Inflation Persistence

Emanuele Franceschi*

September 2022

Preliminary & in progress,
do not circulate

Abstract

We investigate the evolution of persistence in post-WWII US inflation. Besides standard methods, we draw cutting edge methods from deep neural networks and leverage their flexibility in dealing with long trends and short swings to study inflation inertia. We consistently find evidence of decreasing persistence since mid-90's, also controlling for trend inflation and commodities influence. The decrease pre-dates the onset of globalisation forces, post-dates switches in monetary policy, and thus points to longer term transformations unfolding the is US economy, such as structural change.

Keywords: Inflation persistence, machine learning
***JEL* Codes:** E31, E37

1 Introduction

Inflation is one of the main topics in macroeconomics. Its origin, dynamic behaviour, and control have sparked immense strands of research, from microeconomics to forecasting. In the last decade, inflation was – and still is – part of a lively discussion on monetary policy.

During the 2010's decade, inflation has been unexpectedly low and stable in advanced economies. This low trend and mild volatility are even more baffling in light of the large swings in economic activity, commodities price, monetary and fiscal policies. This recent dynamics of inflation is puzzling both if one compares it with historical data and if one looks at the predictions of most macroeconomic theories. From a historical perspective, during the last two decades inflation has become at the same time harder and easier to predict (Stock and Watson, 2007): significantly less volatile than in the post-war period and yet well modelled by a white noise rather than more structured models. Furthermore, according to conventional theories based on simple Taylor rules, the new, unconventional tools

*PSE - Paris School of Economics, Paris 1 Pantheon-Sorbonne University. I wish to thank Fabrizio Coricelli for advice. I also wish to thank Ricardo Reis for kindly providing programs that served as base for Sect. 4. Jaime Montaña, Irene Iodice, Ilja Kuzborskij, Ilya Eryzhenskiy and participants to the Banque de France Chair Workshop at PSE, 1st Ventotene Macro Workshop provided excellent comments and suggestions. All remaining errors are solely Author's own.
Contact: emanuele.franceschi@gmail.com, emanuelefranceschi.com

adopted by central banks in response to the 2008 Global Financial Crisis could have had small or zero effect or have generated inflation spiralling out of control, as in the late '70s (see, for example, Taylor (2014)). In contrast to these predictions, the US economy posted its longest expansion since WWII, until COVID-19 upended it, in a context of moderate inflation.

This paper presents a wide-ranging empirical analysis on the dynamics of inflation mainly based on reduced-form models. We focus on the univariate properties of the inflation series since WWII, abstracting from an analysis of its determinants. Over this period, there were different phases characterised by various stages of structural change, varying degrees of trade openness, and different regimes for macroeconomic policies. To capture and exploit all information present in the data, we use five measures of price change covering consumption and production of goods and services in the US economy. Our contribution to the existing literature relates both to the sample and indicators of inflation, and on the methodology adopted. First, we study inflation persistence using longer time periods and several measures. Second, we implement relatively recent methodologies, including artificial intelligence, which have not yet been fully exploited in the analysis of inflation.

Our analysis starts from simple, but reliable, autoregressive models that impose a structural straightjacket to the data. We then move to more flexible tools, such as a Bayesian state-space autoregressive analysis and the model-free deep learning approach. While Bayesian tools are largely common in macroeconomic analyses, machine learning is still at an exploratory stage, although their use in macroeconomic analysis is rapidly growing (Athey and Imbens, 2019; Varian, 2014), especially because of their excellent forecasting performances (Almosova and Andresen, 2019; Makridakis, Spiliotis, and Assimakopoulos, 2018). In this paper we exploit the capabilities of these tools in order to identify with finer granularity the non-linear properties of inflation. Once the neural network is presented and trained with data, we use its forecasting properties to generate additional data points and assess more precisely how persistence has changed over time. We restrict our attention to inflation persistence and its dynamic changes. A broad definition of persistence relates to inertia, which is the property of an object to not deviate from its past dynamics in absence of external shocks. A highly persistent time series posting a 5% growth rate will likely move in such value's neighbourhood if nothing affects it. On the other hand, when the inertial series is hit by a shock, it will slowly incorporate and dissipate the shock over time. Similarly, weakly persistent series will display more variability and shocks will be depleted relatively quickly. An intuitive implication of inertia is predictability, which goes hand in hand with persistence.

The question of inflation persistence is particularly relevant for monetary policy. Assessing the sensitivity of inflation to changes is crucial for central banks when planning policy interventions: how aggressive should the intervention be to undo an inflationary spiral? Or symmetrically, what is the optimal timing for rate increases during a recovery or expansion? As monetary policy operates through lags, how much will it take for a shock to be transmitted to observed inflation? The degree of persistence also influences the tradeoff between inflation and economic activity: if inflation is persistent and far from the target, it will require corresponding larger output gaps. Likewise, the analysis of persistence provides information on how the sectoral structure of the economy, technology, international finance and trade affect the country's inflation. Analysing inflation persistence also sheds light on the process of price formation: is persistence stable? If not, how does it vary over

time and why? Is there a mutual influence with measures of output persistence and volatility? Moreover, as shown by the large body of studies on the Taylor Rule parametrisation and its changes over time, it is still unclear whether these changes affect in any measurable manner the dynamics of inflation.

Understanding the dynamic properties of inflation improves the decision making of central banks in two crucial ways: before the policy decision, an extended information set helps to calibrate the intervention; and after the policy decision it helps to evaluate its effectiveness. Central banks need to assess whether the sources of movements in inflation are inherited from deep, structural sources like price-setting strategies or, alternatively, whether are due to transient shocks to commodities prices. This is in turn useful also to evaluate the time lag between policy changes and changes in inflation, or the length of time needed to achieve the inflation target. For governments, the knowledge of inflation dynamics is important to design their fiscal policy. For example, with highly inertial inflation, a VAT increase will take several quarters to be fully transferred to final consumer prices. Moreover, if inflation, as in a traditional Phillips Curve, inherits its inertia from output, fiscal authorities might improve the global policy mix with the monetary authority.

Recent research highlights the interplay between trend inflation, inflation target, and persistence, see for example Cogley and Sbordone (2009), Kurozumi and Zandweghe (2019), and Stock and Watson (2007). To account for such interplay, throughout the paper we control for trend inflation and for time-varying trends; nevertheless, such investigation is outside the scope of this study.

Literature overview

Inflation is a central theme in macroeconomics. We therefore contribute to a long and rich literature. Recent inflation dynamics in developed economies is the focus of Ciccarelli and Osbat (2017), Coibion, Gorodnichenko, and Kamdar (2018), and Miles et al. (2017), who apply diverse frameworks but overall report low and stable inflation rates since the 2000. The root causes for such dynamic behaviour are studied in three complementary strands of literature. A large number of studies focuses on the expectations in a Phillips Curve framework. A comprehensive overview of empirical strategies to estimate the effects of inflation expectations in the Phillips Curve is Mavroeidis, Plegborg-Moller, and Stock (2014), which emphasise the uncertainty and difficulties in precisely pinning down a robust specification. On the other hand, Coibion and Gorodnichenko (2015) and Coibion, Gorodnichenko, and Kamdar (2018) propose mechanisms of expectations formation to explore how these affect realised inflation.

An additional cause for inflation dynamics is found in the integration into global value chains, which ease the transmission of foreign shocks: Auer, Borio, and Filardo (2017) and Bianchi and Civelli (2015) fall in this line of research and find evidence of global inflation effects. These effects generally increase with openness but are stable over time. Along these lines, Jarociński and Bobeica (2017) augment a VAR with domestic and global factors to solve the twin puzzle of missing both disinflation and inflation in the Euro Area during the 2008 recession. They find that domestic factors counteracted global ones in the EA and can explain the missing inflation leg of the twin puzzles.

More generally, a third strand of literature has focused on the interplay of monetary policy regimes, inflation, and volatility shocks. Fernández-Villaverde, Guerrón-Quintana, and Rubio-Ramírez (2010, 2015) estimate DSGE models with stochastic volatility and eval-

uate the role of shocks and policies in setting off the Great Moderation. Their findings point towards a minor role for policy in the steady dynamics of aggregates during the 1984-2007 period. Our approach is closely related to Pivetta and Reis (2007) and Fuhrer (2011): the former study inflation dynamics building on Cogley and Sargent (2002, 2005) with a flexible Bayesian approach. The latter offers a review of the state of the art in terms of measures, methods, and theories to evaluate inflation dynamics. The bottom line of both studies, though, is that inflation persistence is relatively stable in the post-WWII period, although both studies predate the 2008 recession and the ensuing policy innovations.

A series of interrelated studies investigated the dynamics of the inflation gap, which is the deviation from trend. Benati and Surico (2008) and Cogley, Primiceri, and Sargent (2008) estimate VAR models with a focus on predictability. Overall they find that US inflation has become less predictable and argue that this broadly corresponds to a more aggressive monetary stance or a change in the inflation target.

We extend the above analyses by using statistical learning tools such as machine learning, which are growing in empirical economic studies. Forecasting is one of the main uses of statistical learning tools. (Jung, Patnam, and Ter-Martirosyan, 2018; Kock and Teräsvirta, 2016; Makridakis, Spiliotis, and Assimakopoulos, 2018; McAdam and McNelis, 2005; Medeiros et al., 2019). In this respect, Nakamura (2005) employed plain neural networks to forecast inflation, while Almosova and Andresen (2019), in close relation to our work, compare recurrent neural networks against workhorse forecasting models to predict monthly inflation at several horizons. Researchers have previously ventured in adapting artificial intelligence to macroeconomic applications (Bajari et al., 2015; Chakraborty and Joseph, 2017; Giannone, Lenza, and Primiceri, 2018; Goulet Coulombe et al., 2019; Korobilis, 2018), econometrics (Athey, 2018; Athey and Imbens, 2015, 2019; Mullainathan and Spiess, 2017; Varian, 2014), or asset pricing (Gu, Kelly, and Xiu, 2020). Lastly, promising applications of sophisticated machine learning models have been proposed in computational economics, see for example Fernandez-Villaverde and Guerron-Quintana (2020), Fernández-Villaverde, Hurtado, and Nuño (2020), and Maliar, Maliar, and Winant (2019) who offer ML-based numerical solutions to DSGEs, or more generally Rackauckas et al. (2020) who incorporates ML for numerical solution of complex dynamic systems.

The rest of the paper is organised as follows: Section (2) presents data, Section (3) overviews the empirical tools and resulting evidence from a plain autoregressive, frequentist approach, Section (4) describes the Bayesian take on persistence, Section (5) presents the results using the machine learning tools and results; finally Section (6) concludes.

2 Data and tools

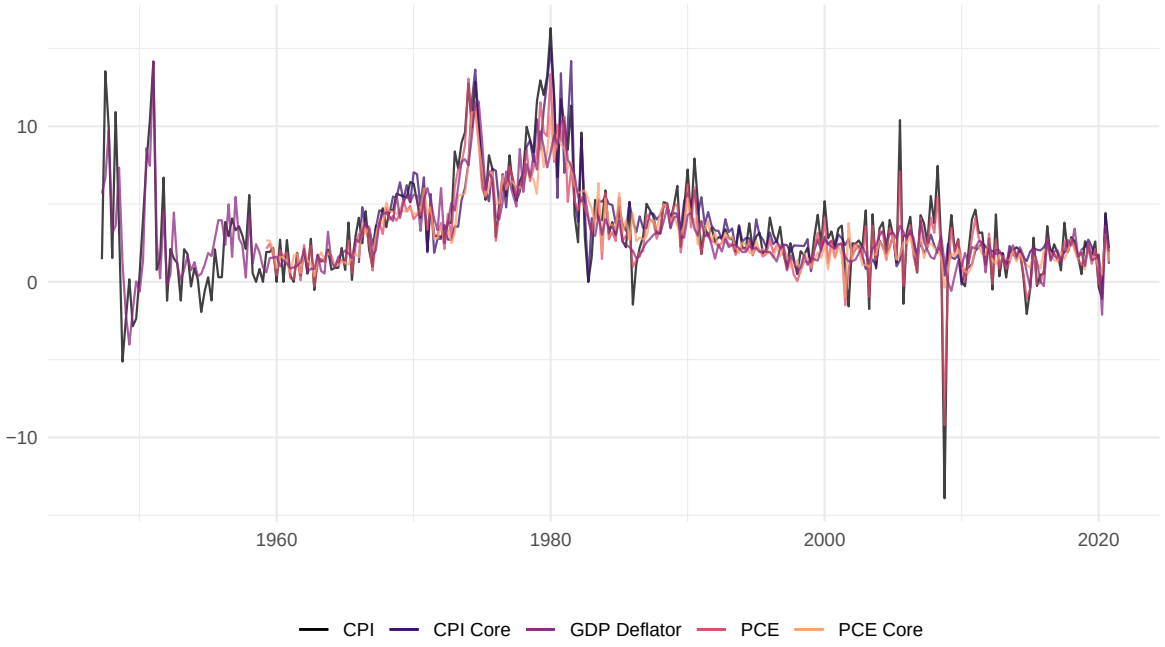
While traditionally only one series is used in analysing inflation persistence, we consider three classes of inflation indexes for the US economy: the Consumer Price Index (CPI), the Personal Consumption Expenditure index (PCE), and finally the Gross Domestic Product Deflator. These three indexes are measured on different baskets of goods, hence discrepancies and deviations are due to the distinct subset of goods and services each index tracks. More precisely, the CPI mainly relates to consumers purchases, the PCE captures business sales, while the GDP deflator is measured on the goods and services produced domestically, abstracting from "imported" inflation. CPI and PCE also differ in the weights for each good and are available as "headline" and "core", with the latter excluding volatile

items like food, energy, and commodities. We cover almost entirely the post-WWII period, as series span 1948Q1:2020Q1 for CPI and GDP deflator, while PCE indexes start in 1960Q1. All series considered are historically revised to track as closely as possible the actual change of prices.[1]

Researchers interested in monetary policy usually prefer PCE and the GDP deflator, as the former is the explicit target of the Federal Reserve Bank (Cogley and Sargent, 2005; Cogley and Sbordone, 2009), while forecasting and statistical analyses often rely on the CPI (Fuhrer, 2011; Pivetta and Reis, 2007). We cover the whole range of indexes to capture common trends in the dynamics of aggregate inflation.

All the series are sourced at a quarterly frequency from the FRED database of the Federal Reserve Bank of St. Louis. We take the raw level of the indexes and compute annualised quarter-on-quarter percentage change, to account for slow-moving trends in the data.[2] Throughout the analysis, our preferred measure is the GDP deflator index, since by design it tracks closely the variation in prices of goods and services produced and supplied within the US economy. This feature allows us to track more closely the underlying macro dynamics. Moreover, it also provides us with more observations, which can be used for checking the robustness of our results. Figure 1 plots all series that we will analyse.

Figure 1: US Inflation Data

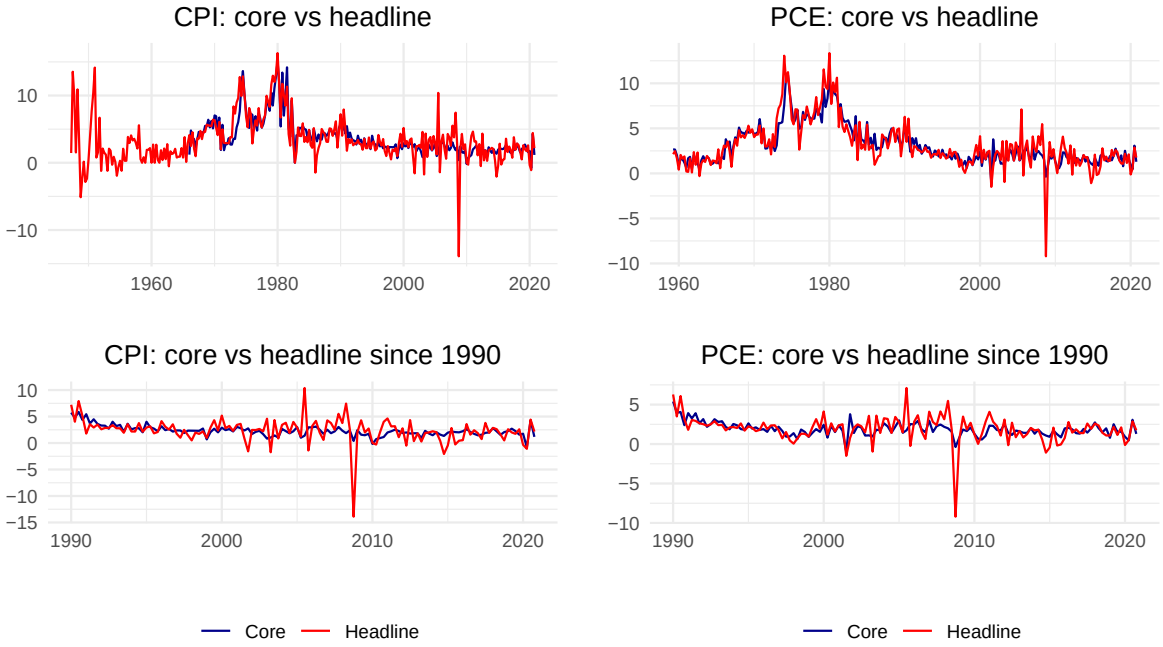


Inflation series: CPI, Consumer Price Index, including and excluding Food and Energy prices (starting 1947Q2, 1958Q4); PCE, Personal Consumption Expenditure index, including and excluding Food and Energy (1959Q2, 1959Q2); US Gross Domestic Product deflator (1947Q2). All series are computed as annualised log differences from previous quarter and end in 2020Q1.

[1]For Core CPI we drop the observations up until 1966, since those are interpolated from lower frequency data and thus carry very little signal to noise ratio.

[2]Quarter on quarter annualized percentage changes are left to the Appendix for comparison: overall trends do not vary significantly, whilst the levels change mildly.

Figure 2: Headline and Core Inflation



Top panel: full sample of headline and core series; bottom panel: headline and core series since 1990.

One stark fact emerges at a simple glimpse of the series. While for most observations headline and core inflation measures move hand in hand, they appear to diverge in volatility after 2000. To better tell apart these discrepancies, Fig.(2) compares headline and core measures for CPI and PCE over the whole sample and since 1990.

This difference in volatility is, at a first pass, due to commodities prices, which sharply fluctuated after 2000. The clearest example is the oil price, which posted threefold increases and contractions since early 2000s (Miles et al., 2017).[3] Core inflation series are unaffected by these swings, as they exclude commodity prices. Therefore, analysing both headline and core series permits to isolate shocks, including their dynamic implications, arising from fluctuations in food and energy prices.

Tools and methods

Distinguishing core and headline inflation does not eliminate other sources of persistence in inflation, in particular the main channels working through forward-looking inflation expectations and the interactions with the level of economic activity, as postulated by several Phillips curve specifications:

$$\pi_t = \beta E_t \left(\pi_{t+i}\right) + \omega \hat{y}_t + \epsilon_t \tag{1}$$

Thus, assuming errors are a zero-mean, drift-less iid process,[4] inflation inertia is fully *inherited* from the dynamics of the output gap $\hat{y}$ actualised at t. Several studies focus separately on expectations (Coibion and Gorodnichenko, 2015; Coibion, Gorodnichenko, and

[3] Fig.(10) in the Appendix plots price level and change for the West Texas Intermediate since mid-80s; Fig.(11) plots the same metrics for a global commodities index.

[4] Potential measurement errors are alike, and omitted in the equation.

Ropele, 2019) or output gap (Mavroeidis, Plegborg-Moller, and Stock, 2014) to explain inflation persistence. The former find that market-implied, consumers', and professional forecasters' expectations drifted apart and match differently the actual value of inflation (Mankiw, Reis, and Wolfers, 2003; Trehan, 2015). In this paper, we take a policymaker point of view and solely assess the transmission of shocks to actual inflation, disregarding effects on expectations or confounding sources such as movements in the output gap. This is clearly a reduced-form approach, but still informative on the underlying dynamics of inflation. Indeed, more general approaches are required to at least match what is observed in simple frameworks such ours.

We employ three classes of models to analyse inflation inertia, spanning increasing levels of refinement and complexity. The first step is a simple autoregressive approach with varying lags. Then, the same structure is extended in a Bayesian framework, and finally we fit a recurrent neural network borrowed from statistical learning. In all parts we measure inertia as the first order serial correlation or the sum of all autoregressive coefficients. These statistics convey enough information that allows us not only to judge inflation persistence but also to complement other common metrics, for example, the largest autoregressive root (LAR) or the halflife of a shock (both employed in Fuhrer (2011) and Pivetta and Reis (2007)).[5]

3 Autoregressive analyses

As a first step to test whether inflation inertia has varied significantly over time, we estimate a simple $AR\,(1)$ model. To capture such variations, we estimate the $AR\,(1)$ model on a 56-quarter rolling window, in line with Fuhrer (2011) and Pivetta and Reis (2007). Unstable estimates or large swings from such exercise would substantiate further analyses, aimed at decomposing the varying weight of past inflation on current price change. For this purpose, we consider inflation observations as drawn from the following process:

$$\pi_t = \beta_{0,t} + \beta_{1,t}\pi_{t-1} + \varepsilon_t \tag{2}$$

This barebone model represents the benchmark for our analysis. In such a framework β_0 represents the steady-state or trend inflation rate, while β_1 encapsulates any form of intrinsic inflation autocorrelation. We will primarily focus on $\beta_{1,t}$, without direct consideration of the intercept. Its consideration in our exercise would require the explicit modelling of trend inflation, which is out of the scope of this work.[6] The error term ε, in this case, mops up new disturbances of any sorts, from expectations to technology, mark-up, demand shocks hitting inflation at time t.

In workhorse modern macro monetary models, the process generating inflation hinges mainly on expected inflation and disturbances in the current output gap: the Phillips curve, in this case, reads as in eq.(1) (Walsh, 2003; Woodford, 2003) and it is fully forward-looking with $i > 0$. The only sources of persistence are the serial correlation of shock to technology and the degree of price flexibility, both originating in the supply block of the economy

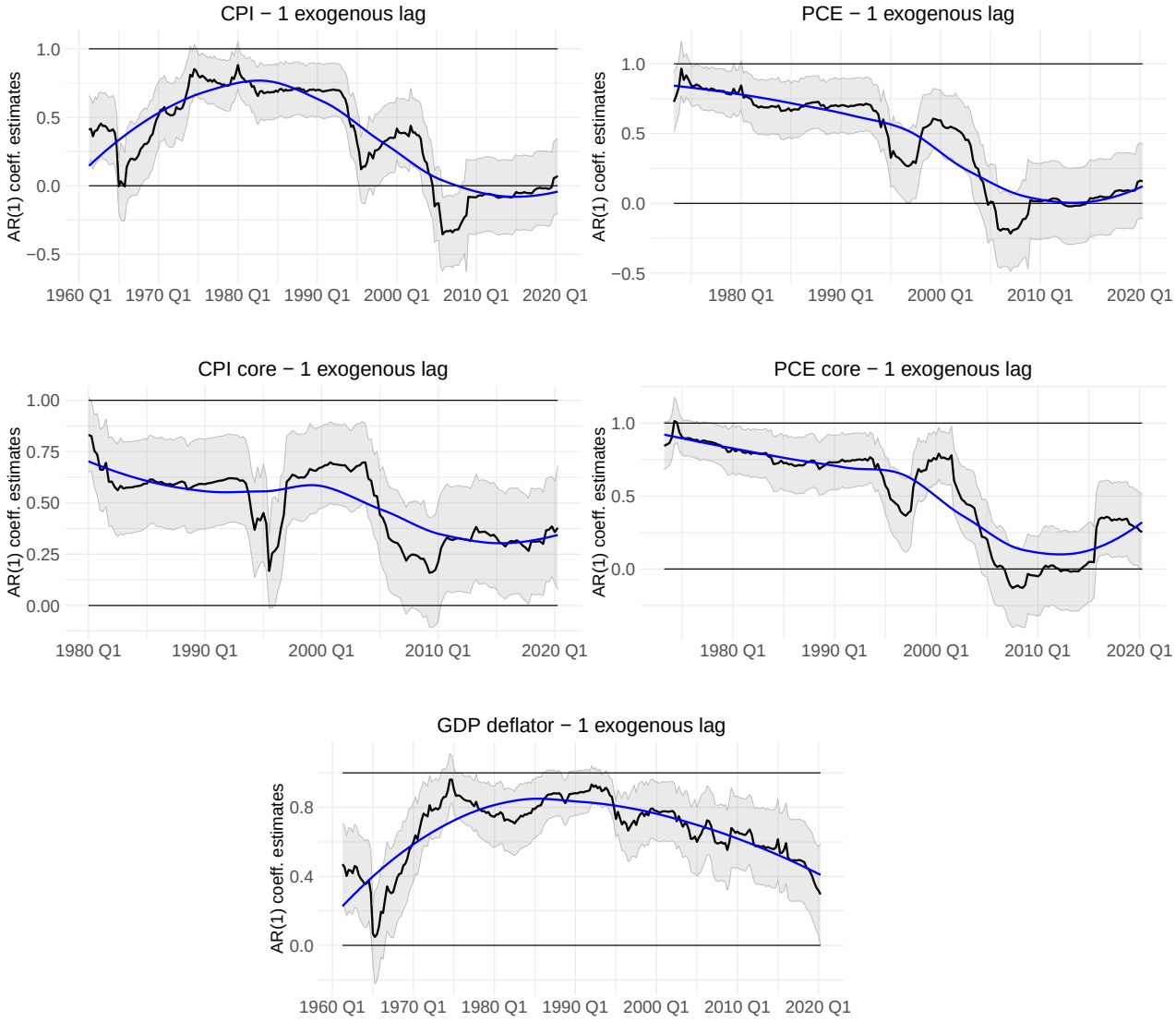
[5] Ideally, though, one would further add measures that are not easily summarised in one scalar, such as the variation of the autocorrelation function over time or the decomposition of permanent and transitory shocks' variances as presented in Stock and Watson (2007).

[6] Nevertheless, Appendix G complements with the full range of trend inflation estimates.

subsumed in $\hat{y}$. This basic framework can be augmented by adding trend inflation and fluctuations around it.[7] In empirical studies this extension helps to bridge the gap between observed and theoretical behaviour of inflation and links central banks' inflation targeting and short term fluctuations around such target.

Fig.(3) collects the estimates from the $AR\,(1)$ process previously described, using the GDP deflator, CPI, and PCE indexes.

Figure 3: $AR\,(1)$ Persistence



$AR(1)$ estimates for CPI, PCE, GDP deflator series, rolling window of 56 quarters (14 years). Black line plots $\beta_{1,t}$ point estimates over time, red lines are 5% confidence intervals, blue line is LOESS fit to polynomially smooth out point estimates, grey bands are its 5% confidence intervals. Headline series are on the top row, core are on the middle row.

[7] Cogley, Primiceri, and Sargent (2008) and Cogley and Sbordone (2009) provide a framework for modelling and estimating time-varying trend inflation and the inflation gap, respectively. Cogley, Primiceri, and Sargent (2008) find that the inflation gap display little persistence, while Cogley and Sbordone (2009) show that allowing trend inflation does away with inflation indexation in NK models.

At a first glance, some regularities emerge: for all series analysed report sensible variation in $\beta_{1,t}$, alongside a generalised downward trend. This trend peaks roughly in the mid-90s in all series considered before decreasing at varying speeds. A 56-quarter window implies that the estimates for this period are based on a subsample that just excluded observations from the early 80s, a period of structural change and monetary (unexpected) intervention, namely the Volcker policy shift. Interestingly, the mid-90s estimates show high levels of variation, in comparison with the rest of the estimates: values for $\beta_{1,t}$ drop significantly before climbing back on trend, common to all five panes, likely reflecting the switch induced by Volcker. The switch takes the form of a debasing of inflation from the previous trend, thus erasing dependence on past realisations.

Consistently with the increased variability displayed by headline series (as opposed to core ones excluding food and energy) the CPI series (left column) display relevant differences between headline (top left) and core series (mid left). This pattern is less evident in the PCE series, which in turn display a similar profile over time both in trend and magnitude. In three cases out of five, zero is included in the confidence interval roughly from 2005, implying a white-noise process.

Overall, these estimates point to a decreased inflation inertia, with significant drops taking place since the 2000s. This is particularly stark for CPI and PCE series but less clear for the GDP deflator: the first autocorrelation coefficient for this latter series starts decreasing in the mid-1990s and five years later displays a mild acceleration, with sensibly more smoothness than other indexes.

Such widespread dynamics allows the exclusion of commodity prices as the root cause of decreased inflation inertia, but the timing of the switches hints at factors like international trade shocks (Autor, Dorn, and Hanson, 2016; Bianchi and Civelli, 2015). Specifically, increasing economic integration at a global scale might foster the transmission of international shocks into domestic inflation, as argued by Auer, Borio, and Filardo (2017). This argument is corroborated by the fact that the GDP deflator displays a slightly different, and slower, decrease, tracking more closely the US national production.

The evidence offered by this simple analysis begs further investigation on the behaviour of inflation and its persistence. A more refined approach within the frequentist domain consists in extracting more information from the inflation time series by using an optimally chosen number of lags. This approach is applied in the next section.

3.1 Optimal lags selection

A natural step forward consists in adding more lags to the model we estimate. This addition allows a better framing of inflation persistence, since longer lags can capture dependence on realisations farther in the past. Two issues arise when comparing multiple lags estimates, though. Firstly, it is not clear whether a process with two lags like .7 and .2 is more, equally, or less persistent than a process with three lags, like .5, .4, .3. To circumvent this issue, we sum over the coefficients, compounding together all estimates. In this way we can compare a measure of persistence independently of the number and magnitude of the single parameters. Secondly, this approach allows for heterogeneity in the lags number for each series. We exploit this feature and compute, on the whole sample, the number of lags that minimises the Bayesian Information Criterion. Formally, the assumed process for inflation is

$$\pi_t = \beta_{0,t} + \sum_{i=1}^{k^*} \beta_{i,t}\pi_{t-i} + \varepsilon_t \quad \text{with} \quad k^* = \text{argmin}_{k \leq \bar{k}} BIC\,(k; 1, \ldots, T) \tag{3}$$

where $\bar{k}$ is set to 18 quarters as an upper bound to the number of admissible lags. Conversely, we measure inflation persistence as follows, as presented by Fuhrer (2011) and Pivetta and Reis (2007):

$$\rho\,(k^*) = \sum_{i=1}^{k^*} \hat{\beta}_i \quad \text{with} \quad \beta_i = \frac{E\,(\pi_t \pi_{t-i})}{V\,(\pi)} \tag{4}$$

where k^* is computed on all available observations, at this stage. We estimate the $AR\,(k^*)$ process with a rolling window in order to study how $\rho\,(k^*)$ evolves over time, using again a width of 56 observations. Tab.(1) presents the optimal lags obtained for each series.[8]

Table 1: Optimal lags selection via BIC

GDP Defl.	CPI headline	CPI core	PCE headline	PCE core
3	3	3	3	2

Fig.(4) shows that the same, generalized downward trend in persistence is found even when more lags are included in the model for inflation dynamics. The values reported are all in the same neighbourhood, corroborating the evidence of initially high but decreasing persistence.

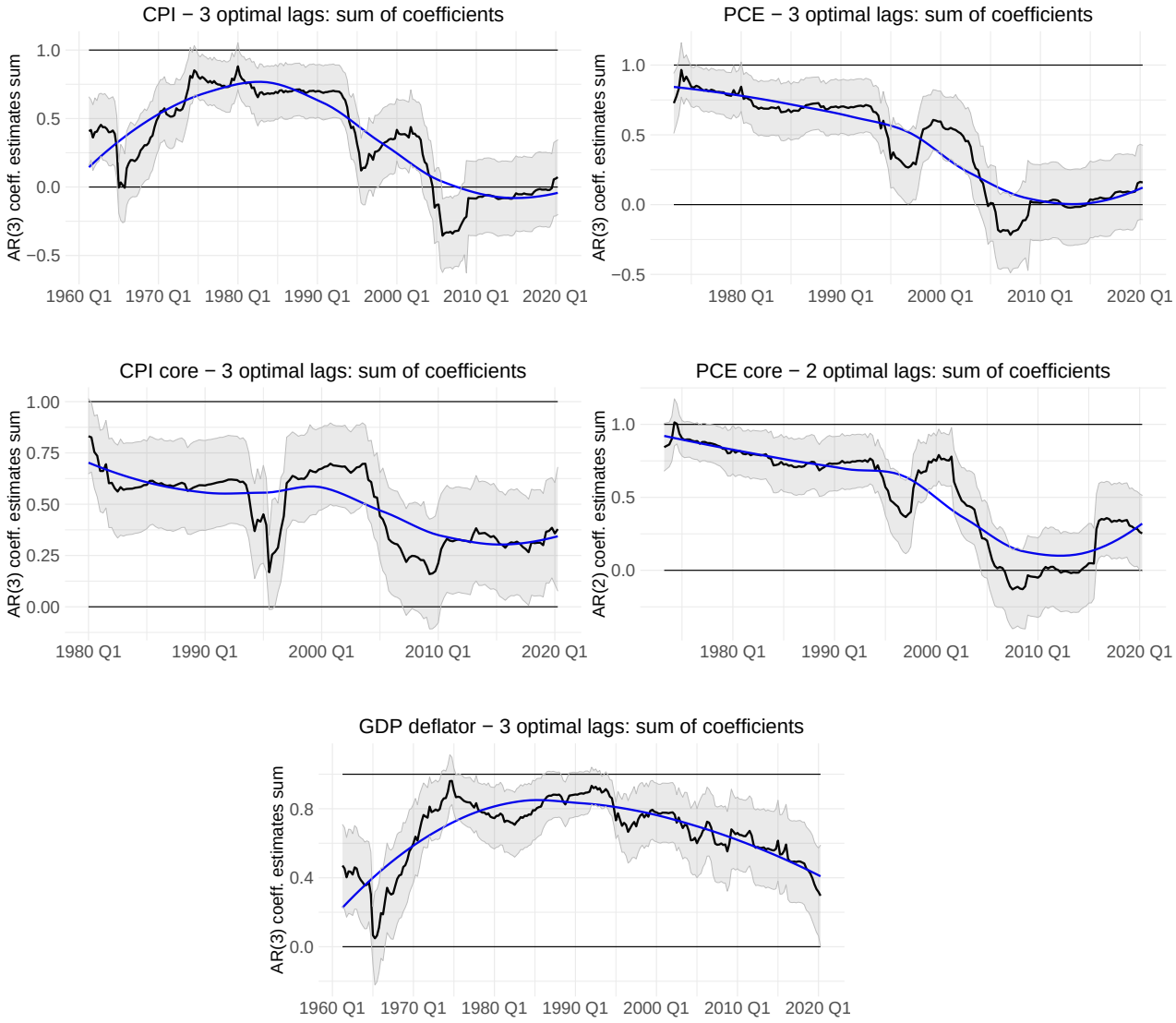
Individual profiles do not differ much from previous plots, with the GDP deflator showing the slowest downward trend in the series, possibly due to slow transformations taking place within the US economy's composition and percolating onto prices. Although less stable than the previous case, the sum of autoregressive coefficients still reports a relevant drop around the mid-90s, when observations associated with Volcker's initial period are phased out of the rolling window.

Again, a sharp decrease in inertia takes place around the year 2000 for CPI and PCE series (both core and headline). The sharp fall emerges from observations from the beginning of the Great Moderation, but for most series is followed by a modest rebound upwards. This last movement, though, does not fully offset the previous decrease and sets inertia on relatively lower values. The GDP deflator, consistently, follows a smoother, hump-shaped path, peaking during the 1970-90 period, with much less volatile point estimates and tighter error bands.

According to these slightly more refined analyses, inflation has become less persistent over the decades and has drastically accelerated this process during the last two decades. The pattern is consistent with two simple methods and is confirmed when we exclude volatile commodity prices, as in core series. The GDP deflator, which tracks more closely the economic activity in the US economy, displays a much smoother dynamics of inertia.

[8] We propose a similar table in the Appendix, including year-on-year growth rates, Tab.3. While year-on-year and annualised quarter-on-quarter series are computed on the same raw data and method, the former present significantly higher levels of autocorrelation: the optimal lags numbers are in all cases between 9 and 18.

Figure 4: $AR\,(k^*)$ Persistence



$AR\,(k^*)$ estimates for $\rho\,(k^*)$, on a 56-quarter rolling window. Black line plots $\rho\,(k^*)$, grey bands are sums of SEs, blue line is LOESS polynomial fit to smooth out point estimates.

All in all, there is room for deeper investigation into the behaviour of inflation dynamics with refined methods. Next section uses a Bayesian approach to deepen the analysis and to exploit more of the information present in the data.

4 A Bayesian estimation of inertia

Adding a layer of sophistication to our inquiry, Bayesian methods helps in efficiently use data information, providing distributions of per-period measures of persistence. To this end, we adapt the approach illustrated in Pivetta and Reis (2007). We build upon this work operating on two margins: first, we use longer time series and, second, we run the

estimations on five series rather than only one. Throughout this exercise, we set the lags to three, consistently with frequentist analyses outlined above.

The main framework of this section dates back to Cogley and Sargent (2002, 2005), subsequently extended in Pivetta and Reis (2007) to allow for degenerate, unit root draws. The assumed state-space model consists of the following components:

$$\begin{aligned} \pi_t &= \beta_{0,t} + \sum_{i=1}^{3} \beta_{i,t} \pi_{t-i} + \varepsilon_t \\ P\left(\beta_{t+1} | \beta_t, V\right) &\propto I\left(\beta_{t+1}\right) MVN\left(\beta_{t+1} | \beta_t, V\right) \\ \implies \beta_{t+1} &= \beta_t + \nu_{t+1} \\ \text{with } \operatorname{var}(\nu) &= Q \end{aligned} \tag{5}$$

where the first equation is the measurement equation we also estimated in the previous section, the second line is the (hidden) state evolution, evolving as a multivariate normal distribution. β_t stacks all parameters at time t, $\beta_t = [\beta_{0,t}, \beta_{1,t}, \beta_{2,t}, \beta_{3,t},]'$. The state equation has the density of the parameters vector β depend on two components, an indicator function I that can be used to optionally exclude unit root draws and a multivariate normal density conditional on past draws of β, with *constant* covariance matrix V. The third line, implied by the Gaussian density, establishes that autoregressive parameters evolve over time as driftless random walks, potentially with unit roots.[9]

In this framework, β values are the model parameters, while the hyper-parameters are collected in the covariance matrix V. This latter gathers the co-variances of measurement and state equations:

$$V = \begin{bmatrix} \sigma_\varepsilon^2 & C' \\ C & Q \end{bmatrix} \tag{6}$$

where $\operatorname{var}(\varepsilon_t) = \sigma_\varepsilon^2$ is the variance of innovations in the measurement equation and Q is that for the state equation. C captures the covariance of measurement and state disturbances, set to zero.

To initialise β we use the first ten years of observations for each series, then the model is estimated on the remaining observations.[10] These estimates are collected in $\left(\bar{\beta}, \bar{P}, \bar{V}, T_0\right)$, with $\bar{\beta}$ and $\bar{P}$ being the OLS based mean and variance of a Gaussian distribution, and $\bar{V}^{-1}$ and T_0 are scale matrix and degrees of freedom of a inverse-Wishart distribution, respectively. Therefore, the prior distribution on β_0 is a draw from the following joint prior

$$P\left(\beta_0, V\right) \propto I\left(\beta\right) MVN\left(\bar{\beta}, \bar{P}\right) IW\left(\bar{V}^{-1}, T_0\right) \tag{7}$$

After initialisation, the algorithm obtains draws covering the past posterior distribution of states and hyper-parameters. These are then used to compute conditional future paths

[9] Ideally, further extensions accommodating for time-varying innovations would provide additional insights on the evolution of shocks and uncertainty in the economy. As a reference, see Bianchi (2013) and Lhuissier (2018), who develop estimated DSGE models with regime-switching uncertainty in volatility. Cogley and Sargent (2002) acknowledge such limitation in their work and tackle stochastic volatility in Cogley and Sargent (2005).

[10] Looking back at Fig.1, though, this step produces potentially biased hyperparameters, in light of the stark heterogeneity with the rest of the sample.

for inflation and state. The final step computes persistence measures on these simulated paths. At each period t, conditional on past information, we simulate distributions for the next 120 periods.

To produce such simulations, we need to draw from the following posterior distribution:

$$P\left(\Pi^{t+1,t+h}, \beta^{t+1,t+h}, \beta^t, V|\Pi^t\right) \tag{8}$$

with Π^t collecting all observations until t. This posterior density can be separated into past and present beliefs and future uncertainty, conditional on time-t information, as follows:

$$\begin{aligned} P\left(\Pi^{t+1,t+h}, \beta^{t+1,t+h}, \beta^t, V|\Pi^t\right) = & \overbrace{P\left(\beta^t, V|\Pi^t\right)}^{\text{beliefs on past and present}} \\ & \times \underbrace{P\left(\Pi^{t+1,t+h}, \beta^{t+1,t+h}|\beta^t, V, \Pi^t\right)}_{\text{future uncertainty}} \end{aligned} \tag{9}$$

The first block can be sampled via a Gibbs sampler. Draws from this sampler are later used to simulate future trajectories conditional on data-informed beliefs up to time t. Additional details on the algorithm to sample from such posterior density are presented in Cogley and Sargent (2002) and Pivetta and Reis (2007).

We set 300 thousand total draws, with a burn-in of 150 thousand to deal with path dependency and to ensure convergence to the posterior. Therefore, we use in our computations of persistence 150 thousand actual draws. Future paths are simulated up to a 120 quarters horizon, equivalent to thirty years of synthetic history. In line with Pivetta and Reis (2007) but in contrast to Cogley and Sargent (2002), we do not rule out explosive roots, so to report the complete, rather than the truncated, distribution of βs.

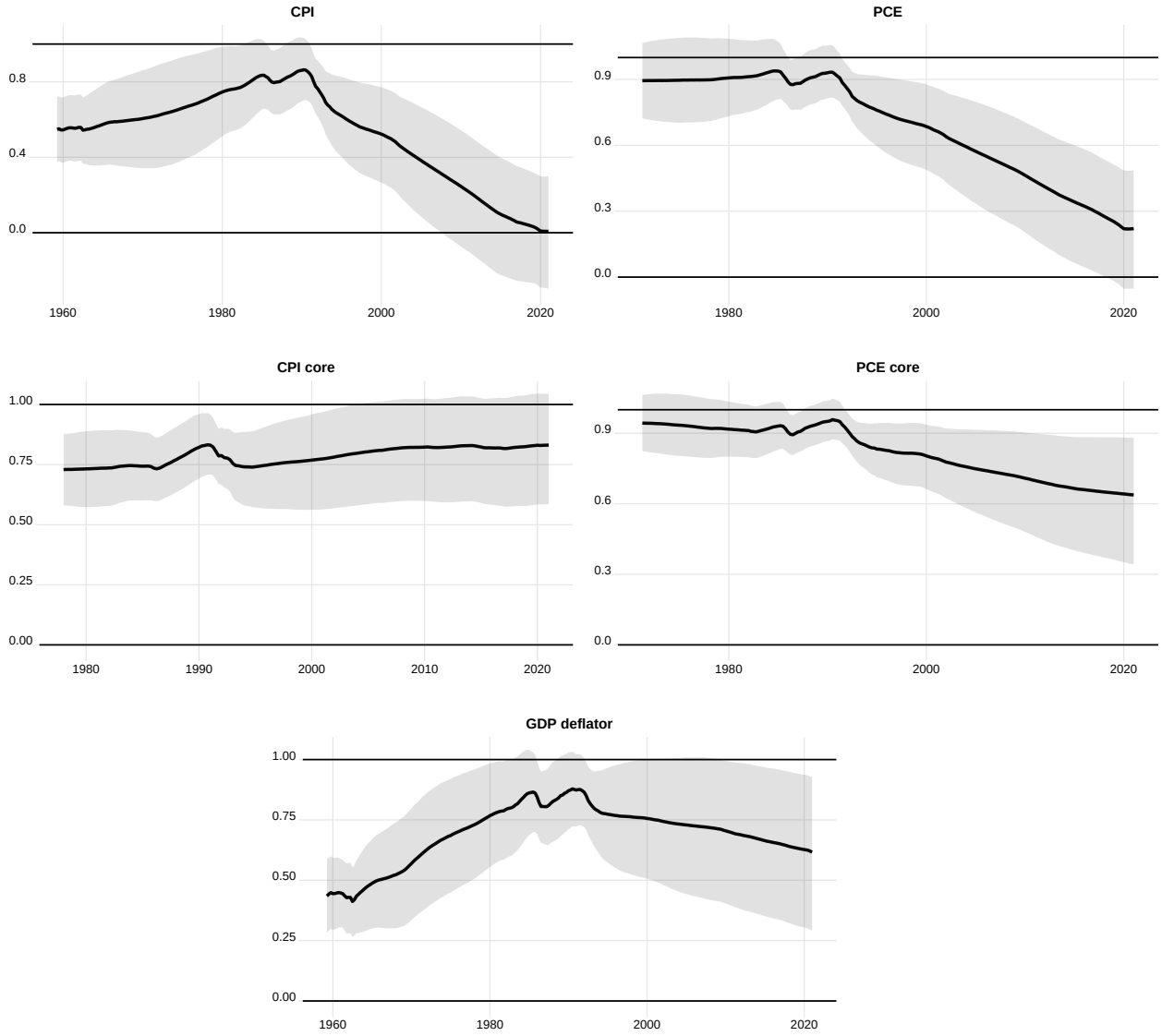
The results are plotted in Fig.(5) for all series in exam, together with upper and lower 5% credibility intervals.[11] The advantage of the Bayesian setup is to summarise all information on persistence at time t from both data and prior in order to extend the sample simulating a wealth of future, consistent paths. At any date t the state incorporates all information available in the past observations. Consistently with this information, the space of future realisations is duly explored far into the future and covering large swaths of the domain of the distributions. Therefore, it pins down more precisely the measurement of inertia at time t rather than relying on a rolling window approach with a fixed number of observations.

The output of this Bayesian exercise broadly corroborates the findings of simpler, frequentist approaches, with some significant departures. Overall, swings in inflation inertia are smoother and more gradual, unfolding over the whole length of the sample.

Scrutinising the general pattern, though, one can easily make contact with previous results: inflation inertia decreases from relatively high levels. Point estimates report generally higher values than Fig. (3) and 4, notwithstanding a common temporal path. The peak is generally reached in the mid-90s, which backdates slightly the onset of the decrease and somewhat weakens the international trade cause for such decline, as China officially entered WTO only in 2001 (Autor, Dorn, and Hanson, 2016). Moreover, this decline appears to be preceded by a phase of increase, as suggested in previous analyses. Core CPI persistence, though, stands out: it shows a rather stable path, if not slightly increasing.

[11] Full draws distributions per quarter are deferred to Appendix (D).

Figure 5: Bayesian Estimates



Sum of $AR\,(3)$ coefficients calculated on the simulated future paths for CPI, PCE (both with and without food and energy), GDP deflator. Black solid lines are the median values, shaded grey areas are 95% credibility intervals. Total draws 300000, burn-in 150000, thus 150000 final draws, per period t.

A second common pattern is an increasing uncertainty around the median: in most series credibility intervals widen visibly toward the end of the period, with GDP deflator's and core series' intervals covering much of the unit space, in line with the higher unpredictability of inflation put forward by Stock and Watson (2007).

This unpredictability is even corroborated as credibility intervals reach zero for the headline CPI and PCE. This levelling down corresponds to the flattening displayed in Figs.(3) and (4), suggesting that after the Global Financial Crisis inflation is much closer to white noise than to an autoregressive process. In turn, it is necessary to couple such reduced form analysis with more interactions with other economic forces to fully unbundle the effects of economic slack, fiscal and monetary policies, international spillovers. A first

sophisticated step towards such setup is carried out in Fernández-Villaverde, Guerrón-Quintana, and Rubio-Ramírez (2010, 2015), which estimate a rich DSGE with volatility shocks and study whether monetary policy switches had more effects than reduced variance on taming inflation.

As a side note, an advantage of this Bayesian approach is the potential to extend this kernel to include more structured and informed models. In fact, Cogley, Primiceri, and Sargent (2008) and Cogley and Sargent (2005) do take this approach to structural models.

5 RNN-LSTM approach to persistence

Based on our reduced form perspective, the lag structure and possible non-linearities play a crucial role. For instance, inflation may display a slow-moving drift that affects realisations at sensibly long horizons. To tackle this possibility, we borrow from a class of models that are precisely designed to handle long, short, and time-varying lags in a flexible and dynamic way.

Long Short-Term Memory models (LSTMs) are machine learning algorithms that exploit the structure and advantages of Recurrent Neural Networks (RNNs). This class of essentially non-parametric models has the advantage of effectively handling a very large set of functional forms under mild regularity requirements (Kidger and Lyons, 2019; Leshno et al., 1993; Tabuada and Gharesifard, 2020). The downside is the infamous black-box nature and the complexity of the inner mechanisms: the resulting estimated network can hardly provide intuitive insights on the connections and relations between data points. Conversely, though, machine learning models produce reliable results in terms of fit and forecasts.

The broader class of neural networks (NN) does not keep up well with persistence, as these models conserve little information about data with potentially long time dependencies. This issue is known as the "vanishing gradient" problem. It originates with the back-propagation algorithm, which, in a nutshell, is an efficient way to minimize a loss function evaluated on data samples by adjusting the NN parameters according to the values of the (chained) gradient. This, coupled with parameters being typically constrained within the $[-1; 1]$ interval, implies that deep networks sequentially multiply small adjustment values, quickly falling to zero. In this way, past information is lost.

Recurrent neural networks (RNN) overcome this shortcoming by explicitly carrying forward relevant information through a hidden state that gates out non-relevant information. This mechanism is reinforced in "stateful" LSTMs, a subset of RNNs.[12] The flip side of this feature is the requirement of long series for training, so much so that the roughly 250 observations present in our quarterly series are barely sufficient.[13] In informing this section we mainly refer to Almosova and Andresen (2019), who first applied stateless RNN-LSTM to inflation forecasting. They show that these models can outperform most traditional forecasting tools and thus are interesting devices to study dynamic properties.

Indeed, they find that these models outperform common forecasting tools at most horizons, prevailing decidedly after the two years horizon. In their investigation, they use

[12] A more detailed and formal introduction to stateful RNN-LSTM is presented in the Appendix.

[13] Properly estimated statistical learning models require about 10^7 data points to train on. With macroeconomic time series, we hardly work with series longer than 300 quarters. Quarterly series are more common in macro applications than weekly and monthly data. The latter are often not available or highly seasonal.

monthly raw data to let the model pick up spontaneously any non-linearities in the data – such as seasonality.

Our approach for this application consists of two steps of increasing granularity. First off, we simply feed the whole sample to the LSTM, let it learn freely and then produce a sufficient number of forecasts to compute the usual statistics on inflation persistence. These forecasts will depend on whatever the LSTM learned from the sequence and will allow for a synthetic extension of the sample size. The output of such trained networks provides insights on likely paths for future inflation and its inertia.

However, to assess the *dynamic change* of persistence we need to train the model on sub-samples of the data. Two options lend themselves to the task: we first split by decades the time series and repeat the analysis just outlined; secondly, we let the LSTM train on a rolling window. This latter will output predictions that can be used to compute persistence and its change over time, much in the spirit of our previous exercises. In the same vein as the Bayesian method, we train the network on a fraction of the data and simulate model-consistent future paths for inflation. These provide additional data points to measure variations in persistence at any given point in time over our sample.

5.1 LSTM forecasting

To produce forecasts for our analyses it is important to decide whether to use a direct or indirect approach to forecasting. The latter consists in feeding the model with data up to time t and subsequently with its own previous forecasts, therefore iterating on data and forecast values. Direct forecasting, on the other hand, use specifically designed models to produce forecasts at a given horizon. Marcellino, Stock, and Watson (2006) compares these two approaches to time series forecasting and find that for linear specifications iterated forecasts perform better than direct ones, and improve with longer forecasting horizons. The case of LSTM differs from the framework of Marcellino, Stock, and Watson (2006) as these models are not strictly linear.

To convey this idea, consider the following simplification. LSTM network links past information to present observation through an arbitrary function F:

$$\pi_t = F\left(\pi_{t-1}, \ldots, \pi_{t-p}; W\right) + \varepsilon_t \tag{10}$$

with p being the lags, W collecting network's parameters, and ε representing an arbitrary error, not necessarily Gaussian nor iid. Then, when the model is trained and $\hat{W}$ is optimised, the model boils down to a possibly non-linear function $\hat{F}$, which can be used to produce forecasts. Naturally, the one period ahead forecast, conditional on time t, reads

$$\hat{\pi}_{t+1|t} = E_t\left[\hat{F}\left(\pi_t; \hat{W}\right)\right] \tag{11}$$

Iterating forward, then, equates to

$$\begin{aligned}
\hat{\pi}_{t+2|t} =& E_t\left[\hat{F}\left(\pi_{t+1}; \hat{W}\right)\right] \\
=& E_t\left[\hat{F}\left(\hat{F}\left(\pi_t; \hat{W}\right); \hat{W}\right)\right] \\
=& E_t\left[\hat{F}\left(\hat{F}\left(F\left(\pi_{t-1}, \ldots, \pi_{t-p}; W\right) + \varepsilon_t; \hat{W}\right); \hat{W}\right)\right].
\end{aligned} \tag{12}$$

Potential non-linearities in F and $\hat{F}$ prevents from taking out ε from the expectations operator directly, but rather calls for computationally intensive integration. Furthermore, no assumption is cast upon the distribution of errors, which in turn accrue over the iterations. A more appropriate approach would consist in fitting one model for each forecast horizon, $\hat{F}^{(t+1)}, \ldots, \hat{F}^{(t+h)}$, based only on information at time t. Such solution is more in the spirit of direct forecasts. While more appropriate, this avenue is computationally demanding, thus we simply assume ε to have mean zero and iterate on previous forecasts, as in Almosova and Andresen (2019).

5.2 LSTM setup

When setting up a LSTM for training, the researcher needs to define its structure, the nodes, and a loss function to evaluate the fit. We study models with one and two layers, and varying numbers of nodes per layer. Satisfactory results can be obtained by a one-layer LSTM with about 75 nodes. Almosova and Andresen (2019) find that the best performance in terms of forecast RSME is produced with 100 nodes.[14] Our preferred loss function is the mean squared error (mse) loss, computed comparing at each step the discrepancies between true and predicted values generated by the network and then used to guide further adjustments in the network's parameters. The choice of such loss function is useful to make direct contact with standard econometric tools, but similar results can be achieved with other compatible loss functions, like mean absolute error (mae).

Weights and biases of the network are optimised to minimize such loss function via the ADAM optimizer (Kingma and Ba, 2014), which is now standard in the field of machine learning (Schmidt, Schneider, and Hennig, 2021). LSTMs feature large numbers of parameters to optimise, usually in the order of thousands if not tens of thousands, and are updated at every iteration. The optimising algorithm explores such highly dimensional parametric space following the gradient of the loss function for as many epochs as the researcher decides to train.[15] This implies that a neural network can be presented several times with the same batch of data and incur in overfitting on the training set with poor out-of-sample performances. We tackle the risk using the early-stopping criterion to govern the adaptive stopping of the optimization. The network is thus shown a 90% subset of the training sample, it is fit with such subsample only. The iterations stop when the loss does not decrease for a given number of iterations on the 10% that was left out for validation. This criterion ensures the generalization of the resulting network. To further improve the generalisation of the results, we impose $L2$ regularisation on the network parameters, nudging weights towards zero in the vein of a Ridge regression.

5.3 Full sample: forecasting inflation and its persistence

We present here the results of a model with one and two layers, 1000 nodes per layer, trained on the full sample of each series. These networks are then used to forecast the inflation rates for the following ten years.Each model is trained on a variety of different

[14] Although it is not clearly stated, one can infer from the parameters count in footnote 8 that such model features a single layer.

[15] Other tuning parameters depend largely on the algorithm of choice: with ADAM, we rely on the default values for the perturbation of the first two moments of the stochastic gradient, namely $\beta_1 = .9$, $\beta_2 = .999$. We also impose $l2$ (Ridge) regularisation on the deeper parameters of the network.

periods, with varying volatility, trend, cyclicity, and monetary policy regimes. This exercise is interesting since LSTMs are geared to capture at the same time short period swings and dependencies that unfold on longer horizons. Extending the sample with predictions from these trained LSMTs allows for a first assessment of the feature learnt and also to propose possible future realisations for inflation inertia going forward.

This first set of results builds on the extension of the sample via forecasts produced by networks trained on the full sample. Comparing models with one or two layers, both with 1000 nodes per layer, the latter does not seem to take advantage of the deeper structure, although forecasts are qualitatively closer to past realisations and smoother overall. It is reasonably due to a shortage of data points: deeper networks, despite the parameters regularization imposed on them, navigate a much more highly dimensional parameter space and thus need more variation in the data as well as more observations to devise a minimum in the loss function.

The next sections present the results of our study employing synthetic data generated by a set of LSTMs on diverse subsamples of the data. From a technical point of view, the LSTMs seem to attain a steady-state-like level when used to iterate forward: despite the absence of clearly defined equilibrium linkages and structural shocks, forecasts converge to the sample mean when iterated for sufficiently long horizons.[16]

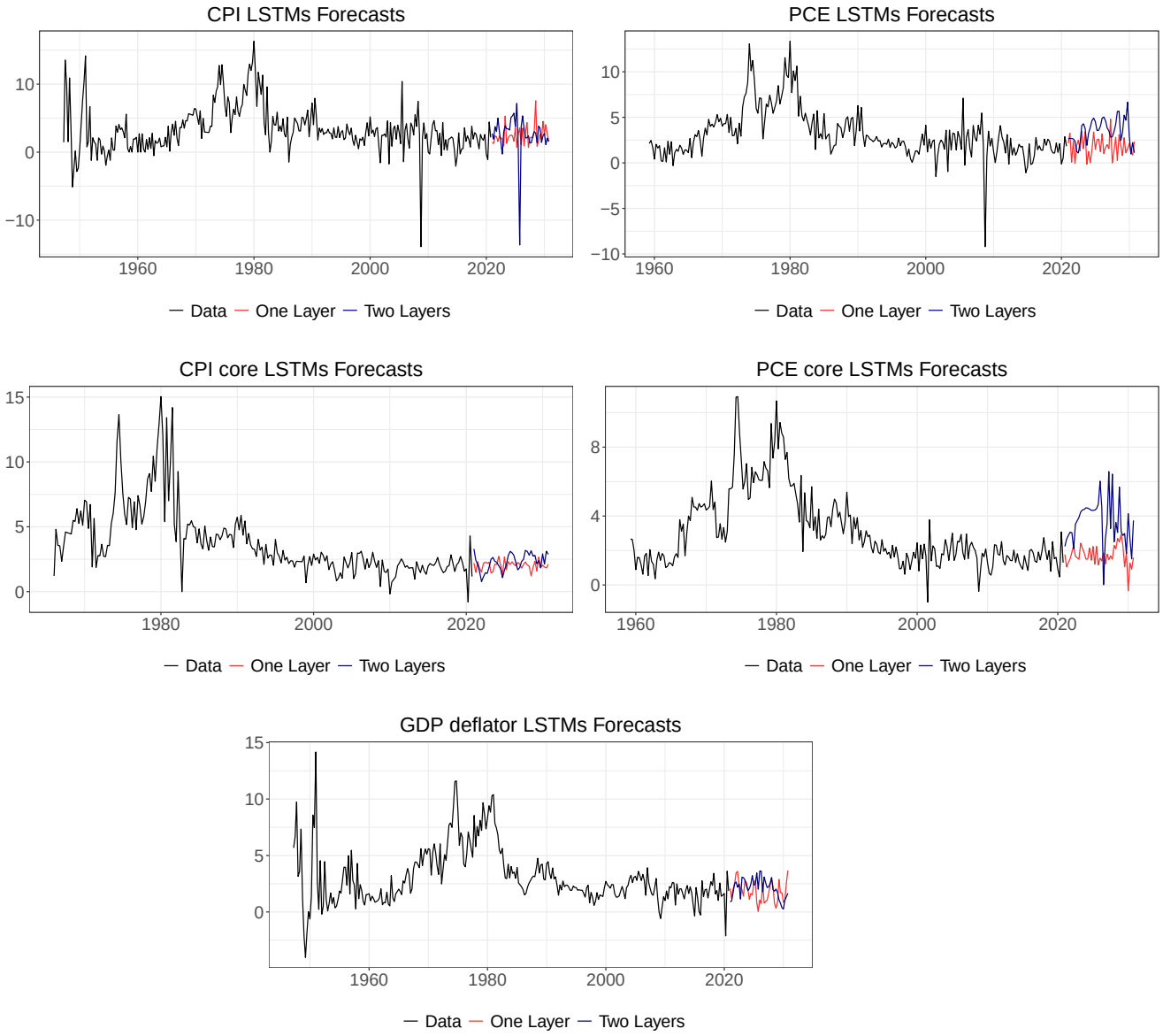
5.4 Regressions on LSTM

The first approach to measure variations in inflation inertia consists in splitting the series into 10-year non-overlapping subsamples (plus optimal lags). Each subsample is used to train a LSTM network that subsequently forecasts on a 40-quarter (ten years) horizon. Then, we fit an autoregressive to elicit the persistence dynamics incorporated into the LSTMs from inflation observations. Panes in the left column of Fig.(8) plot the values from these regressions for the β_1, alongside with confidence intervals. To complete the analysis, right panes in Fig.(8) present the values for the sum of βs for AR(3) models.

Such lag length choice mirrors the optimal lags selected in Section 3 via the BIC minimisation. Broadly, the downward trend for inflation inertia is confirmed, with a sharp drop in all series except the GDP deflator series, which flattens slightly and takes a hump-shaped profile. This holds for both the AR(1) and the AR(3) analyses.[17]

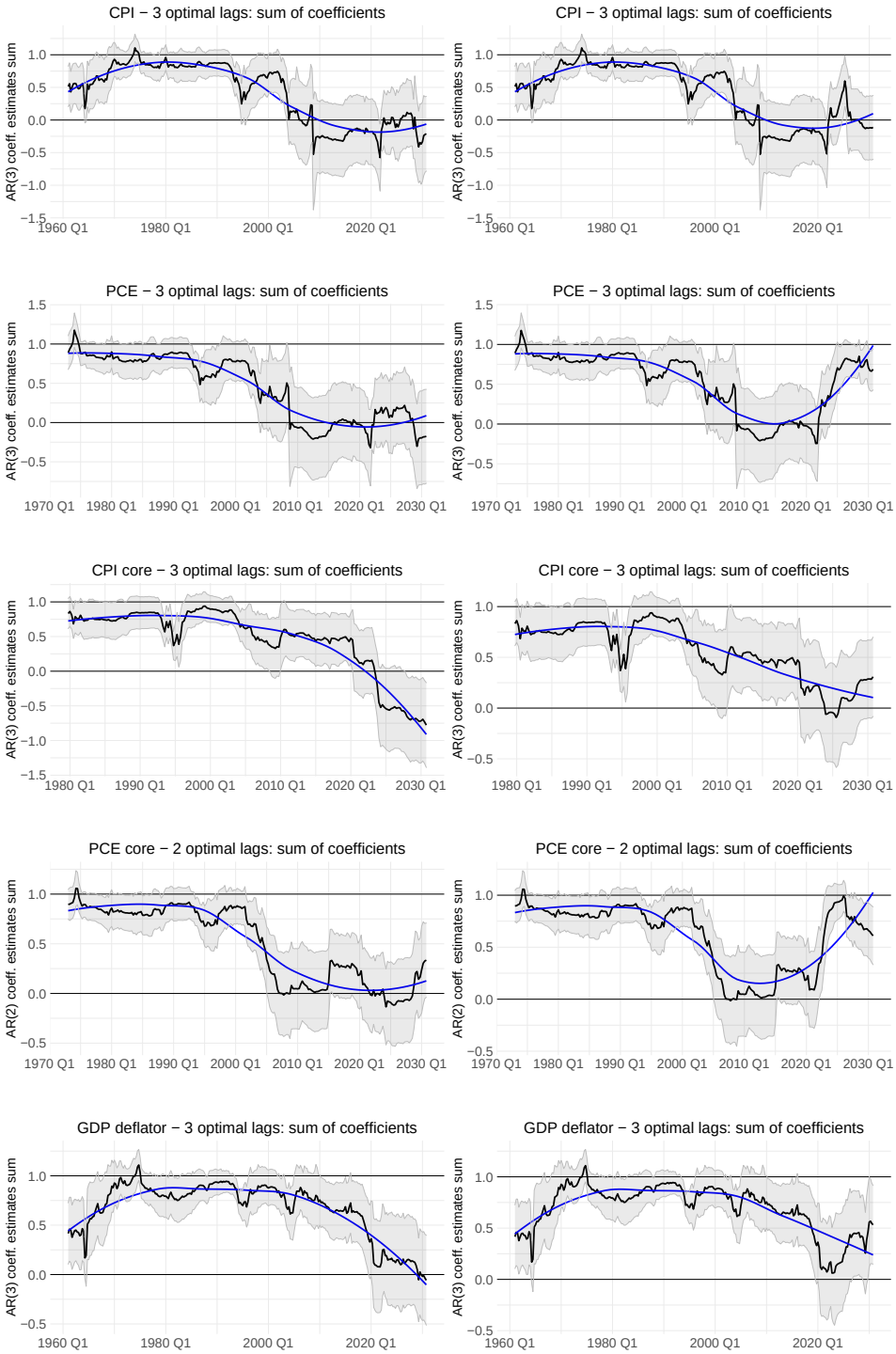
[16] See the presentation of LSTM in the Appendix to clarify the role of the sample mean.
[17] Appendix (F.1) presents detailed OLS regressions results behind such bar plots.

Figure 6: Full Sample Forecasts



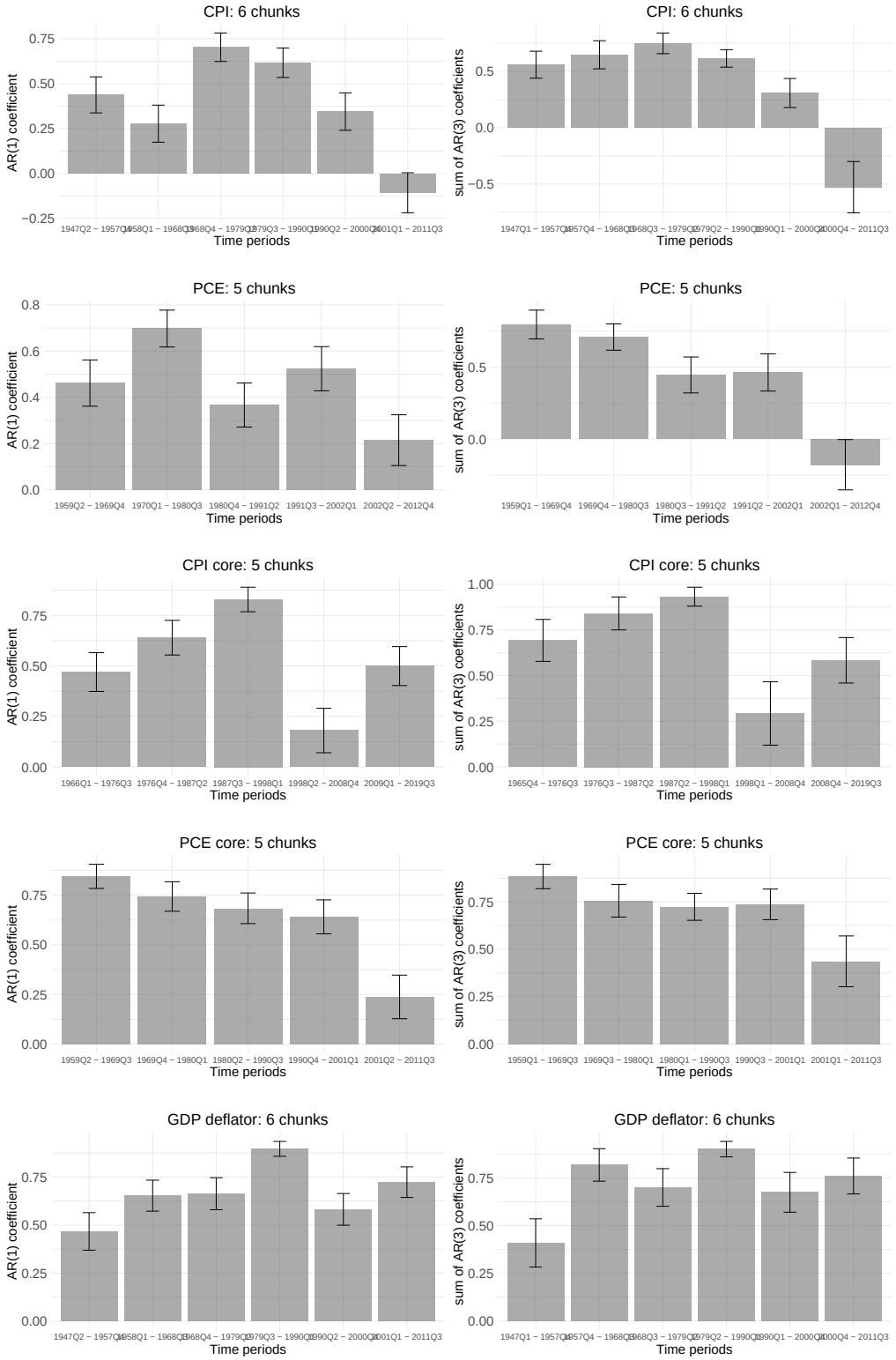
Forecasts from one- and two-layer LSTMs trained on the full sample. Iterative forecasts on 40 quarters.

Figure 7: Full Sample LSTM Predicted Persistence



Left column: AR(k) on data and one-layer LSTM forecast (2020 onward). Right column: AR(k) on data and two-layer LSTM forecast (2020 onward).

Figure 8: LSTM on Decades – Persistence On Data and Forecasts



Left column: $AR\,(1)$ estimates for β_1. Right column: $AR\,(3)$ estimates for sum of β_i. Subsamples are non overlapping and encompass 10 years of data, plus appropriate lags. 'Time periods' report the start-end dates for the subsample of the actual data. Forecasts start from the end date of the sample and run on iteratively for the following decade. Each LSTM network has one layer, 500 neurons, MSE loss function, early-stopping, and is trained for 2000 epochs.

5.5 Rolling LSTMs

This Section presents results from a set of networks trained on a rolling subset of data. Each window spans ten years and is used to train a LSTM net; then, indirect forecasts for the next 40 quarters are produced and used to compute persistence statistics.Once the process is over, the window moves one quarter ahead, drops the oldest observations and a new network is trained. In the same vein of the rolling window in Section (3), using subsets of data and brand new networks allows the assessment of dynamic changes in the underlying data generating process. The idea is to track closely different features of inflation that LSTMs detect in the data and replicate in the forecasts. The advantage over distinct samples is that it allows for a visual detection of such changes – smooth or abrupt ones – at the cost of a precise timing of structural breaks. This procedure is done for AR(1) and AR(3) models.

In general, reported persistence is significantly spikier, with large swings in both higher and lower levels of inertia. Partly, this comes from the relatively small set of data points used to train the networks, which in turn pick up and possibly over-represent local features.[18] Nevertheless, these analyses provide interesting insights on inflation persistence itself, and on different properties of the series employed.

At a first pass, left panes in Fig.(9) display a higher persistence in the initial decades of the covered period, while the scenario is more mixed for recent observations. CPI and PCE (headline and core) show a broad decreasing trend, although the estimates for the first lag coefficient β_{1t} seem to rebound slightly upwards, around 2010. However, the rebound typically starts from negative estimates of the first autocorrelation coefficient. Interestingly, estimates for CPI and PCE display stark U-shaped dynamics in the years from the early 90s to early 00s, as is particularly clear for core PCE. The GDP deflator stands out from the rest of the series, since it posts a decidedly downward trajectory for the last available decade. Consistently with our previous analyses, the estimates for the deflator suggest that some other factor came into play prior to international trade pressures or commodities fluctuations.

A concurrent explanation for such dynamics hinges on policy interventions. With a 10-year rolling window, though, estimates formed in the early 90s are based on a subsample starting in the early 80s, when Volcker impulses a steep turn in the inflation processes. Although appealing, this can explain only part of the dynamics in such decade. Volcker intervention brought down inflation to a moderate level, thus zeroing its structural inertia until its level was under 5%. This policy can well explain the slump and subsequent rebound, while it gives no hints on the second and steady fall in persistence toward the end of the 90s. As previously shown, commodities' volatility plays a minor role, as such dynamics is observed in both core PCE and CPI. Similarly, international trade factors might come at the right timing and act as catalysts to processes already in place. International competition, especially from China, affect tradable sectors, which overlap substantially with manufacturing, thus accelerating a process of sectoral reallocation.

The overall trend, though, can be eyeballed with the help of a polynomial smoother, which points in all plots to a generalised decrease in inflation persistence. It must be noted that during the last decade there seems to be a slow, gradual uptake of inertia, compatible

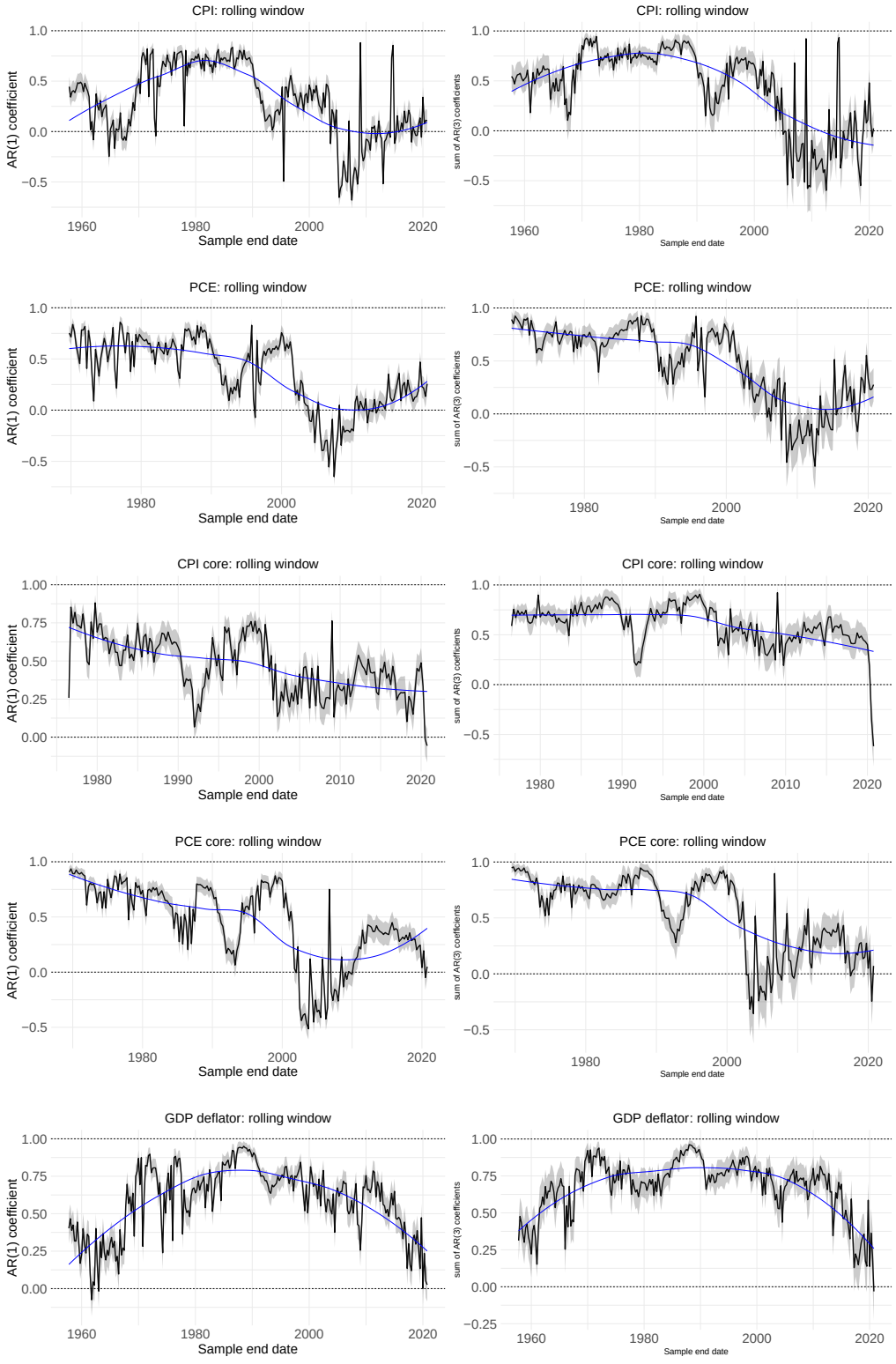
[18] The same exercise can be carried out doubling the window width to 80 quarters – or decreasing significantly the nodes so to avoid over-parametrisation.

with the frequentist analyses presented in Sec.(3). A notable exception is the series for the GDP deflator, which tracks more closely the *composition* of the US economy: it shows all along the sample a concave trend in persistence, and thus substantiates the claim of structural transformation.

Looking at the right panes in Fig.(9), which depict the aforementioned procedure for the sum of the coefficients of an AR(3) process, we can further validate our result of decreasing inflation persistence.

A common pattern is pervasive in all estimates: since early 2000 all estimates present higher volatility in the point estimates and higher uncertainty around these. The end of the Great Moderation period, with fairly stable inflation, appears to set the inflation process on a less predictable ground, with generally large swings in its persistence and basically a quasi-white noise process at times. The root causes for such behaviour are still unclear, but the lively debate on the Phillips Curve, consumers and financiers diverging perception of inflation, and trends at the firm and macro levels do point to some candidate explanations.

Figure 9: Rolling LSTM – Persistence on Data and Forecasts



Left column: plots for AR(1) autocorrelation on a 10-year rolling window augmented with optimal lags. Right column: plots for coefficients sum from an AR(3). Within each window a small LSTMs is trained (2000 epochs, 1 layer, 500 nodes, MSE loss) and then used to iteratively forecast the next 40 quarters (10 years) since last actual observation. Then an autoregressive model is estimated on this extended window, the autocorrelation is stored and plotted as black solid line in correspondence of the last actual data point date, with shaded areas reporting 95% confidence intervals around the point estimate. Blue solid line represent a LOESS polynomial fit to highlight long term trends.

6 Conclusion

Inflation behaviour has been widely investigated in recent years, within Phillips curves and statistical frameworks, yet no conclusive consensus has emerged. Regarding inflation dynamics, it is unclear whether inflation persistence has stabilised (Fuhrer, 2011; Pivetta and Reis, 2007) or declined (Stock and Watson, 2007). Even less established is the debate around the determinants of inflation dynamics (Mavroeidis, Plegborg-Moller, and Stock, 2014). Persistence, or equivalently inertia, is a fundamental property to consider when fiscal or monetary policies are devised and evaluated, as it encapsulates how responsive prices are to interventions.

In this work we revise and extend previous and analyses of inflation persistence for the US macroeconomy. We extend the set of inflation measures to include GDP deflator, which tracks closely the US economy's structure, Consumer Price Index, and Personal Consumption Expenditure index – core and headline. This extension allows the isolation of a number of potential confounding factors: international trade effects from imported goods and services, volatility effects from energy and food items, evolving structure of the US industrial composition. After using autoregressive and Bayesian tools, we extend the methodological toolkit drawing from the deep neural networks field. We adapt Long-Short Term Memory (LSTM) recursive neural networks to leverage their predictive performance and flexible management of nonlinearities – time-varying lag structure, seasonality, short-lived cyclical fluctuations, and long term trends. This class of deep, recursive neural networks already outperforms classic forecasting tools for time series (Almosova and Andresen, 2019; Verstyuk, 2020). We train several of these nets over the full, split, and rolling samples and leverage their flexibility to extend observations and thus study US inflation persistence since WWII.

We show that inflation persistence substantially decreased since the mid-'90s. This pattern holds irrespective of the revised measure of inflation we use. The timing suggests that it is not fully explained by international trade or commodities prices: Persistence peaks around the second half of the '90s, before China's WTO accession and before the increase in energy and food volatility. This is confirmed when we look at the data in a more flexible way. We find evidence that inflation series currently behave similarly to a white noise process, showing a decreasing connection with past values. We also report on the increased statistical uncertainty associated with headline series: volatility in commodities prices further decrease the predictability of overall inflation. Conversely, the GDP deflator displays a smooth, hump-shaped decrease in persistence, suggestive of longer trends in the economy.

In light of the encouraging performance of LSTM models applied to time series econometrics, a promising avenue of research is the further extension of such tools to a wider array of methods. Furthermore, we leave for future research the investigation on the root causes for decreasing inertia.

References

Almosova, Anna and Niek Andresen (2019). "Nonlinear inflation forecasting with recurrent neural networks". In: *WP*.

Athey, Susan (2018). "The Impact of Machine Learning on Economics". In: *The Economics of Artificial Intelligence: An Agenda*. NBER Chapters. National Bureau of Economic Research, Inc, pp. 507–547.

Athey, Susan and Guido W. Imbens (2015). *Machine Learning for Estimating Heterogeneous Causal Effects*. Research Papers 3350. Stanford University, Graduate School of Business.

— (2019). "Machine Learning Methods That Economists Should Know About". In: *Annual Review of Economics* 11.1, pp. 685–725.

Auer, Raphael, Claudio Borio, and Andrew Filardo (2017). *The globalisation of inflation: the growing importance of global value chains*. BIS Working Papers 602. Bank for International Settlements.

Autor, David H., David Dorn, and Gordon H. Hanson (2016). "The China Shock: Learning from Labor-Market Adjustment to Large Changes in Trade". In: *Annual Review of Economics* 8.1, pp. 205–240.

Bajari, Patrick et al. (2015). "Machine Learning Methods for Demand Estimation". In: *American Economic Review* 105.5, pp. 481–85.

Benati, Luca and Paolo Surico (2008). "Evolving U.S. Monetary Policy and The Decline of Inflation Predictability". In: *Journal of the European Economic Association* 6.2-3, pp. 634–646.

Bianchi, Francesco (2013). "Regime switches, agents' beliefs, and post-WWII US macroeconomic dynamics". In: *Review of Economic Studies*.

Bianchi, Francesco and Andrea Civelli (2015). "Globalization and Inflation: Evidence from a Time Varying VAR". In: *Review of Economic Dynamics* 18.2, pp. 406–433.

Chakraborty, Chiranjit and Andreas Joseph (2017). *Machine learning at central banks*. Bank of England working papers 674. Bank of England.

Ciccarelli, Matteo and Chiara Osbat (2017). "Low inflation in the euro area: Causes and consequences". In: *ECB Occasional Paper 181*.

Cogley, Timothy, Giorgio E. Primiceri, and Thomas J. Sargent (2008). "Inflation-gap persistence in the US". In: *American Economic Journal: Macroeconomics*.

Cogley, Timothy and Thomas J. Sargent (2002). "Evolving post-World War II US inflation dynamics". In: *NBER Macroeconomics Annual 2001*. Ed. by Mark Gertler and Kennet Rogoff.

— (2005). "Drift and volatilities: monetary policies ad outcomes in the post World War II US". In: *Review of Economic Dynamics*.

Cogley, Timothy and Argia Sbordone (2009). "Trend inflation, indexation, and inflation persistence in New Keynesian Phillips curve". In: *American Economic Review*.

Coibion, Olivier and Yuriy Gorodnichenko (2015). "Is the Phillips curve alive and well after all? Inflation expectations and the missing disinflation". In: *American Economic Journal – Macroeconomics*.

Coibion, Olivier, Yuriy Gorodnichenko, and Rupal Kamdar (Dec. 2018). "The Formation of Expectations, Inflation, and the Phillips Curve". In: *Journal of Economic Literature* 56.4, pp. 1447–91.

Coibion, Olivier, Yuriy Gorodnichenko, and Tiziano Ropele (2019). "Inflation expectations and firm decisions: new causal evidence". In: *WP*.

Fernandez-Villaverde, Jesus and Pablo Guerron-Quintana (2020). "Estimating DSGE models: recent advances and future challenges". In: *NBER WP w27715*.

Fernández-Villaverde, Jesús, Pablo Guerrón-Quintana, and Juan F. Rubio-Ramírez (Apr. 2010). *Fortune or Virtue: Time-Variant Volatilities Versus Parameter Drifting in U.S. Data*. NBER Working Papers 15928. National Bureau of Economic Research, Inc.

— (2015). "Estimating dynamic equilibrium models with stochastic volatility". In: *Journal of Econometrics* 185.1, pp. 216–229.

Fernández-Villaverde, Jesús, Samuel Hurtado, and Galo Nuño (2020). *Financial Frictions and the Wealth Distribution*. CESifo Working Paper Series 8482. CESifo.

Fuhrer, Jeffrey (2011). "Inflation Persistence". In: *Handbook of Monetary Economics*.

Giannone, Domenico, Michele Lenza, and Giorgio E. Primiceri (2018). *Economic predictions with big data: the illusion of sparsity*. Staff Reports 847. Federal Reserve Bank of New York.

Goulet Coulombe, Philippe et al. (2019). *How is Machine Learning Useful for Macroeconomic Forecasting?* CIRANO Working Papers 2019s-22. CIRANO.

Greff, Klaus et al. (2015). "LSTM: A Search Space Odyssey". In: *CoRR* abs/1503.04069.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu (Feb. 2020). "Empirical Asset Pricing via Machine Learning". In: *The Review of Financial Studies* 33.5, pp. 2223–2273.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning*.

Jarociński, Marek and Elena Bobeica (2017). *Missing disinflation and missing inflation: the puzzles that aren't*. Working Paper Series 2000. European Central Bank.

Jozefowicz, Rafal, Wojciech Zaremba, and Ilya Sutskever (2015). "An Empirical Exploration of Recurrent Network Architectures". In: ed. by Francis Bach and David Blei. Vol. 37, pp. 2342–2350.

Jung, Jin-Kyu, Manasa Patnam, and Anna Ter-Martirosyan (2018). *An Algorithmic Crystal Ball: Forecasts-based on Machine Learning*. IMF Working Papers. International Monetary Fund.

Karpathy, Andrej (2015). *The unreasonable effectiveness of recurrent neural networks*. URL: %7Bhttp://karpathy.github.io/2015/05/21/rnn-effectiveness/%7D.

Kidger, Patrick and Terry J. Lyons (2019). "Universal Approximation with Deep Narrow Networks". In: *CoRR* abs/1905.08539.

Kingma, Diederik P. and Jimmy Ba (2014). *Adam: A Method for Stochastic Optimization*.

Kock, Anders Bredahl and Timo Teräsvirta (2016). "Forecasting macroeconomic variables using neural network models and three automated model selection techniques". In: *Econometric Reviews* 35.8-10, pp. 1753–1779.

Korobilis, Dimitris (2018). *Machine Learning Macroeconometrics: A Primer*. Working Paper series 18-30. Rimini Centre for Economic Analysis.

Kurozumi, Takushi and Willem Van Zandweghe (Aug. 2019). *A Theory of Intrinsic Inflation Persistence*. Working Papers 201916. Federal Reserve Bank of Cleveland.

Leshno, Moshe et al. (1993). "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function". In: *Neural Networks* 6.6, pp. 861–867.

Lhuissier, Stephane (2018). "The regime-switching volatility of Euro Area business cycles". In: *Macroeconomic Dynamics*.

Makridakis, Spyros, Evangelos Spiliotis, and Vassilios Assimakopoulos (2018). "Statistical and Machine Learning forecasting methods: Concerns and ways forward". In: *PLOS ONE* 13.3, pp. 1–26.

Maliar, Lilia, Serguei Maliar, and Pablo Winant (2019). "Will artificial intelligence replace computational economists any time soon?" In: *CEPR discussion paper DP14024*.

Mankiw, N. Gregory, Ricardo Reis, and Justin Wolfers (2003). *Disagreement about Inflation Expectations*. Working Paper 9796. National Bureau of Economic Research.

Marcellino, Massimiliano, James H. Stock, and Mark W. Watson (2006). "A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series". In: *Journal of Econometrics*.

Mavroeidis, Sophocles, Mikkel Plegborg-Moller, and James H. Stock (2014). "Empirical evidence on inflation expectations in the New Keynesian Phillips Curve". In: *Journal of Economic Literature*.

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

Medeiros, Marcelo C. et al. (2019). "Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods". In: *Journal of Business & Economic Statistics*, pp. 1–22.

Miles, David et al. (2017). *And yet it moves: Inflation and the Great Recession*.

Mullainathan, Sendhil and Jann Spiess (2017). "Machine Learning: An Applied Econometric Approach". In: *Journal of Economic Perspectives* 31.2, pp. 87–106.

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

Pivetta, Frederic and Ricardo Reis (2007). "The persistence of inflation in the United States". In: *Journal of Economic Dynamics and Control*.

Rackauckas, Christopher et al. (2020). "Universal Differential Equations for Scientific Machine Learning". In: *arXiv preprint arXiv:2001.04385*.

Ruder, Sebastian (2016). "An overview of gradient descent optimization algorithms". In: *CoRR* abs/1609.04747.

Schmidt, Robin M., Frank Schneider, and Philipp Hennig (2021). *Descending through a Crowded Valley – Benchmarking Deep Learning Optimizers*.

Stock, James H. and Mark W. Watson (2007). "Why Has U.S. Inflation Become Harder to Forecast?" In: *Journal of Money, Credit and Banking* 39.s1, pp. 3–33.

Tabuada, Paulo and Bahman Gharesifard (2020). *Universal Approximation Power of Deep Residual Neural Networks via Nonlinear Control Theory*.

Taylor, John B. (2014). *After unconventional monetary policy. Testimony before the Joint Economic Committee of Congress at the Hearing on "Unwinding quantitative easing: how the Fed should promote stable prices, economic growth and job creation"*.

Trehan, Bharat (2015). "Survey Measures of Expected Inflation and the Inflation Process". In: *Journal of Money, Credit and Banking* 47.1, pp. 207–222.

Varian, Hal R. (2014). "Big Data: New Tricks for Econometrics". In: *Journal of Economic Perspectives* 28.2, pp. 3–28.

Verstyuk, Sergiy (2020). "Modeling multivariate time series in economics: from auto-regressions to recurrent neural networks". In: *WP Mimeo*.

Walsh, Carl E. (2003). *Monetary Theory and Policy*.

Woodford, Micheal (2003). *Interest and Prices: Foundations of a Theory of Monetary Policy*.

Online Appendix

A A primer on the RNN-LSTM framework

In the last section we introduce RNN-LSTM models and use them to gain a deeper understanding of inflation dynamics. This section gives a succinct presentation of these deep-learning tools, presents the optimisation algorithm employed and shortly discusses the hyperparameters tuning assumptions. These paragraphs draw from Almosova and Andresen (2019), Greff et al. (2015), Jozefowicz, Zaremba, and Sutskever (2015), Karpathy (2015), and Verstyuk (2020). For a thorough, formal presentation of Neural Networks within the statistical learning framework, refer to Hastie, Tibshirani, and Friedman (2009).

From statistical learning to long-short term memory models Artificial Neural Networks (ANNs) are systems combining nodes, layers, relationships, biases, and activation functions. By design, they mimic in their structure the human brain with interconnected neurons (nodes) and connections thereof (layers). Recurrent Neural Networks (RNN) are a subclass of ANN, especially devised to deal with *sequences*. Within the RNN class are the Long-Short Term Memory models, which address more complex sequential structures involving varying time dependency and indexed observations.

In a nutshell, each layer is populated with nodes (or neurons) that form a linear combination of the layer's input. Therefore a sequence of layers boils down to a sequence of linear combinations of the original input. Such combination is flexible enough to approximate any nonlinear function with an arbitrary degree of precision.

The general structure of a neural network with M layers and N_m, $m \in 1, \dots, M$, nodes per layer is the following, unrolling the hidden layers

$$\begin{aligned} \mathbf{h_1} &= \mathbf{g_1}\left(\mathbf{b_1} + \mathbf{W_1 x}\right) \\ \mathbf{h_2} &= \mathbf{g_2}\left(\mathbf{b_2} + \mathbf{W_2 h_1}\right) \\ \mathbf{h_3} &= \mathbf{g_3}\left(\mathbf{b_3} + \mathbf{W_3 h_2}\right) \\ &\dots \\ \mathbf{h_M} &= \mathbf{g_M}\left(\mathbf{b_M} + \mathbf{W_M h_{M-1}}\right) \\ \mathbf{\hat{y}} &= \mathbf{g_{M+1}}\left(\mathbf{b_{M+1}} + \mathbf{W_{M+1} h_M}\right) \end{aligned} \tag{13}$$

where $\mathbf{x} \in \mathbb{R}^K$ is a K-dimensional vector of inputs rescaled to have $E\mathbf{x} = \mathbf{0}$ and $V\mathbf{x} = \mathbf{1}$, while $\mathbf{g}_m : \mathbb{R}^{N_m} \mapsto \mathbb{R}^{N_m}$ are activation function mapping layers output from one layer to the following downstream.[19] $\mathbf{b_m} \in \mathbb{R}^{N_m}$ and $\mathbf{W_m} \in \mathbb{R}^{N_m \times (N_m - 1)}$ are the biases and weights of the m-th layer. These last objects will be the target for the optimisation and will be tuned as to minimise a loss function. Finally, $\mathbf{\hat{y}}$ is the predicted vector of the network, to be compared and evaluated against the observed one, $\mathbf{y}$.

We can roll up the network in a more succinct way by function composition

$$\mathbf{\hat{y}} = \mathbf{g_M} \circ \dots \circ \mathbf{g_1}\left(\mathbf{x}\right) \tag{14}$$

[19] For $m \neq 1$, in such case $\mathbf{g}_1 : \mathbb{R}^K \mapsto \mathbb{R}^{N_2}$.

The general idea is that the input $\mathbf{x}$ is passed sequentially through the layers as a conveyor belt and it is transformed – possibly in nonlinear ways – by the mediation of weights, biases and activation functions that are encapsulated in each layer $\mathbf{g_m}$ to finally predict a value $\hat{\mathbf{y}}$. Tuning $\mathbf{b_m}$ and $\mathbf{W_m}$ will eventually improve the predictions of the network.

While the above structure describes a generic RNN, the inner workings of each $\mathbf{g_m}$ make LSTMs apt to dealing with complex time series. In particular, each LSTM layer is composed of gates, states, memory, and output cells. In short, the m-th layer, inherits state and output from the previous one and contains

$$\begin{aligned}
\mathbf{i}_t &= \sigma\left(\mathbf{b}_i + \mathbf{W}_i \mathbf{h}_{m-1,t} + \mathbf{U}_i \mathbf{h}_{t-1}\right) \\
\mathbf{f}_t &= \sigma\left(\mathbf{b}_f + \mathbf{W}_f \mathbf{h}_{m-1,t} + \mathbf{U}_f \mathbf{h}_{t-1}\right) \\
\mathbf{o}_t &= \sigma\left(\mathbf{b}_o + \mathbf{W}_o \mathbf{h}_{m-1,t} + \mathbf{U}_o \mathbf{h}_{t-1}\right) \\
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tanh\left(\mathbf{b}_c + \mathbf{W}_c \mathbf{h}_{m-1,t} + \mathbf{U}_c \mathbf{h}_{t-1}\right) \\
\mathbf{h}_t &= \mathbf{o}_t \odot \tanh\left(\mathbf{c}_t\right)
\end{aligned} \tag{15}$$

where $\mathbf{i}$ is the input cell fed with the output of the previous layer $\mathbf{h}_{m-1,t}$ and past observations $\mathbf{h}_{t-1}$. $\sigma(\cdot)$ is a sigmoid function that squashes its inputs into the $[-1, 1]$ interval to avoid exploding behaviour. Then, $\mathbf{f}_t$ is the forget gate, deciding what part of information retain from the past. $\mathbf{o}_t$ is the output gate, which decides upon the final output of the cell. While these three cells produce activation vectors that signal what to retain, to forget, and to pass on, the last two cells are more involved. $\mathbf{c}_t$ is the cell hidden state, it is updated upon the previous cell state and the new, retained information: this results from simultaneously forgetting something from the previous state: $\mathbf{f}_t \odot \mathbf{c}_{t-1}$; and updating from the current input: $\mathbf{i}_t \odot \tanh(\cdot)$. Finally, the last cell combines all of the above in the final output $\mathbf{h}_t$. In this formulation $\odot$ stands for the element-wise multiplication, while, similarly to $\sigma(\cdot)$, tanh is used to regularise values in a given space. In all of the above, $\mathbf{U_s}, \mathbf{W_s}, \mathbf{b_s}$ are weights matrices and biases relative to that particular cell.

The advantage of LSTMs over other infrastructures is to be found in the additive (instead of multiplicative) update of the hidden state $\mathbf{c}_t$, which prevents the issue of vanishing gradient when performing backpropagation.

Lastly, the training of these networks is performed deciding a loss function L that will be minimised by adjusting the chained weights and biases $\mathbf{W}, \mathbf{b}$ so to obtain predictions $\hat{\mathbf{y}}$ as close as possible to the observed values $\mathbf{y}$. Typical loss functions are Mean Squared Error (`mse`) or Mean Absolute Error (`mae`), which are also helpful to make direct contact to the econometrics field. In the `mse` case, thus, the objective is

$$\min_{\mathbf{W},\mathbf{b}} \; L\left(\mathbf{W}, \mathbf{b}; \hat{\mathbf{y}}, \mathbf{y}\right) = \frac{1}{T} \sum_{t=1}^{T} \left[\hat{\mathbf{y}}_\mathbf{t}\left(\mathbf{W}, \mathbf{b}\right) - \mathbf{y}_\mathbf{t}\right]^2 \tag{16}$$

To minimise such loss function, the full chained gradient is computed and weights and biases are adjusted accordingly, while the networks is evaluated on a variable subset of data (ie, batches of contiguous data points). The adjustment via the chained gradient from the final predicted y back to the earliest layers of the network is precisely backpropagation.

Optimiser The choice of the optimising algorithm is paramount in such framework. In light of the wealth of parameters to fine-tune in order to minimise L, the dimensionality of the parameters space easily scales up to orders of millions of dimensions. Therefore, the optimisation must efficiently explore such space and avoid local minima when possible. Nowadays, the ADAM algorithm has proven to be reliable and efficient in these terms (see Ruder (2016) for a thorough overview of several optimisation algorithms). In a nutshell, it is an adaptation of the stochastic gradient descent algorithm, where on top of the gradient directions there are stochastic perturbations. From Kingma and Ba (2014), it boils down to

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1-\beta_1)\, g_t \\ v_t &= \beta_2 v_{t-1} + (1-\beta_2)\, g_t^2 \\ \hat{m}_t &= \frac{m_t}{1-\beta_1^t} \\ \hat{v}_t &= \frac{v_t}{1-\beta_2^t} \\ [\mathbf{W}, \mathbf{b}]_{t+1} &= [\mathbf{W}, \mathbf{b}]_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \end{aligned} \tag{17}$$

Where g and g^2 are the mean and uncentered variance of the gradient at the current step, while m and v are their moving averages, and $\hat{m}$, $\hat{v}$ are unbiased estimates of these two moments. Finally, the parameters θ are updated as in the last equation. The parameters $\beta_1, \beta_2, \epsilon, \eta$ are the decay rates, a smoothing term, and the learning rate, which essentially governs the change in the parameters.

Implementing this algorithm involves selecting a number of iterations (in ML jargon, *epochs*) and let the algorithm update the parameters for long enough to explore the minima.

Batches, early stopping, validation A handful of choices completes the setup of our exercise with LSTMs: batch size, early stopping on validation, and regularisation.

Batch size governs the subsamples that are fed to the model *at once*. While during each epoch the model is shown the whole dataset, the researcher can choose to pass smaller chunks of data in order to let the model pick up relevant patterns that are common across batches. To grasp the idea, consider seasonality in monthly data: if every May presents a spike and we only train the model on individual months, it will take a longer time to catch such seasonality than if we train it on batches 12 or 24 months at once. The tradeoff in choosing the batch size, thus, is between convergence speed and learning: if the batch size equals the number of observations, the training is faster but less refined and eventually the model is more exposed to overfitting. If the batch size is very small the training is slow and the model might miss key patterns, but out-of-sample performance might benefit.

A key factor of batches is that they are drawn at random from the whole dataset, much in the spirit of bootstrapping. This stochastic subsampling introduces a source of randomness that helps with the out-of-sample generalisation of the trained model.

Once the batch size is defined,[20] the researcher can either set a number of epochs and let the model train, or set a stopping rule and set an upper bound for the iterations. In this work we adopt the Early Stopping rule on validation data. When the model starts training,

[20] It must be noted that, for technical reasons, the batch size must evenly divide training and test samples.

a subsample of the training data is kept apart and not used for learning. At the end of each epoch, the model performance is evaluated on such validation set via the loss function. The training therefore stops when the loss stops decreasing on the validation data or the number of epochs is reached. This procedure ensures a higher level of generalisation of the network and might save some computational time during training.

Lastly, at the end of each epoch, during the parameters update, we impose $L2$ regularisation on the parameters. LSTMs and RNNs in general present several thousands of parameters, and overfitting is often a real risk. To minimise such threat, we add a penalisation to parameters, in the spirit of ridge regressions. In short, the loss function L is augmented to nudge the optimisation to retain only relevant parameters:

$$\min_{\mathbf{W},\mathbf{b}} L\left(\mathbf{W}, \mathbf{b}; \hat{\mathbf{y}}, \mathbf{y}\right) + \lambda \sum_{i=1}^{P} ||\mathbf{W_i}, \mathbf{b_i}||^2 \tag{18}$$

where we take the square of the $l2$ norm, P is the total number of parameters, and λ governs the penalty relevance. In particular, λ can be interpreted as the penalty given to model complexity. It is typically set between 0 and .1: higher values nudge model's weights to be close (but not exactly equal) to 0. λ helps to balance the trade-off between generalisation to new observations and overfitting the training data.

Data transformation To properly train the LSTMs, it is necessary to prepare the dataset with some transformations. After splitting the full sample into two parts, training and test, data must be rescaled to match zero mean and unitary standard deviation. Importantly, this is done *separately* for training and test subsamples. Once the model is trained and produces forecasts, these are still scaled to be of null mean and unitary standard deviation, hence they must be reconverted to the original data magnitude. Interestingly, LSTMs seem to produce forecasts similar to traditional IRFs, in such there is a mean-reversing force when the forecast horizon is long enough.

Taking stock Finally, after reviewing these components of the LSTMs setup, we can sum up the values for the hyperparameters used in our exercise in Tab.(2).

Table 2: Hyperparameters

Instance	Hyperparameter	Value
Full sample, 1 layer	nodes	1000
	epochs	5000
	batch size	highest prime factor of sample size
	lags	15
	early stopping	yes
	trainable parameters	$\sim$ 4mln
Full sample, 2 layers	nodes per layer	750
	epochs	5000
	batch size	highest prime factor of sample size
	lags	15
	early stopping	yes
	trainable parameters	$\sim$ 7mln
10y subsamples, 1 layer	nodes	500
	epochs	2000
	batch size	highest prime factor of sample size
	lags	10
	early stopping	no
	trainable parameters	$\sim$ 1mln
10y rolling window, 1 layer	starting sample	optimal lags + 10 years
	nodes	500
	epochs	2000
	batch size	highest prime factor of sample size
	lags	10
	early stopping	no
	trainable parameters	$\sim$ 1mln
Common across setups		
ADAM optimiser	β_1	.9 (def)
	β_2	.999 (def)
	η	.001 (def)
	ϵ	$1e-7$ (def)
early stopping	validation share	last 10% of batch
	tolerance	$1e-5$
	patience	20% of epochs
forecast horizon	quarters ahead	40

B Year-on-year series

When computing the inflation rate, one can either choose to compute the annualised rate of change between two contiguous quarters (quarter-on-quarter, *qoq*), or compute the change from the corresponding quarter of the previous year (year-on-year, *yoy*). While in principle these methods yield broadly the same inflation rates, *yoy* series display a rather different, higher level of persistence, as found when computing the optimal lags via the BIC minimisation. Formally, the two rates result from

$$\pi_t^{qoq} = 400 \times \ln\left(P_t / P_{t-1}\right) \qquad \pi_t^{yoy} = 100 \times \ln\left(P_t / P_{t-4}\right)$$

To compare the discrepancies in persistence, Tab.(3) reports the values for optimal lags in the two sets of series.

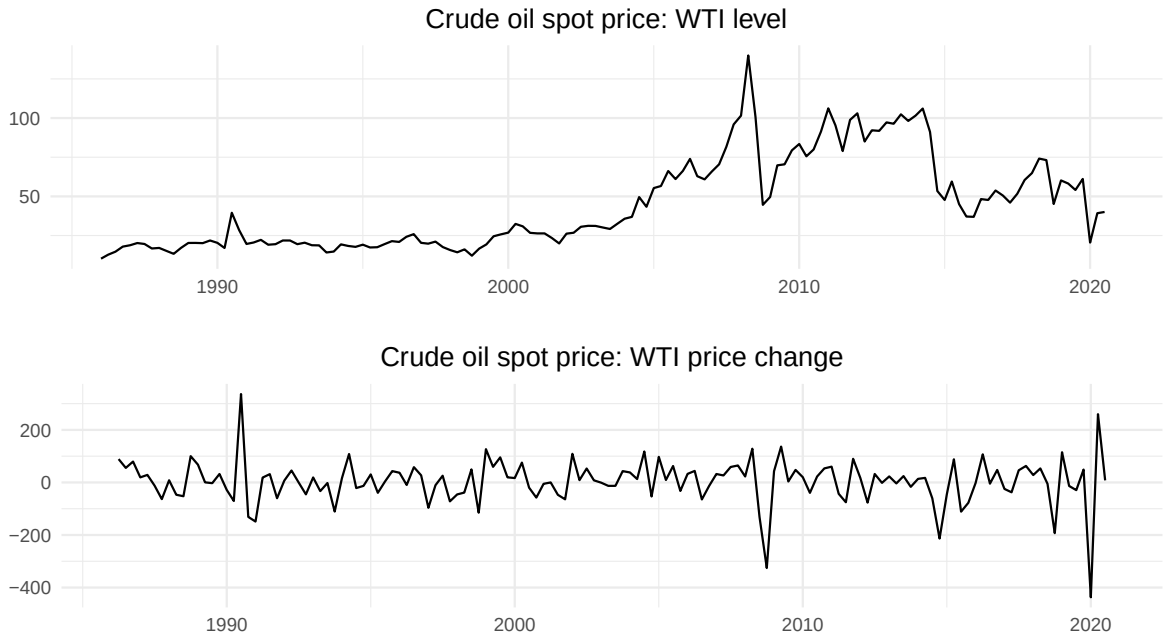
Table 3: Optimal lags from BIC minimisation

	GDP Defl.	CPI headline	CPI core	PCE headline	PCE core
k^*_{qoq}	3	3	3	3	2
k^*_{yoy}	9	9	18	9	18

C Isolating oil and commodities' inflation from headline

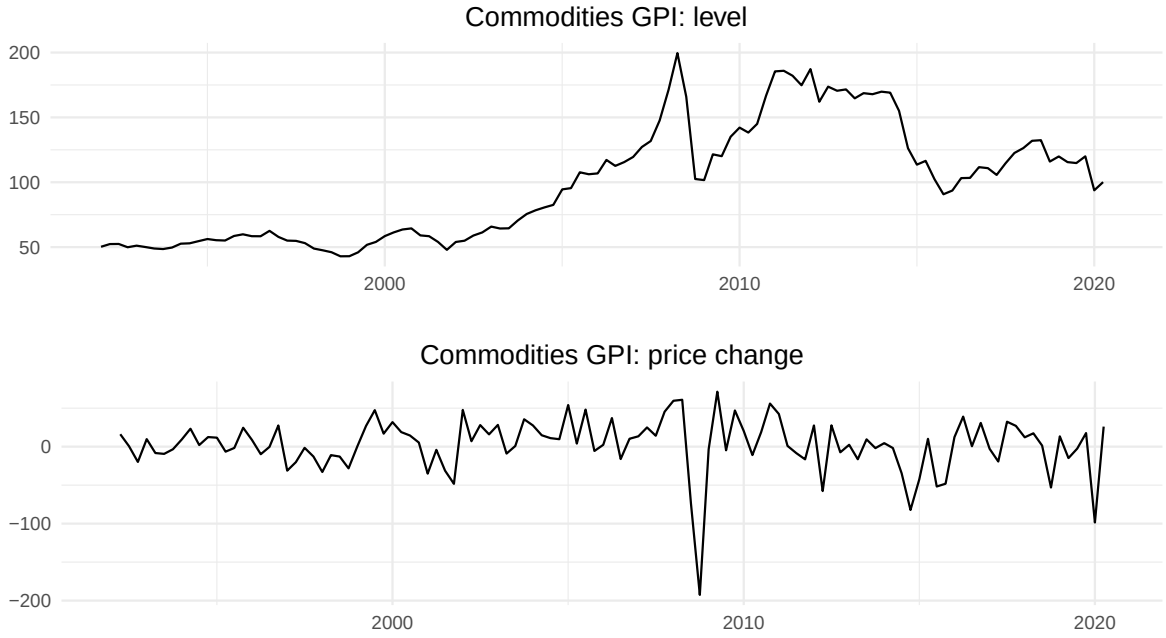
Figs.(10,11) present data on oil and commodities prices, along with their variations. This allows the appreciation of the main differences between headline and core series for CPI and PCE, with core series excluding the items included in these plots.

Figure 10: Oil Prices



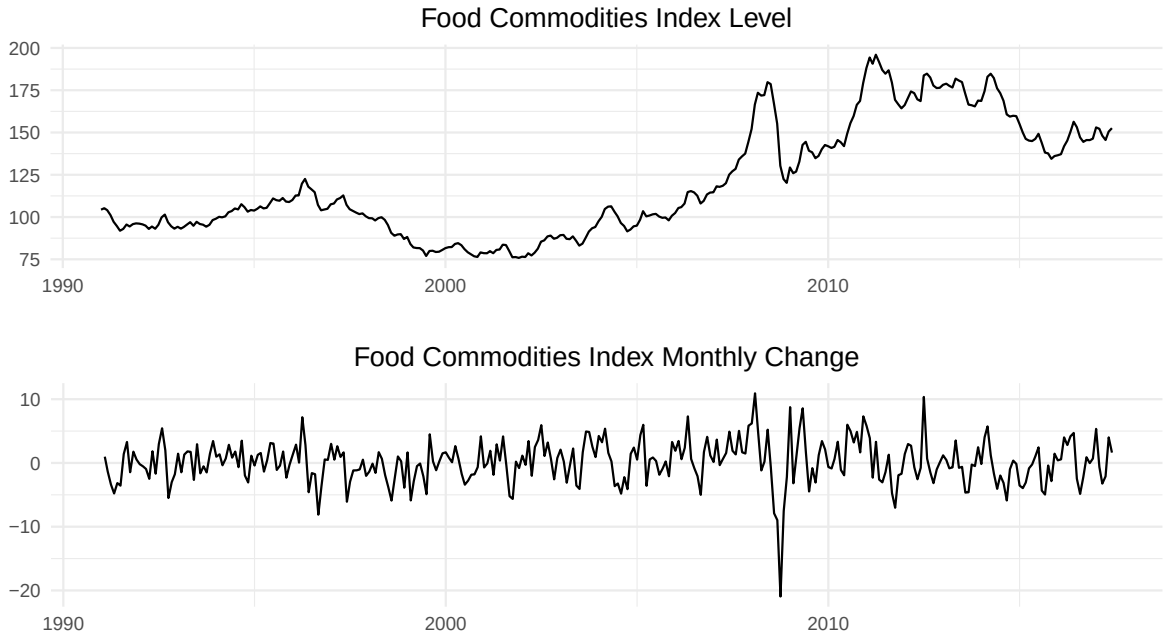
West Texas Intermediate spot price, level (top), and qoq annualized percent change (bottom). Source: FRED, St. Louis Fed.

Figure 11: Commodities Prices



Global Price Index for commodities, level (top), and qoq annualized percent change (bottom). This series includes prices for oil, gas, metals, grains, among others. Source: FRED, St. Louis Fed.

Figure 12: Food Commodities Prices

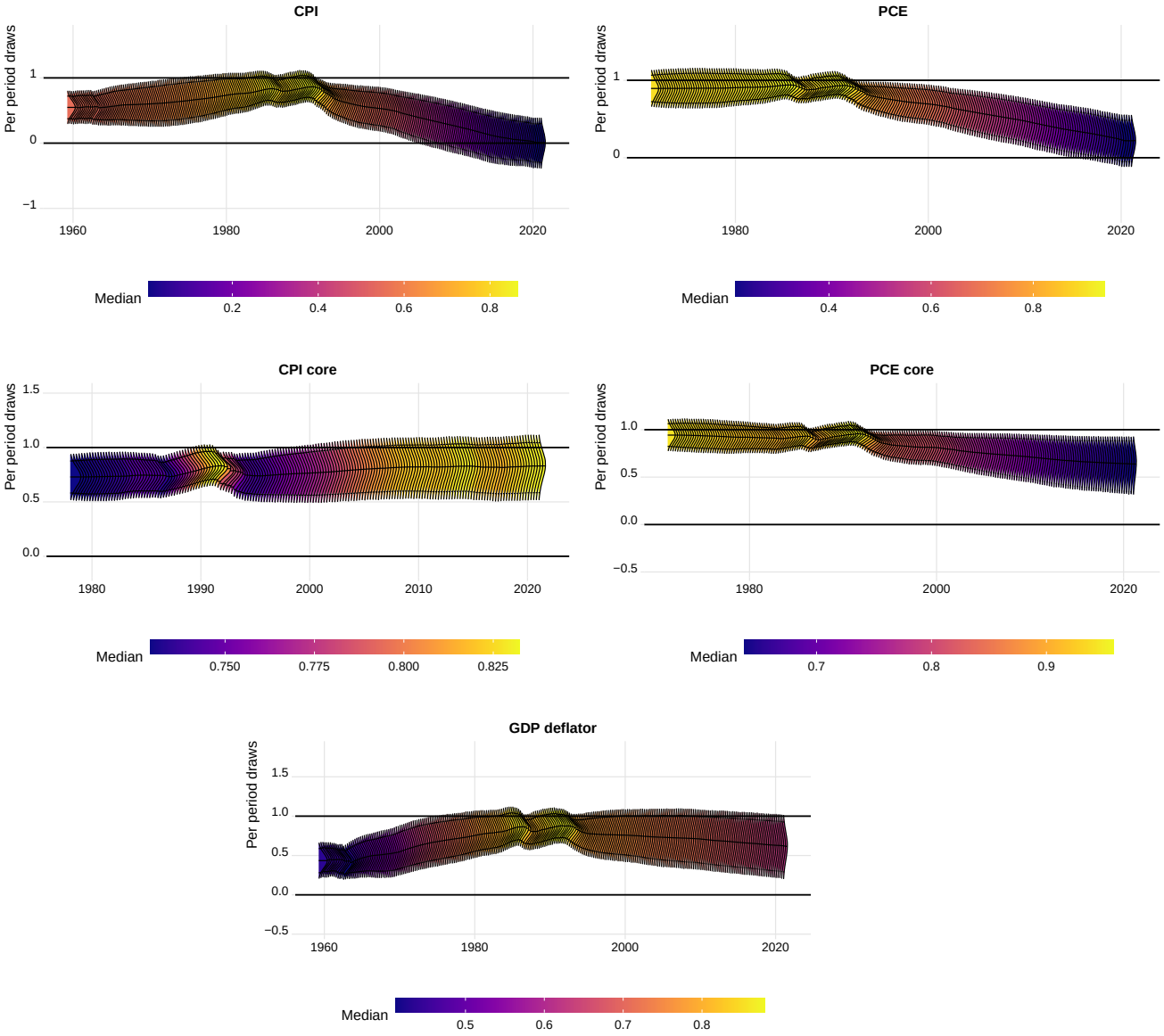


World Food Commodities Index for industrial raw materials, produce, beverage; level (top), and month on month annualized percent change (bottom). Source: IMF.

D Draw distributions

These plots provide further insights on the densities of persistence produced by the Bayesian analysis presented in Sec.(4).

Figure 13: Bayesian Draws – Per Period Full Densities



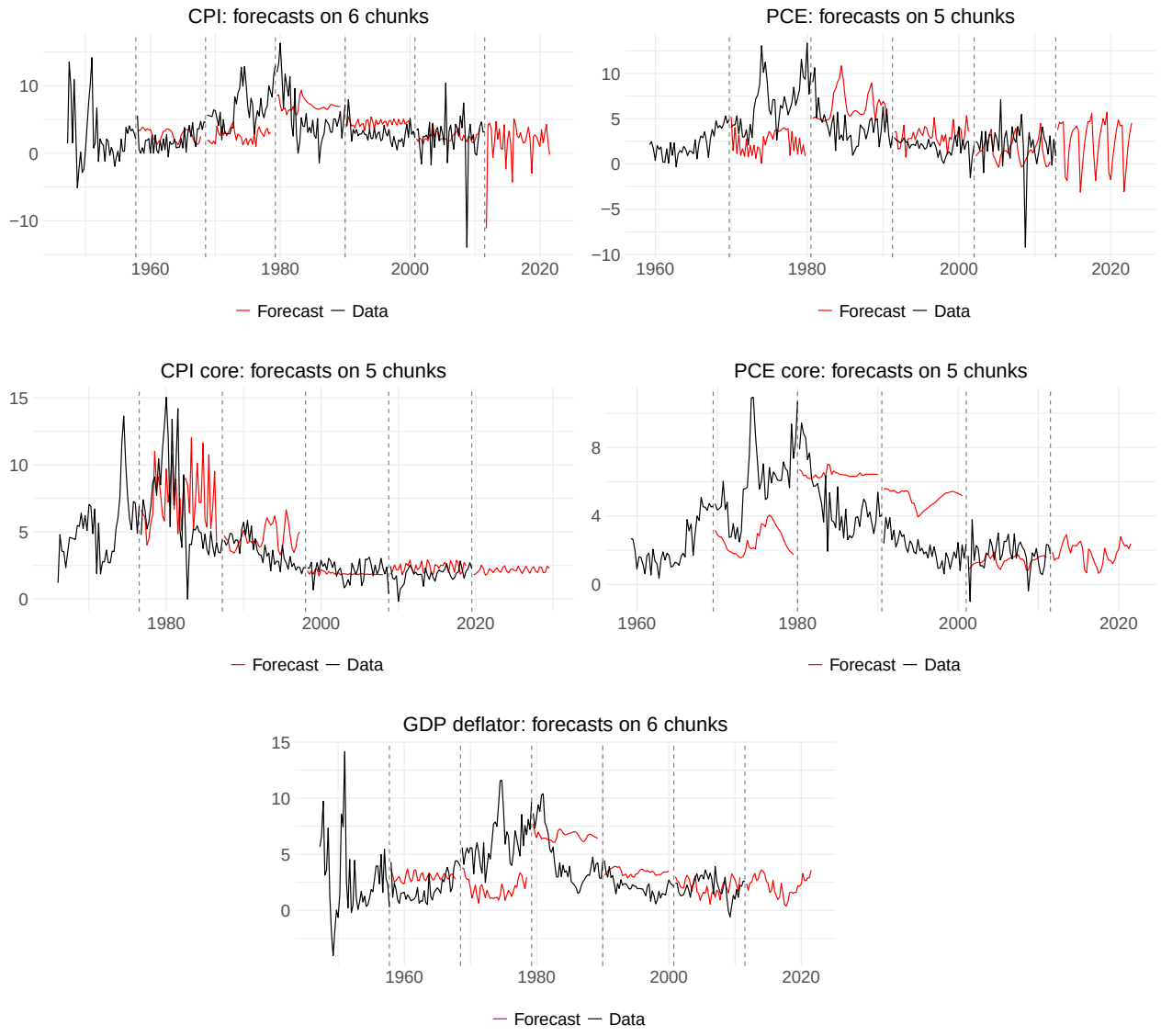
Per-period estimated densities: sum of $AR\,(3)$ draws distributions per quarter. Persistence computed on simulated forward trajectories based on data up to t. 300k total iterations, 150k burn-in, resulting in 150k conserved draws per period. Each density also reports 5%, 50%, and 95% percentiles. Colour depends on median value for each time t density.

E LSTM data and forecasts

This section collects plots for the forecasts of all LSTMs mentioned in the main body of the paper.

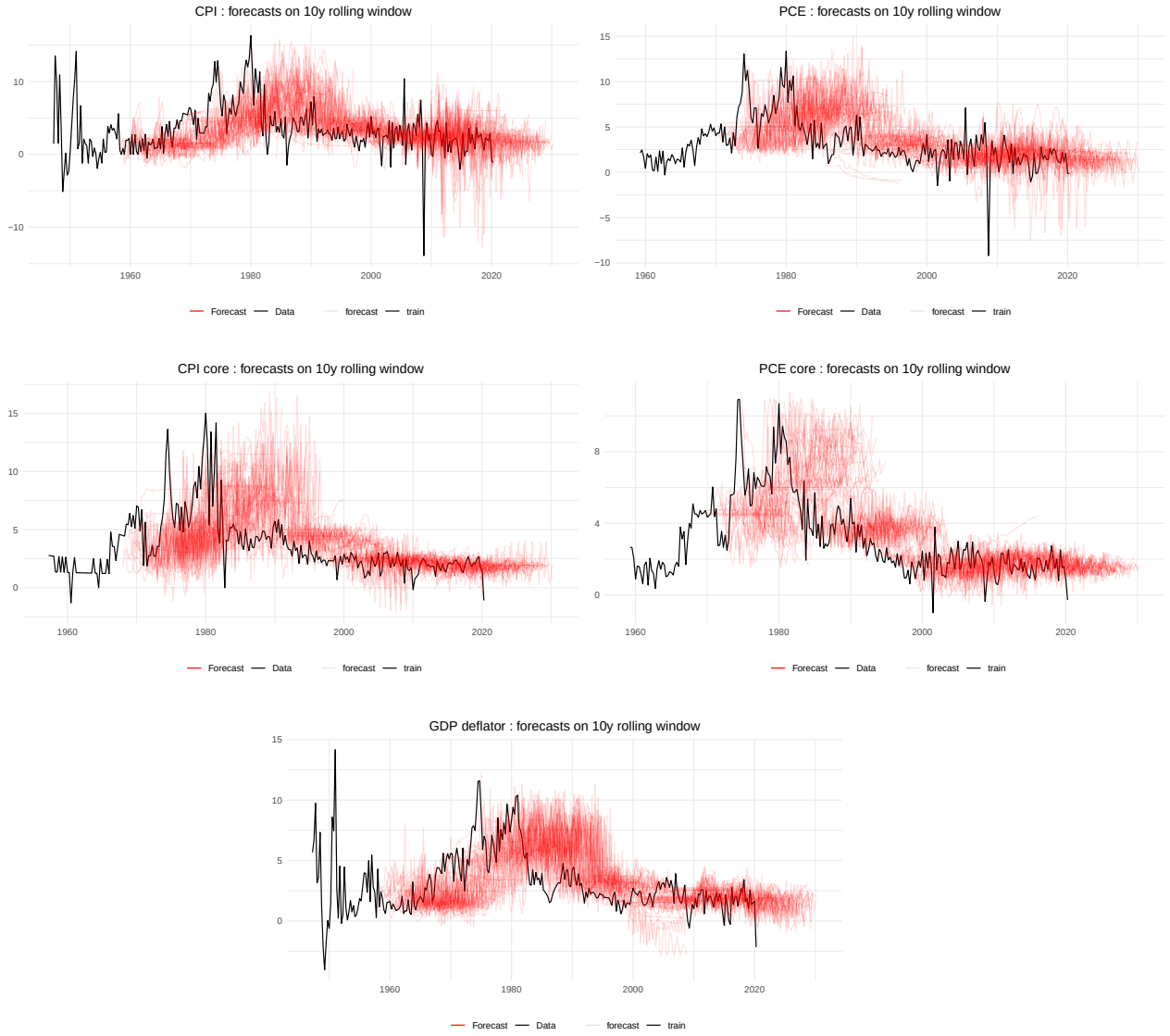
E.1 LSTM predictions on non-overlapping subsamples

Figure 14: LSTM Forecasts on Decade Subsamples



Indirect forecasts from an LSTM trained on 10 years of data, plus appropriate lags defined by BIC minimisation. One layer, 500 nodes, early stopping criterion with 2000 epochs upperbound. Forecast horizon is $h = 40$: first prediction is produced from last available data point and then iterated forward. Dashed vertical lines mark data subsamples' end date.

Figure 15: LSTM Rolling Window Forecasts



Indirect forecasts from a several LSTMs trained on a 10-year rolling window, plus appropriate lags defined by BIC minimisation. One layer, 500 nodes, early stopping criterion with 2000 epochs upper-bound. Forecast horizon is $h = 40$: first prediction is produced from last available data point and then iterated forward.

F LSTM analyses

F.1 OLS regressions on decades

The following tables present detailed information on the OLS regressions that produce the results plotted in subsection 5.4. Each table's column represents a subsample of actual data points augmented with the forecast produced by the trained LSTM. Hence, the number of observations results from $40 + k$ data points and 40 forecasts, where k is the optimal number of lags for each series. The left-most column is the oldest subsample, right-most is the closest in time.

Table 4: CPI decades regressions with LSTM forecasts

	Dependent variable:					
	CPI					
	1947Q2 1957Q4	1958Q1 1968Q3	1968Q4 1979Q2	1979Q3 1990Q1	1990Q2 2000Q4	2001Q1 2011Q3
1^{st} lag	.438***	.277***	.703***	.617***	.345***	−.108
	(.100)	(.103)	(.079)	(.082)	(.104)	(.111)
Constant	1.452***	1.604***	2.050***	1.802***	1.705***	2.313***
	(.403)	(.272)	(.568)	(.465)	(.293)	(.423)
Observations	82	82	82	82	82	82
R^2	.192	.083	.496	.414	.120	.012
Adjusted R^2	.182	.071	.490	.407	.109	−.001
F Statistic (df = 1; 80)	19.017***	7.215***	78.882***	56.537***	10.914***	.932

Note: *p<0.1; **p<0.05; ***p<0.01

Table 5: PCE decades regressions with LSTM forecasts

	Dependent variable:				
	PCE				
	1959Q2 1969Q4	1970Q1 1980Q3	1980Q4 1991Q2	1991Q3 2002Q1	2002Q2 2012Q4
1^{st} lag	.777***	.541***	.410***	.300***	.0004
	(.071)	(.093)	(.093)	(.106)	(.112)
Constant	.526***	3.267***	2.196***	1.174***	2.118***
	(.194)	(.685)	(.382)	(.205)	(.321)
Observations	82	82	82	82	82
R^2	.602	.299	.197	.091	0.00000
Adjusted R^2	.597	.290	.187	.080	−.012
F Statistic (df = 1; 80)	121.230***	34.052***	19.637***	8.005***	.00001
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 6: GDP Deflator decades regressions with LSTM forecasts

	Dependent variable:					
	GDP deflator					
	1947Q2 1957Q4	1958Q1 1968Q3	1968Q4 1979Q2	1979Q3 1990Q1	1990Q2 2000Q4	2001Q1 2011Q3
1^{st} lag	.467***	.653***	.663***	.897***	.582***	.723***
	(.098)	(.081)	(.083)	(.039)	(.082)	(.080)
Constant	1.466***	.659***	2.171***	.352**	.854***	.605***
	(.364)	(.177)	(.550)	(.171)	(.182)	(.184)
Observations	82	82	82	82	82	82
R^2	.222	.451	.442	.870	.383	.507
Adjusted R^2	.212	.444	.435	.869	.376	.501
F Statistic (df = 1; 80)	22.777***	65.679***	63.292***	537.595***	49.747***	82.417***
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 7: Core CPI decades regressions with LSTM forecasts

	Dependent variable:				
	CPI core				
	1966Q1 1976Q3	1976Q4 1987Q2	1987Q3 1998Q1	1998Q2 2008Q4	2009Q1 2019Q3
1^{st} lag	.816***	.656***	.828***	.200*	.501***
	(.064)	(.084)	(.061)	(.110)	(.097)
Constant	1.463**	1.867***	.442**	1.773***	.981***
	(.635)	(.513)	(.181)	(.252)	(.196)
Observations	82	82	82	82	82
R^2	.672	.430	.698	.040	.249
Adjusted R^2	.668	.423	.694	.028	.239
F Statistic (df = 1; 80)	163.902***	60.265***	184.684***	3.330*	26.504***
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 8: Core PCE decades regressions with LSTM forecasts

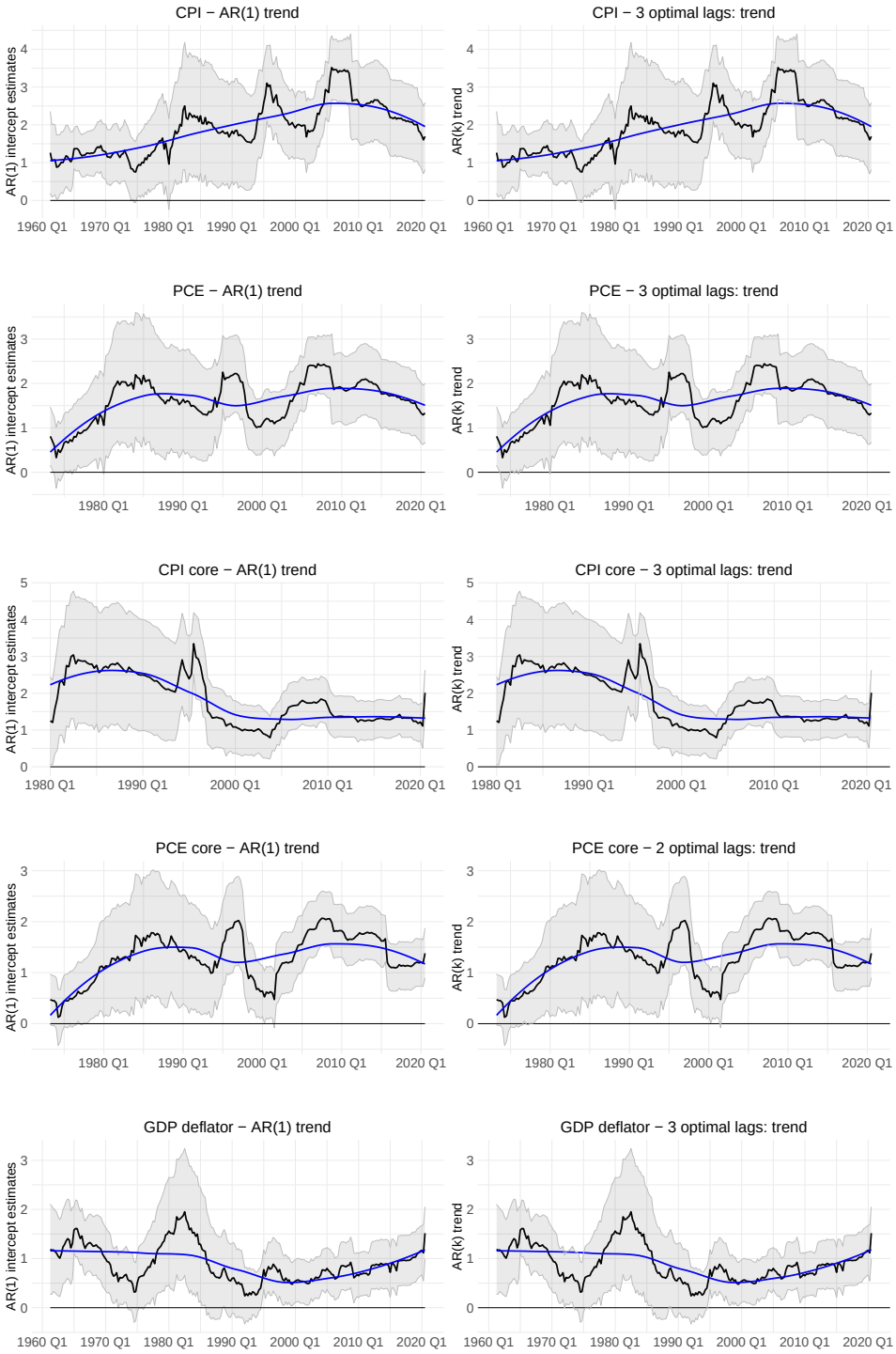
	Dependent variable:				
	PCE core				
	1959Q2 1969Q3	1969Q4 1980Q1	1980Q2 1990Q3	1990Q4 2001Q1	2001Q2 2011Q3
1^{st} lag	.876***	.733***	.684***	.377***	.368***
	(.055)	(.076)	(.076)	(.104)	(.105)
Constant	.368**	1.741***	1.323***	.976***	1.185***
	(.169)	(.500)	(.345)	(.188)	(.214)
Observations	81	81	81	81	81
R^2	.765	.544	.509	.143	.133
Adjusted R^2	.762	.538	.503	.133	.122
F Statistic (df = 1; 79)	257.018***	94.107***	81.810***	13.226***	12.163***
Note:					*p<0.1; **p<0.05; ***p<0.01

G Trend estimates

This section collects estimates on trend inflation that results from two approaches explored in the body of the paper. For both the frequentist and the LSTM applications, a side product of estimating autoregressive models of varying order is the intercept.

G.1 Frequentist application

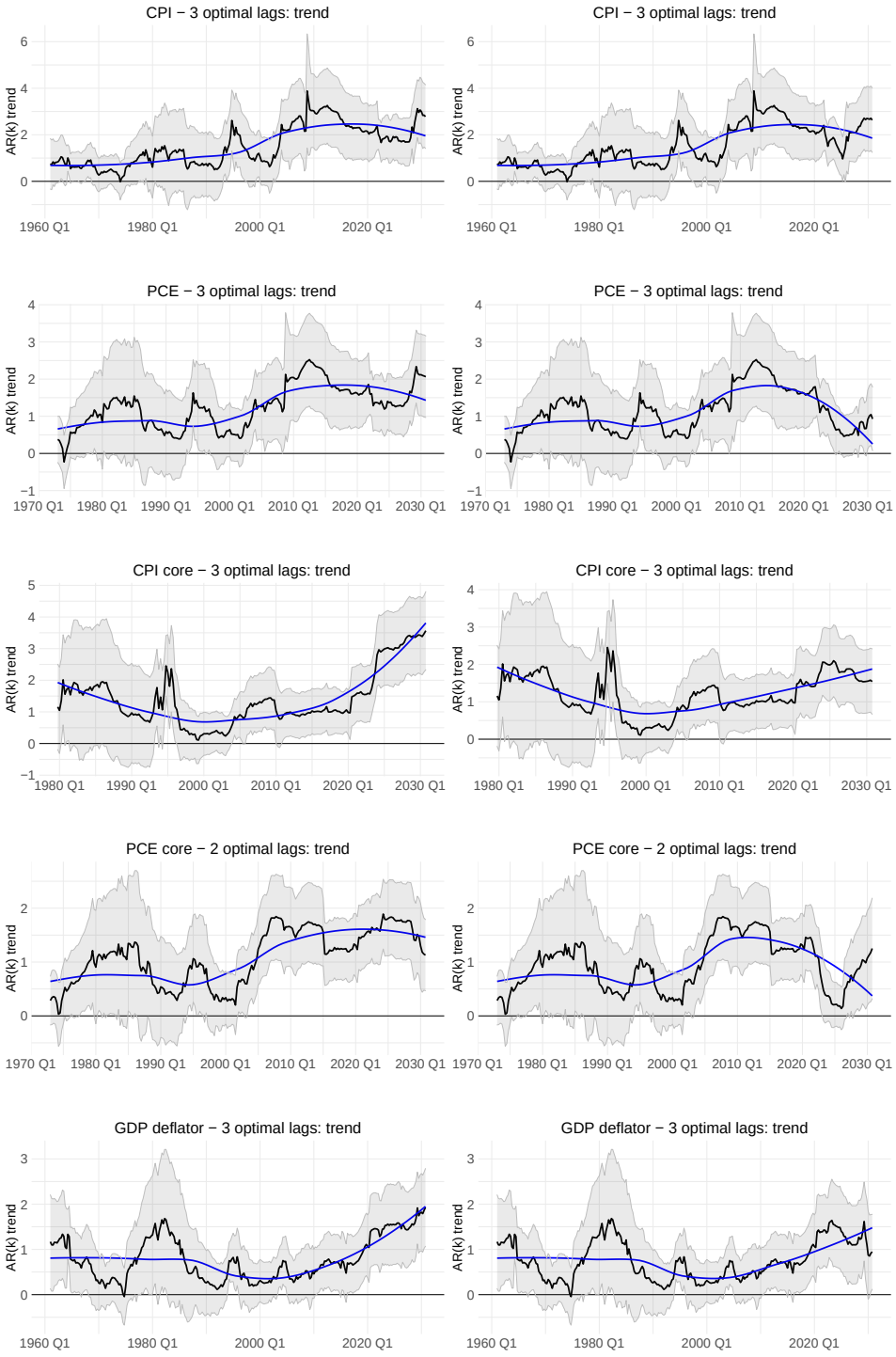
Figure 16: Trend Inflation – Frequentist Measure



Frequentist estimates of trend inflation, computed as intercept of an autoregressive process. Left column: AR(1). Right column: $AR(k^*)$ with lags selected by BIC minimisation.

G.2 LSTM output: full sample

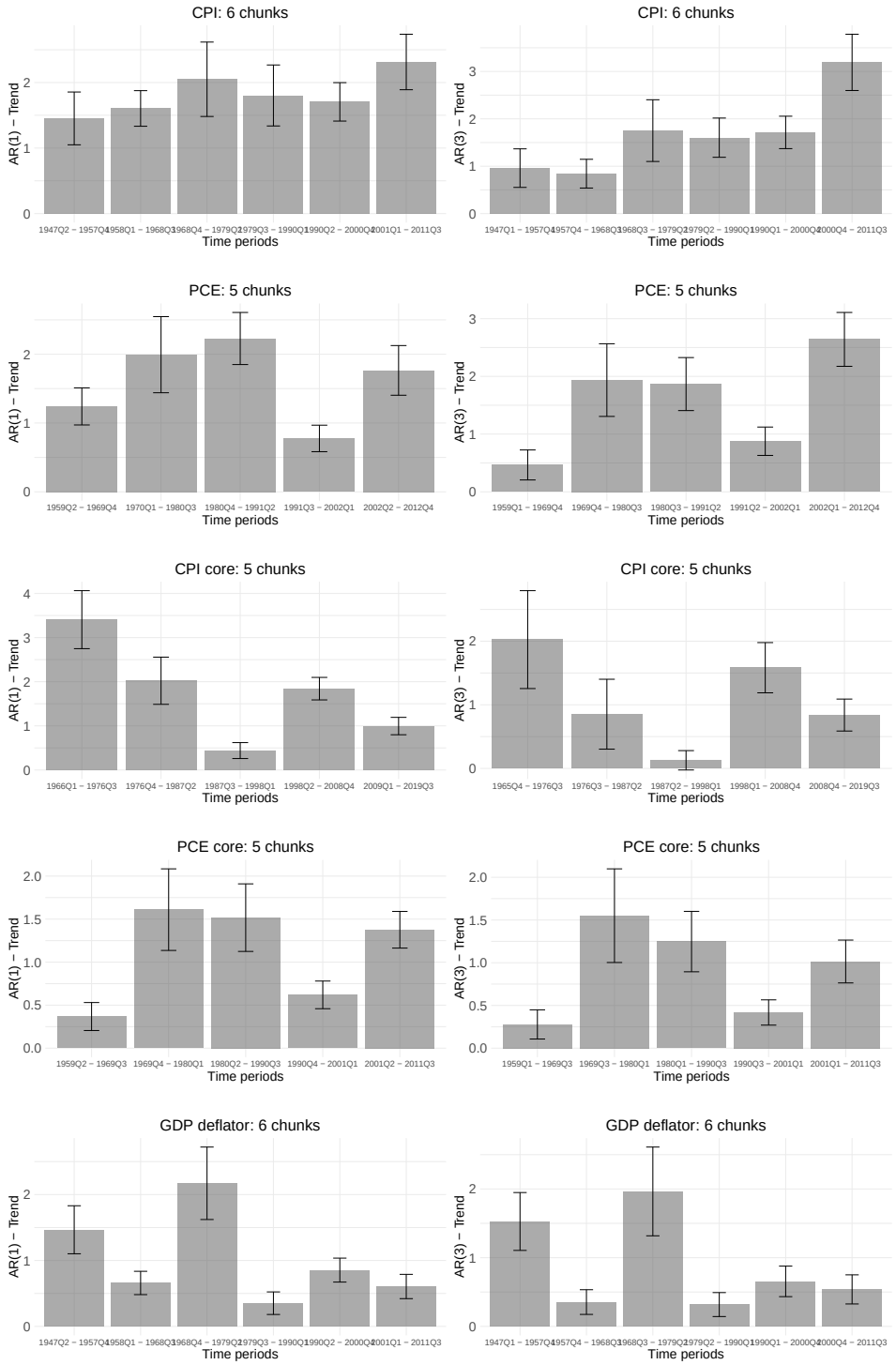
Figure 17: Trend Inflation – LSTM on Full Sample



LSTM trained on the full sample of data and then iterated forward to forecast 40 data points. Left column: one-layer net. Right column: two-layer net. All estimates are from an $AR(k^*)$ with lags minimising BIC.

G.3 LSTM output: subsample analysis

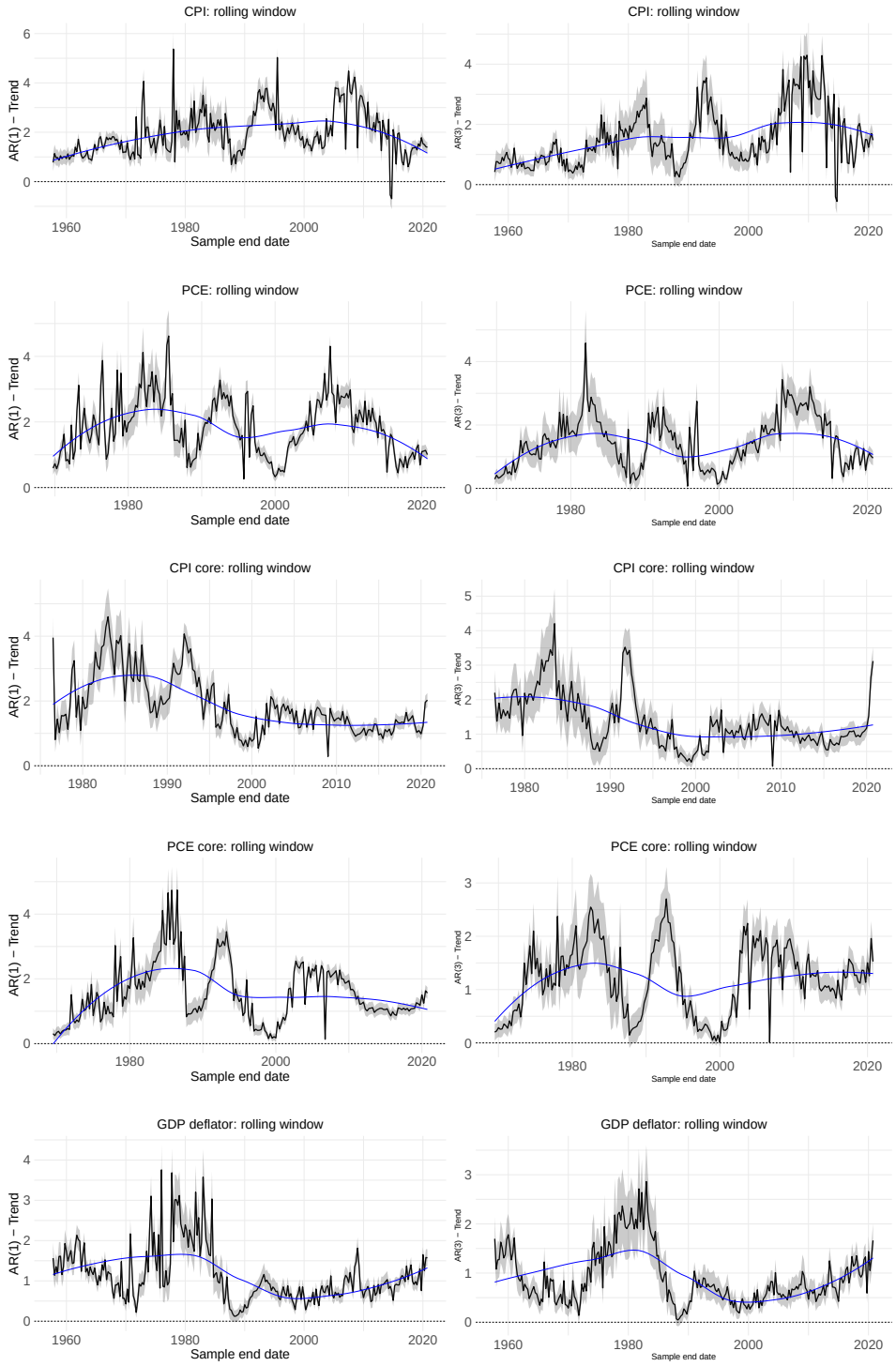
Figure 18: Trend Inflation – LSTM on Decade Subsamples



LSTMs trained on ten-year subsamples, plus appropriate lags, then iterated forward to produce 40 data points. Left column: $AR(1)$. Right column: $AR(k^*)$ with lags minimising BIC.

G.4 LSTM output: rolling window

Figure 19: Trend Inflation – LSTM on Rolling Windows



LSTMs trained on 56-quarter rolling windows, plus appropriate lags, then iterated forward to produce 40 data points. Left column: $AR(1)$. Right column: $AR(k^*)$ with lags minimising BIC.