

As exhibited before, the non-stationary specifications are more accurate than stationary models in forecasting inflation over longer horizons. To illustrate this, I consider Figure 5, which displays the AR, PC-gap, LSTM and UCSV forecasts of CPI inflation over a horizon of two years, as well as the actuals. I observe that the AR projections fluctuate around the sample average of the CPI inflation rates, consistently causing them to be either too high or too low, as the trend of the realized inflation rates tends to vary over time. Thus, the bias caused by the assumption of stationarity affects the accuracy of the AR model.

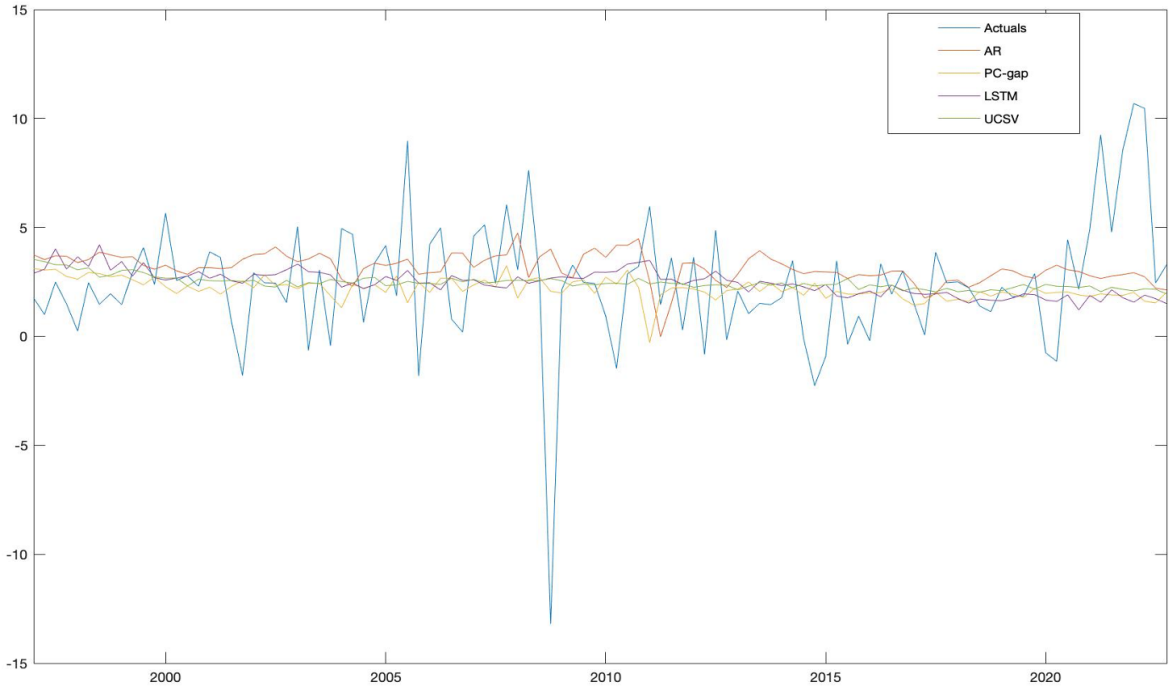


Figure 5: Selection of two-year-ahead CPI inflation forecasts together with actuals, on out-of sample forecasting period 1997Q1 - 2022Q4.

The PC-gap and UCSV forecasts not suffer from this bias, due to their non-stationary specifications. Specifically, these forecasts reckon the slowly-evolving trend in inflation by modelling the inflation gap and current inflation trend. The LSTM forecasts also capture the time-varying mean better than the AR predictions, as a result of the non-linear structure of the LSTM.
To further illuminate the implications from above, I examine Table 5, which shows the bias for the alternative forecasts of the three inflation measures. One should note that the bias of the stationary specifications increases with the forecast horizon, being most poignant over the long horizon. This finding gives reason why the non-stationary models mostly outperform the stationary specifications in terms of predictive accuracy over longer horizons.

Table 5: Bias of the alternative inflation forecasts.

Panel A: CPI inflation				
	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
RW	**0.01**	**0.01**	0.20	0.31
AR	-0.18	-0.30	-0.56	-0.70
AR-gap	0.14	0.22	0.26	0.24
PC	-0.22	-0.25	-0.56	-0.69
PC-gap	0.18	0.24	0.31	0.25
UCSV	0.02	0.02	**0.00**	**-0.03**
Term structure VAR	-0.27	-0.42	-0.64	-0.76
SPF	0.16	0.22	0.13	
ATSIX		0.26	0.20	0.14
RF	0.46	0.37	0.42	0.28
LSTM	0.04	0.04	-0.02	-0.04
Panel B: GDP deflator inflation				
	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
RW	0.08	0.10	0.25	0.36
AR	-0.10	-0.20	-0.47	-0.51
AR-gap	0.13	0.17	0.26	0.20
PC	-0.09	-0.17	-0.47	-0.50
PC-gap	0.15	0.17	0.26	0.25
UCSV	-0.23	-0.23	-0.26	-0.30
Term structure VAR	-0.08	-0.23	-0.52	-0.77
SPF	0.12	0.06	**0.02**	
RF	0.08	0.06	0.15	**0.09**
LSTM	**0.01**	**-0.01**	-0.04	-0.13
Panel C: PCE deflator inflation				
	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
RW	**0.02**	**0.05**	0.18	0.26
AR	-0.15	-0.28	-0.58	-0.63
AR-gap	0.12	0.18	0.23	**0.20**
PC	-0.10	-0.24	-0.57	-0.61
PC-gap	0.23	0.20	0.33	**0.20**
UCSV	-0.33	-0.33	-0.36	-0.41
Term structure VAR	-0.15	-0.30	-0.60	-0.79
RF	0.36	0.31	0.33	0.31
LSTM	-0.11	-0.11	**-0.13**	-0.22

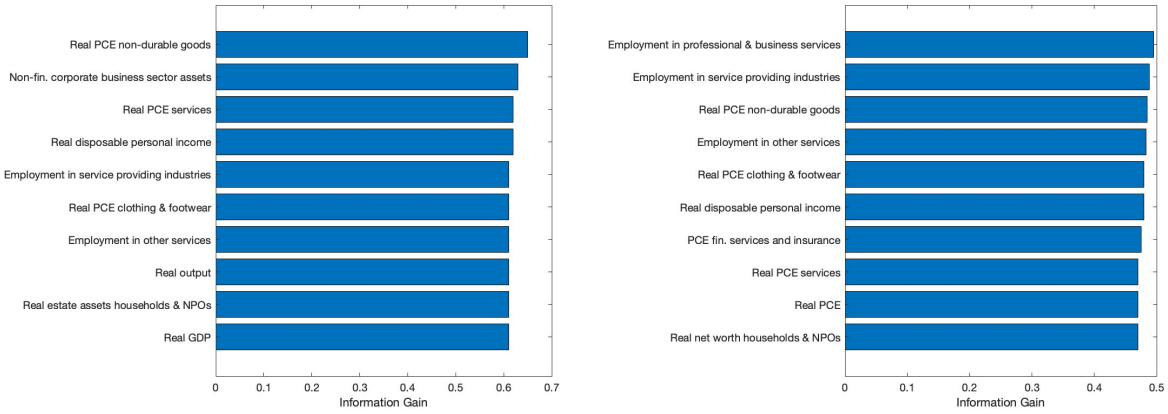
This table discloses the bias of the different forecasts on three measures of inflation, across forecasting horizons of 0, 1, 4 and 8 quarters. The forecasts are constructed for the period from 1997Q1 until 2022Q4 and are based upon an expanding window of observations with data running back to 1948Q1. The bold values denote the lowest (absolute) biases across each forecast horizon.

Moreover, the term structure VAR forecasts suffer from a even higher bias, excepting the nowcasts, even though the specification aims to model the inflation gap. This result suggests that relating the inflation gap to the term structure of government bond yields is rather deficient, since the other models in gap form are considerably less biased. Namely, the AR-gap and PC-gap models perform more stable in terms of bias, generally having a lower bias than the AR and PC model, as already partially suspected above. The RW model logically has relatively low bias over short horizons and high bias over long horizons.
However, it is peculiar that though the judgemental forecasts have relatively low biases, the UCSV model and the LSTM score even lower biases for CPI inflation (and the latter model for GDP deflator inflation as well), yet the judgemental forecasts are substantially more accurate than the projections of both models. Once more, I find evidence that there are other factors leading to the disparities between the judgemental forecasts and the other inflation forecasting methods, such as the advantage of the SPF forecasts in nowcasting.
Similar as for the accuracy of the distinct forecasting approaches, the performance in terms of bias differs for the three inflation measures, due to the differences in dynamics and decompositions of the inflation measures. Nevertheless, the LSTM forecasts are amongst the methods with the smallest biases, across all inflation measures. The low bias of the LSTM model can be attributed to the complexity of the recurrent neural network specification. Furthermore, the relatively small biases for the LSTM forecasts indicate that the LSTM model potentially suffers from overfitting, which adversely affects the out-of-sample forecasting accuracy. At last, the bias of the RF model is stable over all horizons due to its inherent aggregation and ensemble method, but higher for the prediction of CPI and PCE deflator inflation than for the prediction of GDP deflator inflation, corresponding to the results on the predictive accuracy of the model.

5.5 Feature importance

As explained in Section 4.3.2, I apply a dynamic feature selection approach based on the Information Gain in determining which variables of the sizeable FRED-QD dataset to feed to the LSTM neural network in order to prevent overfitting. Figure 6 illustrates the input variables[20] for the LSTM model with the highest Information Gain scores for the first and final CPI inflation forecast over a horizon of two years, on which the LSTM is among the most accurate methods in my forecasting exercise. As can be seen, the leading features in terms of Information Gain are moderately dissimilar for the two forecasts and thus it appears that the importance of the explanatory variables fairly adjusts over time.

[20]For an overview of all variables and their description, I refer to the Appendix of McCracken and Ng (2016).



(a) Forecast for 1997Q1

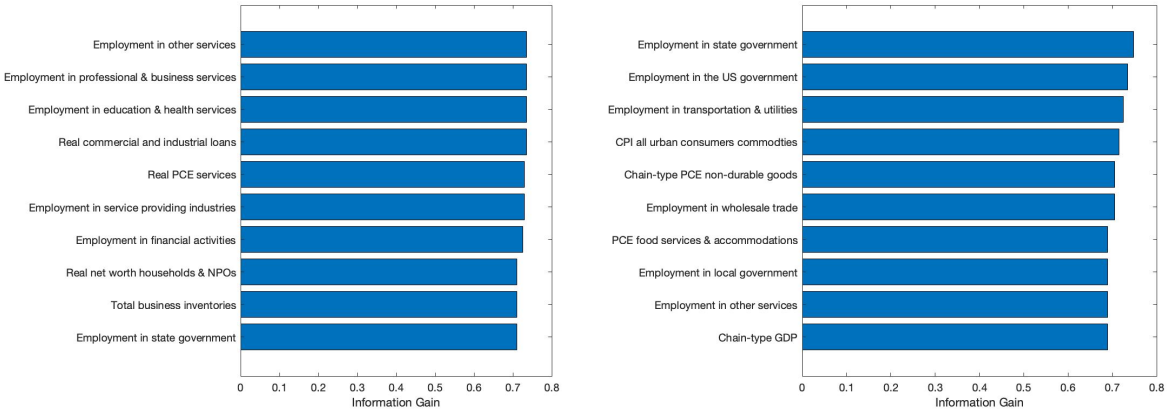
(b) Forecast for 2022Q4

Figure 6: Highest feature importance scores measured by the Information Gain of the LSTM input variables for the first and final two-years-ahead CPI inflation forecast.

That is, for the first two-years-ahead CPI inflation projection the set of variables with the highest importance scores consist of real PCE regarding non-durable goods, clothing and footwear, and services. Additionally, the set consists of real non-financial corporate business sector assets, real disposable personal income, both employment in service-providing industries and in other services, real output, real estate assets of households and non-profit organizations (NPOs), and real GDP. However, for the final long-term CPI forecast, there are four different variables among the set of features with the highest importance scores. Namely, the latter selection includes employment in professional and business services, PCE regarding financial services and insurance, real PCE, and real net worth of households and non-profit organizations.

As a matter of fact, the selection of input variables for the LSTM also differentiates over the forecast sample, once more implying that the explanatory power of the macroeconomic variables evolves over time. However, the input selection of both the first and final two-years-ahead CPI inflation forecast mostly consists of variables describing output, income, interest rates, exchange rates and logically other price indexes. Measures on the stock market, orders and inventories are seemingly less informative for CPI inflation. Accordingly, the dynamic feature selection results in variable sets describing similar aspects of the economy, but the choice of the specific variables representing these characteristics varies over time.

Figure 7 shows the input variables for the LSTM model with the highest Information Gain scores for the first and final PCE deflator inflation forecast over a horizon of two years. As exhibited in the figure, the input variables for the long PCE deflator inflation forecasts with the highest Information Gain scores are not corresponding to the chief features for the long CPI inflation forecasts. That is, the batch of the most important input features for the final two-years-ahead PCE deflator inflation forecast comprises nine alternative variables with respect to the previous figures.



(a) Forecast for 1997Q1

(b) Forecast for 2022Q4

Figure 7: Highest feature importance scores measured by the Information Gain of the LSTM input variables for the first and final two-years-ahead PCE deflator inflation forecast.

Specifically, among the key features are employment in local government, in state government, in the US government, as well as employment in trade, in transportation and utilities, and in wholesale trade. In addition, the chief variables cover the CPI for all urban consumers commodities, chain-type PCE regarding non-durable goods, PCE regarding food services and accommodations, and chain-type GDP.

Similar as noted above, the set of key explanatory features for PCE deflator inflation moderately evolves over time. In particular, the key features for the first PCE deflator inflation forecast over a horizon of two years involve employment in education and health services, real commercial and industrial loans, employment in financial activities and total business inventories, which are not among the chief variables for the final PCE deflator inflation forecast. Furthermore, the input sets for the LSTM in forecasting PCE deflator inflation are rather different. This signifies that the feature importance scores not only vary over time, but also for the distinct inflation measures, which in part unravels why the performance of the LSTM neural network is deviating for the three inflation series.

The feature importance scores as measured by the Information Gain not only give insight on the performance of the LSTM model, but also on the performance of the other models. Namely, the figures above provide evidence that the explanatory power of variables is moderately shifting over time, such that the models incorporating a fixed set of features (e.g., the PC and term structure VAR model) are disadvantaged compared to the methods based on evolving variable sets (i.e., the SPF, RF and LSTM forecasts). Nonetheless, for the LSTM and RF model, the advantage of dynamic variable selection is prevailed by other factors influencing the forecast accuracy, as both models are not consistently more accurate than the models based on a restricted set of features. Besides, the dynamic feature selection based on the Information Gain add more interpretability to the rather black-box LSTM neural network, as the choice of input variables is motivated by explanatory power.

6 Conclusion

In this paper, I aim to compare the accuracy of judgemental forecasts, classical statistical models and machine learning models in forecasting inflation out-of-sample. Firstly, I find that the judgemental forecasts are the most accurate methods in my comparison study. Specifically, the SPF and ATSIX forecasts significantly outperform both the RW and AR benchmark across all horizons and are generally more accurate than the classical statistical models and the machine learning models. The highest disparities in predictive accuracy between the judgemental SPF forecasts and the other models are observed when constructing nowcasts and the superiority of the SPF over longer horizons is (at least partially) caused by this advantage in nowcasting inflation. Accordingly, the ATSIX provides the most accurate CPI inflation forecasts over a horizon of two years.

These findings extend the implications of Ang, Bekaert and Wei (2007) and Faust and Wright (2013), since they demonstrate that the judgemental forecasts are not only superior over classical statistical models but also over machine learning models. The superiority of judgemental forecasts can be ascribed to the expertise of industry professionals and their ability to rapidly recognize shifts in the dynamics of inflation rates.

Secondly, the predictive accuracy of the machine learning methods is the highest over the long-term horizon of two years. Especially for the prediction of GDP and PCE deflator inflation, the RF and the LSTM model perform relatively poor over shorter horizons. Comparing both machine learning methods, I not find that one of the models is strictly superior over the other. Namely, the LSTM is more accurate in forecasting CPI inflation and the RF provides more precise GDP deflator inflation forecasts, whereas the performance of the models in predicting PCE deflator inflation is identical.

Thirdly, considering the classical statistical inflation forecasting models, the non-stationary specifications are generally more accurate than the stationary models. In particular, the accuracy gains of the non-stationary models over the stationary models increase over longer horizons. There is no preferable method to reckon a slowly-evolving trend in inflation series. That is, no specification among the AR-gap, PC-gap and UCSV model is strictly dominating the remaining non-stationary models. However, relating the inflation gap to the term structure of government bond yields is rather defective, considering that the term structure VAR model mostly fails to outperform the stationary AR benchmark.

Although the findings above hold for all three inflation measures considered in this research (i.e., CPI, GDP deflator and PCE deflator inflation), one should note that the performance of the alternative inflation forecasting methods differs for each measure. That is, the methods are generally more accurate for the prediction of CPI inflation than for the prediction of GDP and PCE deflator inflation. Nevertheless, the found results on the forecast accuracy of the alternative methods are not affected by outliers in the forecast errors, as the outcomes are ratified by both the RMSPE and the MAPE metric. Similarly, the results on the accuracy of the different inflation forecasts are robust to the sample period, as excluding periods where volatility is higher and the levels of inflation rates are increased, yields the same implications.

The findings of this study have important implications for central banks, since they can implement the successful methods from this research to predict inflation rates more accurately in

order to improve the potency of monetary policy. Furthermore, more accurate projections of inflation rates aid traders of indexed-linked bonds and inflation derivatives in optimizing their portfolios. This research is also useful for the private sector, which has an inherent stake in making accurate inflation forecasts, since long-term nominal obligations regarding labor, sales, leases, mortgages, and other debts are prevailing in modern economies.

This research has potential limitations. Aforementioned, the survey forecasts have the greatest advantage over the other inflation forecasting methods in nowcasting, which also causes the judgemental forecasts to be superior in constructing forecasts over longer horizons. Future research should investigate how the other methods perform relative to the judgemental forecasts when one accounts for their disadvantage in nowcasting inflation, e.g., by following the trivial approach of Faust and Wright (2013) using the survey nowcasts as jumping-off point for all models.

For follow-up research, it is also relevant to investigate how the alternative inflation forecasting methods compete over an even longer horizon, for example a 10-year horizon, as the disparities in forecast accuracy among the distinct methods in this study appear to be diminishing with the horizon and thus the ranking of the methods in terms of forecast accuracy potentially shifts over longer horizons. Moreover, it is particularly useful to discover which inflation forecasts are the most accurate over such long-term horizons, as these forecasts can then be used as the trend inflation measures in inflation gap models, potentially improving their forecasting performance across all horizons.

Another limitation concerns the LSTM model. That is, I select the hyperparameters of the LSTM neural network by trial-and-error in order to decrease computational complexity and the complementary running time, whereas tuning the hyperparameters through an extensive grid search, which comes at the cost of drastically increasing the computational time, could improve the forecasting performance of the LSTM model. I also recommend future researchers to inspect whether the implications of my paper hold for the price indices in other countries and are not limited to the inflation rates in the US. Finally, as I highlight certain benefits and drawbacks of the distinct inflation forecasting approaches, it is relevant to research how the different forecasts can be optimally combined.

References

[1] Anna Almosova and Niek Andresen. “Nonlinear inflation forecasting with recurrent neural networks”. In: *Journal of Forecasting* 42.2 (2023), pp. 240–259.

[2] Gianni Amisano and Raffaella Giacomini. “Comparing density forecasts via weighted likelihood ratio tests”. In: *Journal of Business & Economic Statistics* 25.2 (2007), pp. 177–190.

[3] Andrew Ang, Geert Bekaert and Min Wei. “Do macro variables, asset markets, or surveys forecast inflation better?” In: *Journal of Monetary Economics* 54.4 (2007), pp. 1163–1212.

[4] George Athanasopoulos and Farshid Vahid. “VARMA versus VAR for macroeconomic forecasting”. In: *Journal of Business & Economic Statistics* 26.2 (2008), pp. 237–252.

[5] Andrew Atkeson, Lee E Ohanian et al. “Are Phillips curves useful for forecasting inflation?” In: *Federal Reserve bank of Minneapolis quarterly review* 25.1 (2001), pp. 2–11.

[6] B Azhagusundari, Antony Selvadoss Thanamani et al. “Feature selection based on information gain”. In: *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* 2.2 (2013), pp. 18–21.

[7] Mitchell J Barnett, Shadi Doroudgar, Vista Khosraviani and Eric J Ip. “Multiple comparisons: To compare or not to compare, that is the question”. In: *Research in Social and Administrative Pharmacy* 18.2 (2022), pp. 2331–2334.

[8] Christoph Behrens, Christian Pierdzioch and Marian Risse. “Testing the optimality of inflation forecasts under flexible loss with random forests”. In: *Economic Modelling* 72 (2018), pp. 270–277.

[9] Christoph Bergmeir and José M Benítez. “On the use of cross-validation for time series predictor evaluation”. In: *Information Sciences* 191 (2012), pp. 192–213.

[10] S Borağan Aruoba. “Term structures of inflation expectations and real interest rates”. In: *Journal of Business & Economic Statistics* 38.3 (2020), pp. 542–553.

[11] Charles S Bos, Philip Hans Franses and Marius Ooms. “Inflation, forecast intervals and long memory regression models”. In: *International Journal of Forecasting* 18.2 (2002), pp. 243–264.

[12] Michael J Boskin. “Causes and Consequences of Bias in the Consumer Price Index as a Measure of the Cost of Living”. In: *Atlantic Economic Journal* 33 (2005), pp. 1–13.

[13] Flint Brayton, John M Roberts and John C Williams. “What’s happened to the Phillips curve?” In: *Available at SSRN 190852* (1999).

[14] Leo Breiman. “Random forests”. In: *Machine learning* 45 (2001), pp. 5–32.

[15] Fabio Canova. “G-7 inflation forecasts: Random walk, Phillips curve or what else?” In: *Macroeconomic Dynamics* 11.1 (2007), pp. 1–30.

[16] Stephen G Cecchetti, Michael Feroli, Peter Hooper, Anil K Kashyap and Kermit L Schoenholtz. “Deflating inflation expectations: The implications of inflation’s simple dynamics”. In: *CEPR Discussion Paper No. DP11925* (2017).

[17] Joshua CC Chan. “Moving average stochastic volatility models with application to inflation forecast”. In: *Journal of Econometrics* 176.2 (2013), pp. 162–172.

[18] Xiaohong Chen, Jeffrey Racine and Norman R Swanson. “Semiparametric ARX neural-network models with an application to forecasting inflation”. In: *IEEE Transactions on neural networks* 12.4 (2001), pp. 674–683.

[19] M Ali Choudhary and Adnan Haider. “Neural network models for inflation forecasting: an appraisal”. In: *Applied Economics* 44.20 (2012), pp. 2631–2635.

[20] Todd Clark and Michael McCracken. “Advances in forecast evaluation”. In: *Handbook of economic forecasting* 2 (2013), pp. 1107–1201.

[21] Todd E Clark. "Real-time density forecasts from Bayesian vector autoregressions with stochastic volatility". In: *Journal of Business & Economic Statistics* 29.3 (2011), pp. 327–341.

[22] Todd E Clark and Michael W McCracken. "Nested forecast model comparisons: a new approach to testing equal accuracy". In: *Journal of Econometrics* 186.1 (2015), pp. 160–177.

[23] Todd E Clark and Michael W McCracken. "The predictive content of the output gap for inflation: Resolving in-sample and out-of-sample evidence". In: *Journal of Money, Credit and Banking* 186.1 (2006), pp. 1127–1148.

[24] Timothy Cogley, Giorgio E Primiceri and Thomas J Sargent. "Inflation-gap persistence in the US". In: *American Economic Journal: Macroeconomics* 2.1 (2010), pp. 43–69.

[25] Timothy Cogley and Thomas J Sargent. "Evolving post-world war II US inflation dynamics". In: *NBER macroeconomics annual* 16 (2001), pp. 331–373.

[26] Francis X Diebold and Canlin Li. "Forecasting the term structure of government bond yields". In: *Journal of Econometrics* 130.2 (2006), pp. 337–364.

[27] Francis X Diebold and Roberto S Mariano. "Comparing predictive accuracy". In: *Journal of Business and Economic Statistics* 13.3 (1995), pp. 253–263.

[28] James Durbin and Siem Jan Koopman. *Time series analysis by state space methods*. Vol. 38. OUP Oxford, 2012.

[29] Jon Faust and Jonathan H Wright. "Comparing Greenbook and reduced form forecasts using a large realtime dataset". In: *Journal of Business & Economic Statistics* 27.4 (2009), pp. 468–479.

[30] Jon Faust and Jonathan H Wright. "Forecasting inflation". In: *Handbook of economic forecasting*. Vol. 2. Elsevier, 2013, pp. 2–56.

[31] Jonas DM Fisher, Chin T Liu and Ruilin Zhou. "When can we forecast inflation?" In: *Economic Perspectives-Federal Reserve Bank of Chicago* 26.1 (2002), pp. 32–44.

[32] Jeffrey C Fuhrer. "The Phillips curve is alive and well". In: *New England Economic Review* (1995), pp. 41–57.

[33] Márcio GP Garcia, Marcelo C Medeiros and Gabriel FR Vasconcelos. "Real-time inflation forecasting with high-dimensional models: The case of Brazil". In: *International Journal of Forecasting* 33.3 (2017), pp. 679–693.

[34] Tilmann Gneiting and Roopesh Ranjan. "Comparing density forecasts using threshold- and quantile-weighted scoring rules". In: *Journal of Business & Economic Statistics* 29.3 (2011), pp. 411–422.

[35] Philippe Goulet Coulombe, Maxime Leroux, Dalibor Stevanovic and Stéphane Surprenant. "How is machine learning useful for macroeconomic forecasting?" In: *Journal of Applied Econometrics* 37.5 (2022), pp. 920–964.

[36] Alan P Grant and Lloyd B Thomas. "Inflationary expectations and rationality revisited". In: *Economics Letters* 62.3 (1999), pp. 331–338.

[37] Jan JJ Groen, Richard Paap and Francesco Ravazzolo. "Real-time inflation forecasting in a changing world". In: *Journal of Business & Economic Statistics* 31.1 (2013), pp. 29–44.

[38] Bruce W Hamilton. "Using Engel's Law to estimate CPI bias". In: *American Economic Review* 91.3 (2001), pp. 619–630.

[39] David Harvey, Stephen Leybourne and Paul Newbold. "Testing the equality of prediction mean squared errors". In: *International Journal of forecasting* 13.2 (1997), pp. 281–291.

[40] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[41] Semin Ibisevic. "Diebold-Mariano Test Statistic". In: *MathWorks* (2024). URL: `https://nl.mathworks.com/matlabcentral/fileexchange/33979-diebold-mariano-test-statistic`.

[42] Atsushi Inoue and Lutz Kilian. "How useful is bagging in forecasting economic time series? A case study of US consumer price inflation". In: *Journal of the American Statistical Association* 103.482 (2008), pp. 511–522.

[43] Swati Jadhav, Hongmei He and Karl Jenkins. "Information gain directed genetic algorithm wrapper feature selection for credit rating". In: *Applied Soft Computing* 69 (2018), pp. 541–553.

[44] Scott Joslin, Marcel Priebsch and Kenneth J Singleton. "Risk premiums in dynamic term structure models with unspanned macro risks". In: *The Journal of Finance* 69.3 (2014), pp. 1197–1233.

[45] John T Kent. "Information gain and a general measure of correlation". In: *Biometrika* 70.1 (1983), pp. 163–173.

[46] Chang-Jin Kim, Pym Manopimoke and Charles R Nelson. "Trend inflation and the nature of structural breaks in the New Keynesian Phillips curve". In: *Journal of Money, Credit and Banking* 46.2-3 (2014), pp. 253–266.

[47] Martin Längkvist, Lars Karlsson and Amy Loutfi. "A review of unsupervised feature learning and deep learning for time-series modeling". In: *Pattern recognition letters* 42 (2014), pp. 11–24.

[48] Shang Lei. "A feature selection method based on information gain and genetic algorithm". In: *2012 international conference on computer science and electronics engineering*. Vol. 2. IEEE. 2012, pp. 355–358.

[49] Yan Liu and Jing Cynthia Wu. "Reconstructing the yield curve". In: *Journal of Financial Economics* 142.3 (2021), pp. 1395–1425.

[50] Massimiliano Marcellino, James H Stock and Mark W Watson. "A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series". In: *Journal of econometrics* 135.1-2 (2006), pp. 499–526.

[51] Peter McAdam and Paul McNelis. "Forecasting inflation with thick models and neural networks". In: *Economic Modelling* 22.5 (2005), pp. 848–867.

[52] Michael W McCracken and Serena Ng. "FRED-MD: A monthly database for macroeconomic research". In: *Journal of Business & Economic Statistics* 34.4 (2016), pp. 574–589.

[53] Marcelo C Medeiros and Eduardo F Mendes. "1-regularization of high-dimensional time-series models with non-Gaussian and heteroskedastic errors". In: *Journal of Econometrics* 191.1 (2016), pp. 255–271.

[54] Marcelo C Medeiros, Gabriel FR Vasconcelos, Álvaro Veiga and Eduardo Zilberman. "Forecasting inflation in a data-rich environment: the benefits of machine learning methods". In: *Journal of Business & Economic Statistics* 39.1 (2021), pp. 98–119.

[55] Yash P Mehra. "Survey measures of expected inflation: revisiting the issues of predictive content and rationality". In: *FRB Richmond Economic Quarterly* 88.3 (2002), pp. 17–36.

[56] Elmar Mertens and James M Nason. "Inflation and professional forecast dynamics: An evaluation of stickiness, persistence, and volatility". In: *Quantitative Economics* 11.4 (2020), pp. 1485–1520.

[57] James Morley, Jeremy Piger and Robert Rasche. "Inflation in the G7: Mind the Gap (s)?" In: *Macroeconomic Dynamics* 19.4 (2015), pp. 883–912.

[58] Emi Nakamura. "Inflation forecasting using a neural network". In: *Economics Letters* 86.3 (2005), pp. 373–378.

[59] Nicholas N Noe and George M von Furstenberg. "The upward bias in the consumer price index due to substitution". In: *Journal of Political Economy* 80.6 (1972), pp. 1280–1286.

[60] Rodrigo Peirano, Werner Kristjanpoller and Marcel C Minutolo. "Forecasting inflation in Latin American countries using a SARIMA–LSTM combination". In: *Soft Computing* 25.16 (2021), pp. 10851–10862.

[61] Thomas V Perneger. "What's wrong with Bonferroni adjustments". In: *Bmj* 316.7139 (1998), pp. 1236–1238.

[62] Giorgio E Primiceri. "Time varying structural vector autoregressions and monetary policy". In: *The Review of Economic Studies* 72.3 (2005), pp. 821–852.

[63] Adolfo Rodríguez-Vargas. "Forecasting Costa Rican inflation with machine learning methods". In: *Latin American Journal of Central Banking* 1.1-4 (2020), p. 100012.

[64] Kenneth J Rothman. "No adjustments are needed for multiple comparisons". In: *Epidemiology* 1.1 (1990), pp. 43–46.

[65] Carolyn M Rutter. "Looking back at prospective studies". In: *Academic radiology* 15.11 (2008), pp. 1463–1466.

[66] Christopher A Sims. "A nine-variable probabilistic macroeconomic forecasting model". In: *Business cycles, indicators, and forecasting*. University of Chicago press, 1993, pp. 179–212.

[67] Christopher A Sims. "The role of models and probabilities in the monetary policy process". In: *Brookings Papers on Economic Activity* 2002.2 (2002), pp. 1–40.

[68] Christopher A Sims and Tao Zha. "Were there regime switches in US monetary policy?" In: *American Economic Review* 96.1 (2006), pp. 54–81.

[69] Tom AB Snijders. "On cross-validation for predictor evaluation in time series". In: *On Model Uncertainty and its Statistical Implications: Proceedings of a Workshop, Held in Groningen, The Netherlands, September 25–26, 1986*. Springer. 1988, pp. 56–69.

[70] James H Stock and Mark W Watson. "Evidence on structural instability in macroeconomic time series relations". In: *Journal of Business & Economic Statistics* 14.1 (1996), pp. 11–30.

[71] James H Stock and Mark W Watson. "Forecasting inflation". In: *Journal of monetary economics* 44.2 (1999), pp. 293–335.

[72] James H Stock and Mark W Watson. "Has the business cycle changed and why?" In: *NBER macroeconomics annual* 17 (2002), pp. 159–218.

[73] James H Stock and Mark W Watson. *Modeling inflation after the crisis*. Tech. rep. National Bureau of Economic Research, 2010.

[74] James H Stock and Mark W Watson. "Phillips curve inflation forecasts". In: *Understanding Inflation and the Implications for Monetary Policy: A Phillips Curve Retrospective* (2008).

[75] James H Stock and Mark W Watson. "Why has US inflation become harder to forecast?" In: *Journal of Money, Credit and banking* 39 (2007), pp. 3–33.

[76] Lloyd B Thomas Jr. "Survey measures of expected US inflation". In: *Journal of Economic perspectives* 13.4 (1999), pp. 125–144.

[77] Peter Tulip. "Has the economy become more predictable? Changes in Greenbook forecast accuracy". In: *Journal of Money, Credit and Banking* 41.6 (2009), pp. 1217–1231.

[78] Dick Van Dijk, Siem Jan Koopman, Michel Van der Wel and Jonathan H Wright. "Forecasting interest rates with shifting endpoints". In: *Journal of Applied Econometrics* 29.5 (2014), pp. 693–712.

A Results on forecast accuracy relative to the AR benchmark

Table 6: Relative root mean square prediction errors of the alternative inflation forecasts with respect to the AR forecast.

Panel A: CPI inflation

	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
RW	1.24	1.27	1.25	1.27
AR-gap	1.00	0.99*	0.98*	0.95**
PC	1.06	1.03	1.00	1.01
PC-gap	1.06	1.01	0.98*	0.96**
UCSV	0.94*	0.94*	0.98*	0.94*
Term structure VAR	0.99*	0.97*	1.04	0.97**
SPF	**0.78****	0.93**	0.97*	
ATSIX		**0.92***	**0.96***	**0.92****
RF	0.99*	0.99*	1.03	0.99**
LSTM	0.96*	0.96*	1.01	0.95**

Panel B: GDP deflator inflation

	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
RW	1.15	1.11	1.05	1.11
AR-gap	1.01	1.01	0.96***	0.95*
PC	1.11	1.09	1.01	1.02
PC-gap	1.11	1.09	0.97*	0.95*
UCSV	1.30	1.22	1.00	**0.92****
Term structure VAR	1.17	1.17	1.19	1.07
SPF	**0.94***	**0.95***	**0.86*****	
RF	1.13	1.11	1.00	0.94*
LSTM	1.35	1.25	1.07	1.00

Panel C: PCE deflator inflation

	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
RW	1.13	1.18	1.14	1.17
AR-gap	**1.02**	1.01	**0.95***	**0.92***
PC	1.08	1.06	1.01	1.02
PC-gap	1.09	1.07	**0.95***	**0.92****
UCSV	1.09	**0.97**	0.96*	0.93*
Term structure VAR	1.05	1.01	1.05	1.01
RF	1.17	1.03	1.00	0.99
LSTM	1.15	1.03	1.01	0.99

This table discloses the RMSPE of the different forecasts on three measures of inflation, across forecasting horizons of 0, 1, 4 and 8 quarters. The forecasts are constructed for the period from 1997Q1 until 2022Q4 and are based upon an expanding window of observations with data running back to 1948Q1. The RMSPEs are reported as fractions of the RMSPEs of the AR model. The asterisks indicate the significance level of found accuracy gains with respect to the AR model; one, two and three asterisks correspond to an accuracy improvement at the 10%, 5% and 1% significance level, respectively. The bold values denote the lowest RMSPEs across each forecast horizon.

Table 7: Relative mean absolute prediction errors of the alternative inflation forecasts with respect to the AR forecast.

Panel A: CPI inflation

	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
RW	1.21	1.20	1.26	1.21
AR-gap	1.00	0.99*	0.93*	0.90**
PC	1.05	1.03	1.00	1.02
PC-gap	1.06	1.01	0.93*	0.90**
UCSV	0.85*	**0.91***	0.91*	0.88*
Term structure VAR	1.01	0.97*	1.04	0.96**
SPF	**0.78****	0.92**	**0.90***	
ATSIX		0.92*	0.91*	**0.86****
RF	0.93*	1.00	0.97	0.96**
LSTM	0.86*	0.92*	0.96	0.89**

Panel B: GDP deflator inflation

	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
RW	1.10	1.08	1.01	1.05
AR-gap	1.00	1.02	0.86***	0.95*
PC	1.05	0.99	1.00	1.02
PC-gap	1.02	1.00	0.87*	**0.85***
UCSV	1.22	1.10	0.96	0.89**
Term structure VAR	1.14	1.10	1.15	1.09
SPF	**0.88***	**0.87***	**0.82*****	
RF	1.11	1.01	0.97	0.87**
LSTM	1.24	1.11	1.02	0.95**

Panel C: PCE deflator inflation

	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
RW	1.12	1.17	1.11	1.12
AR-gap	**1.01**	1.01	**0.92***	**0.86***
PC	1.07	1.05	1.00	1.02
PC-gap	1.08	1.05	**0.92***	**0.86****
UCSV	1.15	**0.99**	0.97*	0.91*
Term structure VAR	1.10	1.04	1.06	1.02
RF	1.19	1.02	0.99	0.95
LSTM	1.20	1.02	0.99	0.95

This table discloses the MAPE of the different forecasts on three measures of inflation, across forecasting horizons of 0, 1, 4 and 8 quarters. The forecasts are constructed for the period from 1997Q1 until 2022Q4 and are based upon an expanding window of observations with data running back to 1948Q1. The MAPEs are reported as fractions of the MAPEs of the AR model. The asterisks indicate the significance level of found accuracy gains with respect to the AR model; one, two and three asterisks correspond to an accuracy improvement at the 10%, 5% and 1% significance level, respectively. The bold values denote the lowest MAPEs across each forecast horizon.

B Results on absolute accuracy of the benchmark forecasts

Table 8: Root mean square prediction errors of the benchmark inflation forecasts.

Panel A: CPI inflation				
	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
RW	3.80	3.88	3.72	3.97
AR	3.07	3.07	2.99	3.13
Panel B: GDP deflator inflation				
	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
RW	1.35	1.42	1.70	1.93
AR	1.18	1.28	1.62	1.74
Panel C: PCE deflator inflation				
	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
RW	1.85	2.14	2.15	2.33
AR	1.63	1.82	1.89	2.00

This table discloses the absolute RMSPE of the benchmark forecasts on three measures of inflation, across forecasting horizons of 0, 1, 4 and 8 quarters. The forecasts are constructed for the period from 1997Q1 until 2022Q4 and are based upon an expanding window of observations with data running back to 1948Q1.

Table 9: Mean absolute prediction errors of the benchmark inflation forecasts.

Panel A: CPI inflation				
	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
RW	2.69	2.52	2.65	2.69
AR	2.22	2.10	2.11	2.23
Panel B: GDP deflator inflation				
	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
RW	0.98	1.09	1.20	1.38
AR	0.89	1.01	1.19	1.31
Panel C: PCE deflator inflation				
	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
RW	1.27	1.53	1.53	1.68
AR	1.12	1.31	1.38	1.49

This table discloses the absolute MAPE of the benchmark forecasts on three measures of inflation, across forecasting horizons of 0, 1, 4 and 8 quarters. The forecasts are constructed for the period from 1997Q1 until 2022Q4 and are based upon an expanding window of observations with data running back to 1948Q1.

C Results on forecast accuracy before the Great Recession

Table 10: Relative root mean square prediction errors of the alternative inflation forecasts with respect to the AR forecast, over the period before the Great Recession.

Panel A: CPI inflation

	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
RW	1.30	1.38	1.25	1.24
AR-gap	1.00	0.98	0.95	0.94
PC	1.03	0.95	1.00	1.02
PC-gap	1.00	**0.94***	0.95	0.95
UCSV	0.84**	0.95	0.95	**0.90***
Term structure VAR	1.02	1.04	1.04	0.98
SPF	**0.81*****	0.96	0.95	
ATSIX		0.96	**0.93***	**0.90***
RF	0.86*	1.03	0.95	0.97
LSTM	0.84**	0.98	0.98	0.92*

Panel B: GDP deflator inflation

	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
RW	1.10	1.13	0.91	1.01
AR-gap	1.01	0.98	0.78*	**0.81****
PC	0.99	0.98	1.00	1.02
PC-gap	1.01	0.95	0.78*	**0.81****
UCSV	1.20	1.09	0.93	0.94
Term structure VAR	1.06	1.09	1.08	1.14
SPF	**0.82****	**0.83****	**0.73*****	
RF	1.02	0.95	0.84	0.84*
LSTM	1.24	1.07	1.00	1.00

Panel C: PCE deflator inflation

	Horizon:			
Forecast	$h = 0$	$h = 1$	$h = 4$	$h = 8$
RW	1.15	1.20	1.00	0.92
AR-gap	1.03	1.01	0.83*	**0.78***
PC	0.98	0.98	1.01	1.02
PC-gap	1.02	0.98	**0.81***	0.79*
UCSV	1.04	0.99	0.95	0.94
Term structure VAR	1.07	1.08	1.03	1.07
RF	**0.96**	**0.91**	**0.81***	0.90
LSTM	1.15	1.02	0.95	1.03

This table discloses the RMSPE of the different forecasts on three measures of inflation, across forecasting horizons of 0, 1, 4 and 8 quarters. The forecasts are constructed for the period from 1997Q1 until 2007Q3 and are based upon an expanding window of observations with data running back to 1948Q1. The RMSPEs are reported as fractions of the RMSPEs of the AR model. The asterisks indicate the significance level of found accuracy gains with respect to the random walk model; one, two and three asterisks correspond to an accuracy improvement at the 10%, 5% and 1% significance level, respectively. The bold values denote the lowest RMSPEs across each forecast horizon.

D Description of the programming files

The attached ZIP-file contains the programming code files that I implement in my research. A description of each of these files follows below.

Data files for MATLAB code:

- CPI.mat: This MAT-file stores the CPI inflation rates.
- GDP.mat: This MAT-file stores the GDP deflator inflation rates.
- PCE.mat: This MAT-file stores the PCE deflator inflation rates.
- Beta_q.mat: This MAT-file stores the dynamic yield curve factors.
- RUC.mat: This MAT-file stores the unemployment rates.
- CPI_trend.mat: This MAT-file stores the proxy trend CPI inflation rates.
- PCE_trend.mat: This MAT-file stores the proxy trend PCE deflator inflation rates.
- forecastsSPF_mean_CPI.mat: This MAT-file stores the mean SPF forecasts of CPI inflation.
- forecastsSPF_mean_GDP.mat: This MAT-file stores the mean SPF forecasts of GDP deflator inflation.
- forecastsATSIX_matrix.mat: This MAT-file stores the ATSIX forecasts of CPI inflation.
- forecasts$X$_LSTM_matrix.mat: This MAT-file stores the LSTM forecasts of inflation measure X. Generated through PY files explained below.
- forecasts$X$_RF_matrix.mat: This MAT-file stores the RF forecasts of inflation measure X. Generated through PY files explained below.

MATLAB Functions:

- rw.m: This function generates the RW forecasts.
- ar_p.m: This function generates the AR forecasts.
- arGAP_p.m: This function generates the AR-gap forecasts.
- pc_p.m: This function generates the PC forecasts.
- pcGAP_p.m: This function generates the PC-gap forecasts.
- ucsv.m: This function generates the UCSV forecasts.
- termStructureVAR.m: This function generates the term structure VAR forecasts.
- dmtest.m: This function generates the augmented Diebold-Mariano test statistics. The ready-to-use function is shared by Ibisevic (2024) and publicly accessible.

MATLAB Scripts:

- mainResults_X.m: This script constructs the alternative forecasts of inflation measure X and subsequently computes the RMSPE, MAPE and bias of the different forecasts. Moreover, the script executes the augmented Diebold-Mariano test for the comparison of the alternative forecasts with a selected benchmark.
- DataSection.m: This script plots the figures from Section 3.

Data files for Python code:

- CPI_ML.csv: This CSV-file stores the CPI inflation rates.
- GDP_ML.csv: This CSV-file stores the GDP deflator inflation rates.
- PCE_ML.csv: This CSV-file stores the PCE deflator inflation rates.
- dataset_ML.csv: This CSV-file stores the FRED-QD data.

Python files:

- LSTM_DynamicFeatureSelection.py: This script generates the LSTM forecasts.
- feature_selection.py: This script plots the figures illustrating the Information Gain scores from Section 5.5.
- RandomForest_model.py: This module is implemented in tuning the hyperparameters of the RF model.
- forecast_RandomForest.py: This script generates the RF forecasts.