Forecasting Economic Growth with Regime Changes and Uncertainty Quantification

Guilherme Piantino and Hedibert F. Lopes

Insper Institute of Education and Research

ARTICLE HISTORY
Compiled April 6, 2025

ABSTRACT
We evaluate the performance of Gaussian Process (GP) models of varying depth (standard GP, two-layer and three-layer Deep Gaussian Process, DGP) in forecasting U.S. real Gross National Product growth, with a particular focus on capturing regime changes and tackling uncertainty quantification (UQ). We compare these models to traditional benchmarks, including the Markov-Switching (MS) models and ARIMA-type models. The results show that the three-layer DGP consistently yields the most accurate forecasts and best-calibrated predictive distributions across horizons. Notably, the ARIMA model emerges as a competitive benchmark, frequently rivaling deep models in both point prediction and probabilistic accuracy. These findings underscore the value of hierarchical Bayesian modeling in economic forecasting and highlight the potential of DGPs as a tool for capturing nonstationarity and uncertainty in macroeconomic time series.

KEYWORDS
Deep Gaussian Processes; Regime Switching; Macroeconomic Forecasting; Nonstationarity; Bayesian Nonparametrics; Uncertainty Quantification

1. Introduction

Macroeconomic time series, such as the U.S. real Gross National Product (GNP) growth rate, are often subject to structural changes driven by evolving economic conditions and policy changes. These transitions, frequently aligned with turning points in the business cycle, impose challenges for traditional econometric models that assume stationarity or linearity. As a result, accurate forecasting in such environments requires flexible modeling frameworks and robust uncertainty quantification to support risk-aware decision-making.

Traditional approaches, including ARIMA models, deterministic trends, and linear state-space models, often impose restrictive assumptions that fail to capture the nonlinear transitions observed during business cycles, e.g.Beveridge and Nelson (1981), Nelson and Plosser (1982), and Campbell and Mankiw (1987).

In response, a foundational strand of time series econometrics has focused on structural change through regime-switching models. These frameworks allow the parameters of the data-generating process to shift across regimes, thereby accommodating

Corresponding author: Guilherme Piantino, guilhermejlp@al.insper.edu.br

structural breaks, policy changes, and cyclical dynamics, overcoming key limitations of constant-parameter models in macroeconomic applications.

1.1. Regime-Switching Models: Theory and Evolution

One early strand of this literature is based on threshold models, introduced by Tong (1983) and surveyed by Potter (1999), where regime changes are triggered deterministically when observable variables cross pre-specified thresholds. These models are particularly appealing when regime shifts are believed to be governed by known economic conditions, such as interest rates exceeding a policy threshold.

A more widely adopted and flexible class of models is the family of Markov-switching models, in which regime transitions are governed by an unobserved discrete-state Markov process. This framework was pioneered by Goldfeld and Quandt (1973) and Cosslett and Lee (1985), and brought to prominence by Hamilton (1989), who demonstrated that U.S. real GDP growth could be characterized by a mean-shifting autoregressive process, with the latent regime closely aligned with NBER business cycle classification. These models allow key parameters (mean, variance, and coefficients) to switch across regimes, capturing structural changes and nonlinear dynamics in macroeconomic behavior.

However, despite their theoretical appeal and strong in-sample performance, these models often perform poorly in out-of-sample forecasting tasks. As highlighted by Dacco and Satchell (1999), even when the true model is correctly specified, small misclassifications in predicting the active regime can lead to higher forecast errors than simpler models such as a random walk or a random walk with drift. Their findings suggest that the usefulness of regime-switching models in forecasting is limited by the difficulty of accurately identifying future regimes, especially when transitions are latent and not directly observable.

Subsequent developments extended this framework in several directions. Multivariate extensions, such as the Markov-Switching Vector Autoregression (MS-VAR) model by Krolzig (1997), enabled the joint modeling of multiple macroeconomic indicators. Meanwhile, regime-switching volatility models, including Markov-switching GARCH frameworks by Cai (1994), Hamilton and Susmel (1994), and Gray (1996), provided tools to model abrupt changes in financial market volatility.

Recognizing the limitation of fixed transition dynamics, Filardo (1994) and Diebold et al. (1994) introduced time-varying transition probability (TVTP) models, in which the switching behavior evolves with observable covariates. Additionally, Kim et al. (2003) developed models with endogenous switching, where transitions are correlated with contemporaneous shocks, enhancing the ability to capture asymmetries and nonlinear responses in macroeconomic series.

Bayesian estimation has played a central role in regime-switching analysis, particularly in addressing parameter uncertainty and latent state inference. Key contributions include the Gibbs sampling approaches of Kim and Nelson (1998), and the Bayesian change-point model by Chib (1998), which incorporates a latent Markov process with constrained transitions. These methods offer robust tools for estimating complex models with hidden structures.

More recently, advances in Bayesian nonparametrics and machine learning have further expanded the regime-switching paradigm. Fox et al. (2011) proposed a Bayesian nonparametric hidden Markov model using a Dirichlet process prior, allowing for an unbounded number of latent regimes. Wu et al. (2018) introduced a deep generative

state-space framework that integrates neural networks with regime-switching dynamics, improving the capacity to learn nonlinear and time-varying patterns directly from data.

Liu and Nguyen (2015) developed a tree approach to option pricing under a regime-switching jump diffusion framework. Their method uses a trinomial tree structure to efficiently model regime-dependent jump dynamics, offering a tractable and interpretable alternative to continuous-time diffusion models, and showcasing the potential of tree-based approaches in capturing regime-sensitive financial behaviors.

Similarly, Bie et al. (2024) contributed a tree-based macroeconomic regime-switching model in the context of the Nelson-Siegel yield curve, using a Bayesian method to choose optimal split candidates based on the marginal DNS likelihood. Their model reveals regime-dependent predictability in U.S. Treasury yields, particularly when short-term rates are high and offers interpretability and computational simplicity relative to traditional Markov-switching methods.

In a related development, Hauzenberger et al. (2024) proposed a Gaussian Process Vector Autoregression (GP-VAR) framework, combining the flexibility of GPs with full Bayesian inference to capture nonlinearities in multivariate time series.

Equally important to accurate modeling regime changes is the ability to quantify uncertainty around estimates and predictions. Fully Bayesian approaches offer a principled framework for robust inference under structural change by producing posterior distributions over both latent states and model parameters.

In this sense, we decided to adopt a Gaussian Processes (GPs) method. Their probabilistic formulation makes them especially well-suited for uncertainty quantification, as they provide full posterior distributions over functions rather than point estimates. This allows uncertainty to be explicitly propagated through forecasts. GPs are particularly advantageous in settings with small to moderate datasets, a common feature of macroeconomic time series, which are often observed quarterly or monthly. Unlike many machine learning models that require large sample sizes to generalize effectively, GPs naturally adjust their uncertainty in data-scarce regions, reducing the risk of overfitting.

The kernel function, which defines the covariance structure of the model, encodes prior beliefs about the underlying data (e.g. smoothness, periodicity, or long-term trends), enhancing interpretability and allowing for the modular integration of economic theory. Moreover, the kernel also acts as a built-in regularizer by controlling the function space that the GP can explore, helping to prevent overfitting, especially in noisy or data-sparse environments.

However, standard GPs typically rely on stationary kernels and pairwise input distances, which assume a constant covariance structure across the input space. This assumption limits their capacity to capture structural changes, abrupt regime transitions, or time-varying volatility, features often present in macroeconomic data.

1.2. Nonstationary Gaussian Processes: Advances and Approaches

To address the limitations of standard GPs, several approaches have been developed to accommodate nonstationary behavior in the data. One solution is to replace stationary kernels with spatially-varying ones. Higdon et al. (1999) introduced process convolutions, which were later extended by Paciorek and Schervish (2003) and Katzfuss (2013) using Matérn kernels, allowing local variation in smoothness and enabling full Bayesian inference via MCMC. Another approach involves modeling functional hy-

perparameter, such as input-dependent lengthscales or variances from Heinonen et al. (2016) or the heteroskedastic GP framework from Binois et al. (2018), which allows the noise variance to vary across the input space.

Another approach is to partition the input space into subregions and fits local GPs with distinct kernels or hyperparameters. Known as divide-and-conquer, this includes Dirichlet process partitioning (Rasmussen and Ghahramani, 2001), partitions defined from a Voronoi tesselation (Kim et al., 2005), and regression-tree-based such as treed GPs (Gramacy and Lee, 2008). Local approximate GPs (Gramacy and Apley, 2015) extend this framework by using local conditioning sets to achieve scalability in large datasets. While these models offer local adaptivity and computational efficiency, they may sacrifice global coherence, making uncertainty quantification more challenging across the full input space.

A more recent and increasingly powerful approach involves learning nonlinear transformations (or warpings) of the input space to induce stationarity in a transformed representation. Foundational work by Sampson and Guttorp (1992) and Schmidt and O'Hagan (2003) laid the groundwork for these techniques. Building on this idea, Damianou and Lawrence (2013) introduced Deep Gaussian Processes (DGPs), which model hierarchical warpings by stacking multiple GP layers. Each layer transforms its input through a learned GP, allowing the final prediction to be made in a deeply warped and highly adaptive space. This hierarchical composition enables DGPs to capture both smooth and abrupt transitions, making them particularly well-suited to settings characterized by complex, time-varying dynamics.

Inspired by both spatial modeling and deep learning, DGPs have gained popularity in the machine learning community due to their conceptual connection to deep neural networks. Modern implementations of DGPs use variational inference or MCMC to infer the latent transformations and hyperparameters, which retains the fully Bayesian nature of standard GPs. Although computationally intensive, DGPs have demonstrated strong performance in nonstationary and high-noise settings.

While kernel-based models emphasize interpretability and divide-and-conquer approaches prioritize scalability, DGPs stand out for their capacity to model complex nonstationary structure within a coherent Bayesian framework. Unlike the other methods, DGPs provide a unified, globally consistent architecture capable of learning intricate, hierarchical structures directly from data. They eliminate the need for handcrafted kernels or defined partitions and can flexibly adapt to both local and global variations in the underlying process. Importantly, DGPs retain the fully probabilistic treatment of uncertainty that characterizes Gaussian Processes, making them particularly well-suited for applications where uncertainty quantification is critical for decision-making.

1.3. Contributions and Forecasting Framework

In this paper, we adopt DGPs to model macroeconomic data subject to regime shifts, leveraging their ability to learn nonlinear, time-varying relationships directly from data. Unlike traditional regime-switching models that rely on predefined states or transition matrices, DGPs can infer regime changes in a flexible, data-driven manner. Through hierarchical warping, they offer a unified way to model both gradual trends and abrupt transitions without imposing rigid parametric structures. At the same time, they maintain full Bayesian inference, enabling uncertainty quantification around trend estimates and regime shifts.

This approach aligns with the view that macroeconomic outcomes are shaped by latent, evolving forces. As emphasized by Neftci (1984) and Sichel (1987), accounting for asymmetries and nonlinear dynamics is essential for understanding macroeconomic fluctuations. We apply DGPs to model and forecast U.S. real GNP growth, the same dataset used in the foundational work of Hamilton (1989). Our findings demonstrate that DGPs are capable of effectively capturing the complex, nonlinear, and time-varying dynamics characteristic of macroeconomic data, generating improved predictive performance and more reliable uncertainty quantification.

Building on this foundation, our paper contributes to the literature by employing DGPs for macroeconomic forecasting in the presence of structural changes. This approach allows the model to learn both smooth and abrupt regime transitions directly from the data, infer nonstationary patterns without the need for manually specified latent states or transition rules, and generate coherent predictive distributions. As shown in the results section, the DGP framework consistently outperforms standard GP and traditional forecasting models such as ARIMA and Markov-Switching, particularly in its ability to generalize across forecast horizons and quantify predictive uncertainty more effectively.

2. Methodology

In this section we define and discuss GPs and DGPs, as well as their posterior inference based on Markov chain Monte Carlo (MCMC) approximations. In what follows, the U.S. real Gross National Product (GNP) growth rate is the y_i variable, while the time index is the x_i variable. More specifically, the relationship between the response variable y and the vector of explanatory variables x is as follows, assuming a Gaussian noise, for $i = 1, \ldots, n$,

$$y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

for an yet to be specified function f and noise variance σ^2. More compactly,

$$\mathbf{y} = f(\mathbf{x}) + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}), \tag{2}$$

which yields the likelihood for the entire dataset:

$$p(\mathbf{y}|\mathbf{f}, \mathbf{x}, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}(y_i|f(x_i), \sigma^2) \tag{3}$$

2.1. Gaussian Processes

GPs provide a flexible, nonparametric Bayesian framework for modeling functions. A GP defines a distribution over functions $f(\mathbf{x})$, such that any finite collection of function values follows a multivariate normal distribution. Formally, a GP is fully specified by a mean function and a covariance (kernel) function. The mean function is:

$$\mu(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] \tag{4}$$

And the covariance function:

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - \mu(\mathbf{x}))(f(\mathbf{x}') - \mu(\mathbf{x}'))] \tag{5}$$

This is written in the form of a prior distribution for the unknown f function

$$f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{6}$$

Given a data set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ are input vectors (in our case the time index) and $y_i \in \mathbb{R}$ are corresponding scalar responses, the goal is to infer the underlying latent function $f(x_i)$ that generated observation y_i.

2.1.1. Posterior inference

Posterior inference proceeds by modeling the observed data through a joint Gaussian prior over the latent function values, combined with a likelihood function that relates those latent values to the observations. This yields a tractable posterior distribution for prediction and uncertainty quantification.

Applying Bayes' rule, the posterior distribution over the latent function is obtained by combining the prior with the likelihood:

$$p(f|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|f)\, p(f)}{p(\mathbf{y})} \tag{7}$$

Due to the conjugacy between the Gaussian prior (Eq. 6) and the Gaussian likelihood (Eq. 3), the posterior distribution remains Gaussian and can be computed in closed form. This posterior not only provides point predictions but also quantifies uncertainty around them via predictive variances.

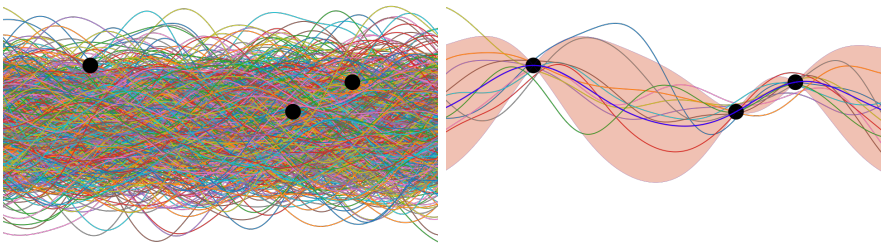


Figure 1.: Samples from a GP prior (left) and posterior (right) conditioned on observed data (black points). The right panel uses a Radial Basis Function (RBF) kernel with optimized hyperparameters. The shaded red region denotes the 95% credible intervals.

In GP regression, making predictions at new points involves conditioning the joint prior distribution on the observed training data. Assuming the standard Gaussian observation model (Eq. 2), we model the observations $\mathbf{y} \in \mathbb{R}^n$ as noisy realizations of an underlying latent function f, evaluated at input locations $\mathbf{X} \in \mathbb{R}^{n \times d}$. Assuming $\mu(x)$ is equal to zero, we can express joint beliefs over training outputs $\mathbf{y}$ and predicted outputs $\mathbf{f}_*$ at new test locations $\mathbf{X}_* \in \mathbb{R}^{n_* \times d}$ as a single multivariate Gaussian distribution:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(\mathbf{X},\mathbf{X}) + \sigma^2\mathbf{I} & K(\mathbf{X},\mathbf{X}_*) \\ \\ K(\mathbf{X}_*,\mathbf{X}) & K(\mathbf{X}_*,\mathbf{X}_*) \end{bmatrix}\right).$$

Here, $K(\cdot,\cdot)$ denotes the covariance matrix constructed using the kernel function. Conditioning this joint Gaussian on the observed data $(\mathbf{X},\mathbf{y})$, we obtain the predictive distribution for the function values $\mathbf{f}_*$ at test inputs $\mathbf{X}_*$. The resulting conditional distribution is also Gaussian, given by:

$$\mathbf{f}_* \mid \mathbf{X},\mathbf{y},\mathbf{X}_* \sim \mathcal{N}\left(\bar{\mathbf{f}}_*, \operatorname{cov}(\mathbf{f}_*)\right), \tag{8}$$

where the mean and covariance are computed analytically as:

$$\bar{\mathbf{f}}_* = K(\mathbf{X}_*,\mathbf{X})\left[K(\mathbf{X},\mathbf{X}) + \sigma^2\mathbf{I}\right]^{-1}\mathbf{y}, \tag{9}$$

$$\operatorname{cov}(\mathbf{f}_*) = K(\mathbf{X}_*,\mathbf{X}_*) - K(\mathbf{X}_*,\mathbf{X})\left[K(\mathbf{X},\mathbf{X}) + \sigma^2\mathbf{I}\right]^{-1}K(\mathbf{X},\mathbf{X}_*). \tag{10}$$

The predictive mean $\bar{\mathbf{f}}_*$ represents the expected value of the function at the test points, while the predictive covariance $\operatorname{cov}(\mathbf{f}_*)$ quantifies the associated uncertainty, taking into account both the prior and the observed data.

2.1.2. Choice of the kernel

In this work, we employ the Squared Exponential (SE) kernel, also known as the Radial Basis Function (RBF) kernel, to define the prior covariance structure. This kernel is widely used for its flexibility and analytical tractability. It is given by:

$$K(\mathbf{x},\mathbf{x}') = \sigma^2 \exp\left(-\frac{||\mathbf{x}-\mathbf{x}'||^2}{2\theta^2}\right) + g\,\mathbb{I}_{i=j}, \tag{11}$$

where:

- σ^2 is the signal variance, controlling the vertical amplitude of function values;
- θ is the characteristic length-scale, determining how quickly correlation decays with distance;
- g is the noise variance (or nugget), included on the diagonal to account for observation noise or to improve numerical stability.

The SE kernel ensures smooth, infinitely differentiable function realizations and is often a default choice in many applications. Its hyperparameters, denoted by $\vartheta = (\sigma^2,\theta,g)'$, are typically learned from data by maximizing the marginal likelihood or through fully Bayesian treatment.

Despite its strong performance in many settings, the SE kernel assumes homogeneity and stationarity across the input domain, which may be too restrictive. In the context of macroeconomic time series, where structural changes and local behaviors

are common, these limitations motivate the exploration of nonstationary extensions such as Deep Gaussian Processes, which are introduced in the following subsection.

2.2. Deep Gaussian Processes

To capture complex, hierarchical relationships in data that may exhibit non-stationarity, heteroskedasticity, or structural changes, we employ Deep Gaussian Processes. DGPs extend the flexibility of standard GPs by stacking multiple layers of latent functions, where the outputs of one GP serve as the inputs to the next. This recursive structure allows the model to learn a series of non-linear transformations, effectively warping the input space and capturing deeper abstractions in the data.

Formally, in a two-layer DGP, we define the model as follows:

$$\mathbf{W} \sim \mathcal{GP}(0, K_1(\mathbf{X}, \mathbf{X}')), \quad \mathbf{Y} \sim \mathcal{GP}(0, K_2(\mathbf{W}, \mathbf{W}')), \tag{12}$$

where $\mathbf{X}$ represents the observed input data, $\mathbf{W}$ is a latent variable capturing an intermediate, warped representation of the input space, and $\mathbf{Y}$ is the observed response. Each layer has its own covariance function, which may differ in functional form or hyperparameterization, thus increasing the model's flexibility.

This hierarchical construction enables DGPs to capture highly non-linear and non-stationary relationships while maintaining the Bayesian benefits of standard GPs, such as principled uncertainty estimation. Moreover, the composition of multiple layers allows the model to represent functions with varying degrees of smoothness or local structure, which are otherwise difficult to model using shallow architectures or stationary kernels alone.

The basic two-layer DGP can be naturally extended to deeper architectures by stacking additional Gaussian Process layers. In a three-layer DGP, for instance, the model is defined as:

$$\mathbf{Z} \sim \mathcal{GP}(0, K_1(\mathbf{X}, \mathbf{X}')), \quad \mathbf{W} \sim \mathcal{GP}(0, K_2(\mathbf{Z}, \mathbf{Z}')), \quad \mathbf{Y} \sim \mathcal{GP}(0, K_3(\mathbf{W}, \mathbf{W}')), \tag{13}$$

where $\mathbf{Z}$ and $\mathbf{W}$ are intermediate latent layers that encode representations of the input $\mathbf{X}$. Each layer applies a GP prior with potentially distinct kernel functions, allowing the model to successively warp and enrich the representation of the data. This recursive construction can be generalized to any number of layers, thereby increasing the capacity of the model to capture complex, hierarchical, and non-stationary patterns.

However, deeper architectures also introduce greater computational and inferential complexity, along with an increased risk of overfitting. In the results section, we detail the regularization strategies employed to mitigate these challenges and ensure robust generalization. In the next section, we will detail the algorithm used for sampling from the posterior.

2.2.1. Posterior Sampling Algorithm

Exact inference in DGPs is analytically intractable due to the nested GP structure and the requirement to integrate over multiple layers of latent variables. To make inference feasible, we adopt a hybrid MCMC approach as proposed by Sauer et al. (2023), which effectively balances computational tractability and statistical efficiency.

Specifically, our approach combines two sampling strategies:

- **Metropolis-Hastings (MH)** is used to update the hyperparameters of the covariance functions, including those governing both observed and latent layers.
- **Elliptical Slice Sampling (ESS)** is employed to sample the latent layer $\mathbf{W}$, exploiting its Gaussian process prior for efficient posterior exploration.

This hybrid sampler allows us to perform fully Bayesian inference over both the latent variables and model hyperparameters, even in the presence of non-Gaussian posteriors and complex hierarchical dependencies.

To facilitate posterior inference in our hierarchical model, we derive the log marginal likelihoods needed for evaluating acceptance probabilities in our MCMC sampler, integrating out the scale parameter under a reference prior. These likelihoods are computed conditionally on the latent variables and form the backbone of the MH and ESS steps.

The log marginal likelihood of the model, conditional on the latent layer $\mathbf{W}$, is given by:

$$\log \mathcal{L}(\mathbf{Y} \mid \mathbf{W}, \theta_y, g) \propto -\frac{n}{2} \log(n\hat{\sigma}^2) - \frac{1}{2} \log \left| K_{\theta_y}(\mathbf{W}) + gI_n \right|, \tag{14}$$

where

$$\hat{\sigma}^2 = \frac{\mathbf{Y}^\top (K_{\theta_y}(\mathbf{W}) + gI_n)^{-1} \mathbf{Y}}{n}, \tag{15}$$

and $K_{\theta_y}(\mathbf{W})$ is the covariance matrix from the second-layer GP, parameterized by θ_y, and g is the noise variance.

For the full two-layer model, the joint log-likelihood becomes:

$$\log \mathcal{L}(\mathbf{Y} \mid \mathbf{W}, \mathbf{X}, \theta, g) = \log \mathcal{L}(\mathbf{Y} \mid \mathbf{W}, \theta_y, g) + \log \mathcal{L}(\mathbf{W} \mid \mathbf{X}, \theta_w). \tag{16}$$

The joint log-likelihood decomposes into two distinct components: the likelihood of the observed response given the latent representation, and the likelihood of the latent representation given the original inputs. Specifically, the term $\log \mathcal{L}(\mathbf{Y} \mid \mathbf{W}, \theta_y, g)$ corresponds to the second GP layer and measures how well the latent variables $\mathbf{W}$ explain the observed outputs $\mathbf{Y}$, conditioned on the output-layer kernel hyperparameters θ_y and noise variance g. The second term, $\log \mathcal{L}(\mathbf{W} \mid \mathbf{X}, \theta_w)$, represents the marginal likelihood of the latent layer $\mathbf{W}$ under a GP prior with input $\mathbf{X}$ and hyperparameters θ_w.

The joint posterior distribution over the latent variables and hyperparameters in the two-layer DGP is expressed as:

$$\pi(\mathbf{W}, \theta, g \mid \mathcal{D}_n) \propto \mathcal{L}(\mathbf{Y} \mid \mathbf{W}, \mathbf{X}, \theta, g) \cdot \pi(\theta, g), \tag{17}$$

where $\mathcal{L}(\mathbf{Y} \mid \mathbf{W}, \mathbf{X}, \theta, g)$ denotes the joint marginal likelihood of the observed data, and $\pi(\theta, g)$ represents the prior distribution over the model's hyperparameters.

In our framework, we place Gamma priors on the kernel length-scale parameters and noise variance:

$$\theta \sim \text{Gamma}(3/2, b_\theta), \quad g \sim \text{Gamma}(3/2, b_g), \tag{18}$$

where b_θ and b_g are rate parameters selected based on empirical analysis, depending on the nature and scale of each hyperparameter. A more detailed explanation can be

found in Booth (2024).

As said before, we adopt a hybrid MCMC approach to perform posterior inference. This method combines MH updates for the hyperparameters with ESS for the latent variables. Posterior sampling proceeds iteratively as follows:

(1) Sample the noise variance $g^{(t)} \sim \pi(g \mid \mathbf{Y}, \mathbf{W}^{(t-1)}, \theta_y^{(t-1)})$ using MH.
(2) Sample the output-layer kernel hyperparameters $\theta_y^{(t)} \sim \pi(\theta_y \mid \mathbf{Y}, \mathbf{W}^{(t-1)}, g^{(t)})$ using MH.
(3) Sample the latent-layer kernel hyperparameters $\theta_w^{(t)} \sim \pi(\theta_w \mid \mathbf{W}^{(t-1)})$ using MH.
(4) Sample the latent variables $\mathbf{W}^{(t)} \sim \pi(\mathbf{W} \mid \mathbf{Y}, \mathbf{X}, \theta_y^{(t)}, g^{(t)})$ using ESS, which leverages the Gaussian prior structure on $\mathbf{W}$ to generate rejection-free proposals.

This inference scheme allows us to jointly learn both the structural properties of the GP layers and the latent representations, accessing uncertainty across all model components.

The ESS step for sampling $\mathbf{W}$ proceeds by:

(1) Drawing a prior sample $\mathbf{W}^{\text{prior}} \sim \mathcal{N}(0, K_{\theta_w}(\mathbf{X}))$.
(2) Sampling a rotation angle $\gamma \sim \text{Uniform}(0, 2\pi)$.
(3) Proposing $\mathbf{W}^* = \mathbf{W}^{(t-1)} \cos\gamma + \mathbf{W}^{\text{prior}} \sin\gamma$.
(4) Accepting $\mathbf{W}^*$ with probability:

$$\alpha = \min\left(1, \frac{\mathcal{L}(\mathbf{Y} \mid \mathbf{W}^*, \theta_y, g)}{\mathcal{L}(\mathbf{Y} \mid \mathbf{W}^{(t-1)}, \theta_y, g)}\right).$$

(5) If rejected, the bracket is shrunk and another angle is proposed until acceptance.

ESS offers a robust and efficient alternative to traditional Metropolis-Hastings algorithms, particularly in the context of DGPs. Unlike Metropolis-Hastings, which often suffers from poor mixing and low acceptance rates in hierarchical models, ESS provides rejection-free proposals that are well-suited for exploring high-dimensional, multi-modal, and highly correlated posteriors. Its tuning-free nature enhances computational efficiency and makes it especially effective for latent Gaussian models. By improving mixing and scalability, ESS enables a fully Bayesian treatment of DGPs in complex applications such as nonstationary time series forecasting. A fully Bayesian MCMC implementation of this framework is available through the deepgp R package, which is publicly accessible on CRAN Booth (2024).

3. Results

3.1. Dataset Overview

We utilize quarterly U.S. Real GNP data obtained from the Federal Reserve Economic Data (FRED) database, spanning the period from 1947 to 2024. This dataset is widely used in macroeconomic modeling and policy analysis and notably served as the basis for Hamilton (1989).

To contextualize the regime-switching nature of macroeconomic fluctuations, we incorporate the National Bureau of Economic Research (NBER) recession indicator (USREC), which classifies each quarter as either a **recession (1)** or an **expansion (0)**. This binary labeling provides a meaningful framework to assess the ability of probabilistic models to detect structural shifts in economic activity.

Figure 2 displays the real GNP growth rate over time alongside shaded regions corresponding to NBER defined recession periods. These episodes of economic contraction are clearly associated with sharp declines in output growth, highlighting the relevance of regime aware modeling strategies for capturing such nonlinear dynamics.

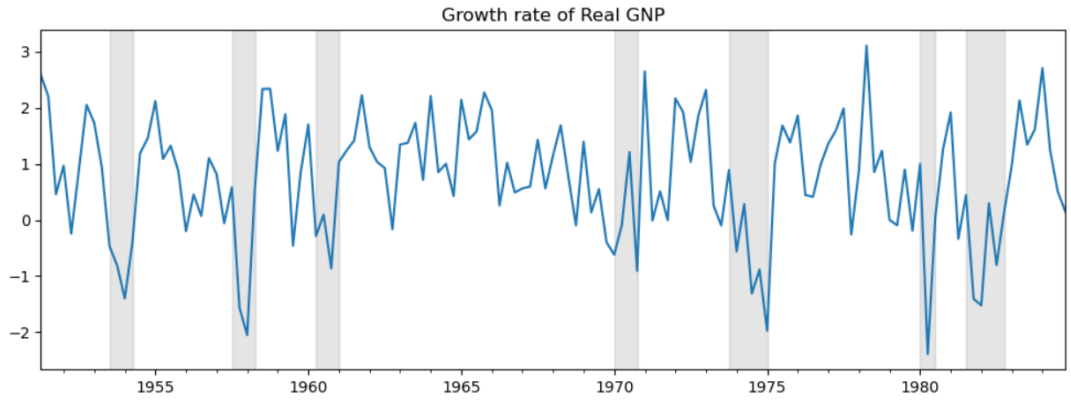


Figure 2.: Real GNP growth with NBER recession classification.

The results section is organized into two main parts: an in-sample model fit and an out-of-sample forecasting evaluation. In the first part, we assess the in-sample performance of a standard GP, a two-layer DGP, and a three-layer DGP, using quarterly U.S. Real GNP data over the period from 1952Q2 to 1984Q4, consistent with the sample window analyzed in Hamilton (1989). This evaluation focuses on the models' ability to fit historical data and capture regime-dependent dynamics. In the second part, we conduct a comprehensive out-of-sample forecasting exercise using a 12-step-ahead rolling window approach from 1984 to 2024. To benchmark performance, we include the classical Markov-Switching model and ARIMA as a references.

While deeper DGP architectures offer enhanced expressiveness and the capacity to model nonstationary and hierarchical relationships, they also introduce an increased risk of overfitting. To mitigate this, and in line with the framework proposed by Sauer et al. (2023), we constrain the depth of the DGP to a maximum of three layers, limit the dimensionality of latent variables (typically matching or remaining below the input dimensionality), and adopt a fully Bayesian inference approach via MCMC to integrate over the posterior distribution and avoid overconfident point estimates.

For evaluation, we computed two key metrics: Root Mean Squared Error (RMSE), which measures the average accuracy of point predictions, and the Continuous Ranked Probability Score (CRPS), which assesses the quality of the entire predictive distribution and is particularly informative for evaluating uncertainty quantification. The formulas and a more detailed explanation of the metrics is provided in the Appendix.

This dual analysis allows us to compare models in terms of both in-sample flexibility and out-of-sample generalization, with a particular focus on their ability to accommodate nonlinearities, non-stationarity, and structural changes in macroeconomic time series.

3.2. In-Sample Model Fit

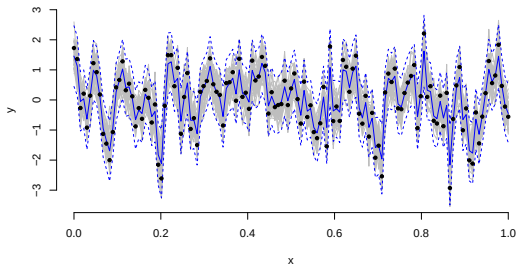
To evaluate each model's ability to represent historical macroeconomic dynamics, we begin by examining their in-sample performance over the period 1952Q2–1984Q4 using the specified architectures. We first analyze the posterior plots of each model's fit.

Figure (a) shows the posterior fit of a standard GP with a single layer. The model captures local fluctuations in the data but appears overly sensitive to short-term noise. The fit is highly variable, with narrow credible intervals that suggest overconfidence in uncertain regions. This behavior is typical of shallow GPs with stationary kernels applied to nonstationary data. They may struggle to capture broader structural patterns and instead focus on short-range correlations.

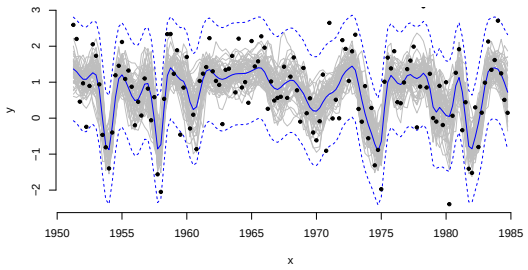
Figure (b) presents the output of a two-layer DGP. Compared to the shallow GP, this model captures smoother transitions and exhibits greater robustness to noise. The posterior mean more clearly tracks medium-run variations, while the credible intervals appropriately widen during volatile periods, indicating more realistic uncertainty quantification. This layered architecture enables the model to learn latent representations that better capture the evolving dynamics of the time series.

Figure (c) illustrates the predictive distribution of a three-layer DGP. This model provides the smoothest and most structured fit among the three. It successfully captures both long-term trends and localized changes in the data, while maintaining well-calibrated uncertainty bounds.

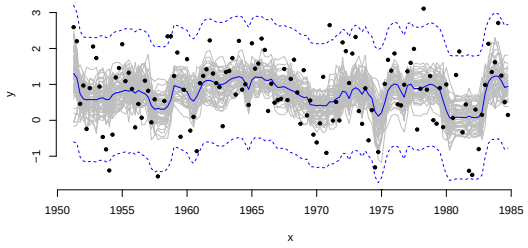
Overall, model depth appears to enhance the capacity to model macroeconomic dynamics with greater flexibility and more credible uncertainty estimates. While the shallow GP is prone to overfitting high frequency noise, deeper DGPs demonstrate superior smoothing, adaptability, and uncertainty quantification.



(a) Standard GP model with 1 layer



(b) Two-layer Deep Gaussian Process



(c) Three-layer Deep Gaussian Process

Figure 3.: Comparison of GP models by depth: (a) shallow, (b) two-layer, and (c) three-layer. The blue line represents the posterior mean function, the blue dashed lines denote the credible intervals, the gray lines correspond to prior GP samples, and the black dots indicate the observed values.

Now let´s analyze the performance metrics. As shown in Table 1, the standard GP achieves the lowest in-sample RMSE and a competitive CRPS, suggesting a strong fit

Model	RMSE	CRPS
One-layer GP	1.159	0.689
Two-layer DGP	1.304	0.773
Three-layer DGP	1.194	0.679

Table 1.: In-sample RMSE and CRPS for GP models trained on U.S. Real GNP (1952Q2–1984Q4).

to the training data.

The two-layer DGP exhibits higher in-sample RMSE and CRPS, which may reflect its greater model complexity and the added uncertainty introduced by latent transformations. The three-layer DGP offers a more favorable trade-off: it marginally improves upon the RMSE of the standard GP while achieving the lowest CRPS among all models. This indicates better-calibrated uncertainty estimates and more reliable predictive distributions.

However, these results should be interpreted with caution. A good in-sample fit may result from overfitting the training data, which can lead to poor generalization and degraded performance in out-of-sample forecasting scenarios.

3.3. Out-of-Sample Forecasting (12-Step Horizon)

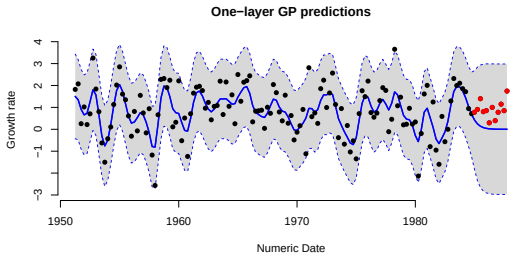
To evaluate the models' generalization capabilities beyond the training window, we perform a 12-step-ahead rolling forecast exercise over the period 1984 to 2024. At each step of the rolling window, the models are re-estimated and forecasts are generated for the subsequent twelve quarters. This iterative procedure ensures that each prediction is made using only past information, thereby mimicking real-world forecasting conditions and reducing look-ahead bias.

Forecasts are produced using the same time windows across all models. The GP models vary in depth but are otherwise matched in terms of noise assumptions and kernel families. This controlled setup allows us to isolate the effect of hierarchical composition and the use of non-stationary modeling on forecast performance.

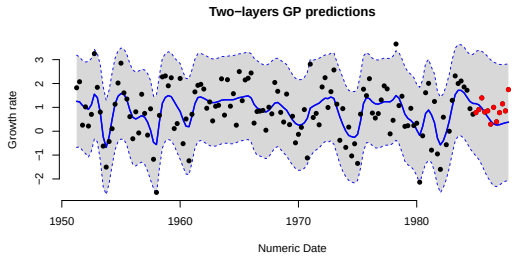
To build intuition about the behavior of each model during the initial stages of forecasting, we visualize predictions for the first 12-step horizon. These plots offer qualitative insights into how well each model captures short-term dynamics and handles the transition from in-sample fitting to out-of-sample forecasting.

Figure (a) presents the forecast produced by the standard GP. This model produces highly fluctuating forecasts with overly narrow credible intervals that reflect underestimation of predictive uncertainty. The forecast trajectory appears to extrapolate the local behavior of the last observed points rather than capturing underlying economic shifts.

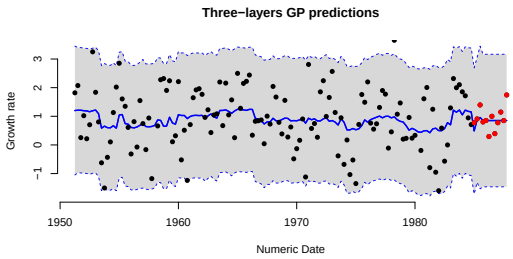
Figure (b) shows the forecasts from the two-layer DGP, which demonstrates a clear improvement in both predictive smoothness and uncertainty calibration. The posterior mean exhibits a more coherent pattern that aligns with medium-run dynamics, and the credible intervals expand in regions of heightened uncertainty. Notably, the red points fall mostly within the credible bands, highlighting the model's improved generalization capacity relative to the shallow GP.



(a) Standard GP model with 1 layer



(b) Two-layer Deep Gaussian Process



(c) Three-layer Deep Gaussian Process

Figure 4.: Comparison of GP models by depth: (a) shallow, (b) two-layer, and (c) three-layer. The blue line represents the posterior mean function, while the blue dashed lines and gray shaded areas indicate the credible intervals. The black dots correspond to observed values in the training sample, and the red dots represent the actual out-of-sample observations.

Figure (c) presents the forecasts from the three-layer DGP, which delivers the most stable and realistic predictions among all models. The posterior mean effectively captures long-term trends while remaining resilient to short-term fluctuations. The credible intervals display adaptive behavior, narrowing in periods of high confidence and appropriately widening during more uncertain transitions. The majority of the actual observations (red points) lie within the posterior predictive intervals, reflecting accurate uncertainty quantification and strong forecasting performance.

In summary, the rolling 12-step-ahead forecast evaluation highlights the strengths and weaknesses of each model architecture. While shallow models may overfit to recent trends and fail to extrapolate meaningfully, deeper DGPs demonstrate greater resilience and robustness when confronted with complex, evolving macroeconomic dynamics.

Model	**RMSE**	**CRPS**
One-layer GP	0.9222	0.5570
Two-layer DGP	0.5775	0.3934
Three-layer DGP	0.3776	0.3217

Table 2.: Out-of-sample RMSE and CRPS for GP models trained on U.S. Real GNP (1952Q2–1984Q4) and evaluated using a 12-step-ahead forecast.

Table **??** presents the out-of-sample forecasting performance of the Gaussian Process models, evaluated using RMSE and CRPS over a 12-step-ahead horizon.

The one-layer GP performs the worst across both metrics, with an RMSE of 0.9222 and a CRPS of 0.5570. These values suggest that the shallow GP, constrained by its stationary kernel and limited representational power, struggles to generalize beyond the training period, despite having shown strong in-sample performance. The two-layer DGP significantly improves upon the shallow GP, reducing the RMSE to 0.5775 and the CRPS to 0.3934. This improvement reflects the model's ability to better capture nonstationary and hierarchical features in the data through its latent representation layer.

The three-layer DGP yields the best performance, achieving the lowest RMSE (0.3776) and CRPS (0.3217), which demonstrates its superior ability to model complex macroeconomic dynamics and uncertainty. The results corroborate the graphical analysis and suggest that the additional depth enables the model to learn nuanced transformations of the input space and produce more robust long-horizon forecasts.

To comprehensively evaluate each model's predictive capacity, now we analyze forecast accuracy across all out-of-sample windows using step specific metrics. Then we report RMSE and the CRPS averaged by forecast horizon for the 12-step-ahead rolling forecasts spanning from 1984 to 2024.

Steps	M-S	ARIMA	GP	2-DGP	3-DGP
1	0.7014	0.5921	0.6392	0.6945	0.5466
2	0.8025	0.5801	0.6890	0.7552	0.5826
3	0.8174	0.5912	0.7574	0.7851	0.6285
4	0.8138	0.5830	0.7744	0.7668	0.5672
5	0.8470	0.5797	0.8031	0.7721	0.5581
6	0.8054	0.5768	0.8177	0.7718	0.5490
7	0.8338	0.5629	0.8343	0.7001	0.5540
8	0.8169	0.5723	0.8482	0.7757	0.5681
9	0.8208	0.5790	0.8629	0.7293	0.5611
10	0.8351	0.5822	0.8687	0.7160	0.5556
11	0.8248	0.5809	0.8692	0.7435	0.5843
12	0.8221	0.5811	0.8700	0.7505	0.5805

Table 3.: Average RMSE by forecast horizon (1984–2024)

The evaluation of forecast accuracy over a rolling 12-step horizon (Tables **??** and **??**) highlights substantial performance differences among the models.

The RMSE results show that the three-layer DGP consistently achieves the best predictive accuracy across all forecast horizons, maintaining the lowest error values from short- to long-term forecasts. At horizon 1, for instance, it yields an RMSE of 0.5466, outperforming all other models, including the standard GP (0.6392), the M-S model (0.7014), the two-layer DGP (0.6945), and ARIMA (0.5921). This performance advantage persists throughout the evaluation window. By horizon 12, the three-layer DGP still leads with an RMSE of 0.5805, followed closely by ARIMA at 0.5811, while the two-layer DGP (0.7505), standard GP (0.8700), and M-S model (0.8221) lag further behind. The ARIMA model demonstrates particularly strong and stable performance, ranking second overall and consistently outperforming the M-S, GP, and even the two-layer DGP models at nearly every step.

This performance advantage demonstrates the three-layer DGP's ability to general-

Steps	M-S	ARIMA	GP	2-DGP	3-DGP
1	0.5180	0.4872	0.5104	0.5786	0.4603
2	0.6226	0.4907	0.5583	0.7064	0.4786
3	0.6260	0.4960	0.5988	0.7653	0.5136
4	0.6299	0.4938	0.6133	0.7715	0.4771
5	0.6539	0.4908	0.6271	0.7395	0.4718
6	0.6195	0.4925	0.6336	0.7728	0.4724
7	0.6384	0.4794	0.6444	0.7164	0.4908
8	0.6212	0.4877	0.6502	0.7602	0.4944
9	0.6289	0.4912	0.6539	0.7199	0.4813
10	0.6380	0.4901	0.6568	0.6798	0.4836
11	0.6342	0.4892	0.6567	0.7190	0.5105
12	0.6335	0.4888	0.6569	0.7676	0.4819

Table 4.: Average CRPS by forecast horizon (1984–2024)

ize well over longer horizons, capturing both short-term and persistent nonlinearities. In contrast, the standard GP, constrained by its stationary kernel, tends to extrapolate local trends without adapting to structural changes, which leads to increasing forecast errors over time. The two-layer DGP performs moderately, outperforming the other models in mid-range horizons, but remaining inferior to the three-layer DGP overall. It is worth mentioning that the ARIMA model shows robust and consistent accuracy across all forecast steps, offering a strong parametric benchmark that frequently surpasses the M-S and GP models, and approaches the performance of the three-layer DGP.

Turning to the CRPS, the three-layer DGP again dominates, yielding the lowest values across nearly all forecast horizons, which underscores its superior ability to quantify predictive uncertainty. For example, at horizon 1, it achieves a CRPS of 0.4603, outperforming all other models, including the standard GP (0.5104), the M-S model (0.5180), and ARIMA (0.4872). As the forecast horizon increases, the three-layer DGP continues to deliver stable and relatively low CRPS values, adapting well to growing uncertainty in long-term forecasts. Among the remaining models, ARIMA consistently shows strong and reliable performance, often ranking just behind the three-layer DGP and surpassing both the M-S and standard GP models at every step.

Interestingly, both the standard GP and the M-S model exhibit relatively flat CRPS trajectories across forecast horizons, suggesting a limited capacity to adapt to the increasing uncertainty associated with longer-term predictions. The two-layer DGP, despite its added complexity, often yields higher CRPS values than the standard GP, indicating that the introduction of intermediate latent layers may inject additional uncertainty without providing sufficient representational power to effectively capture the underlying dynamics. In contrast, the ARIMA model displays a consistently low and stable CRPS profile, outperforming the M-S and GP models at all horizons and frequently surpassing the two-layer DGP as well.

Overall, the results reinforce the utility of deeper Gaussian Process architectures in macroeconomic forecasting. The three-layer DGP clearly emerges as the most effective model, offering the best trade-off between accuracy and calibrated uncertainty across all horizons. This supports the hypothesis that nonstationary, hierarchical latent structures are critical to capturing the evolving dynamics of real-world economic data, especially over long forecasting windows.

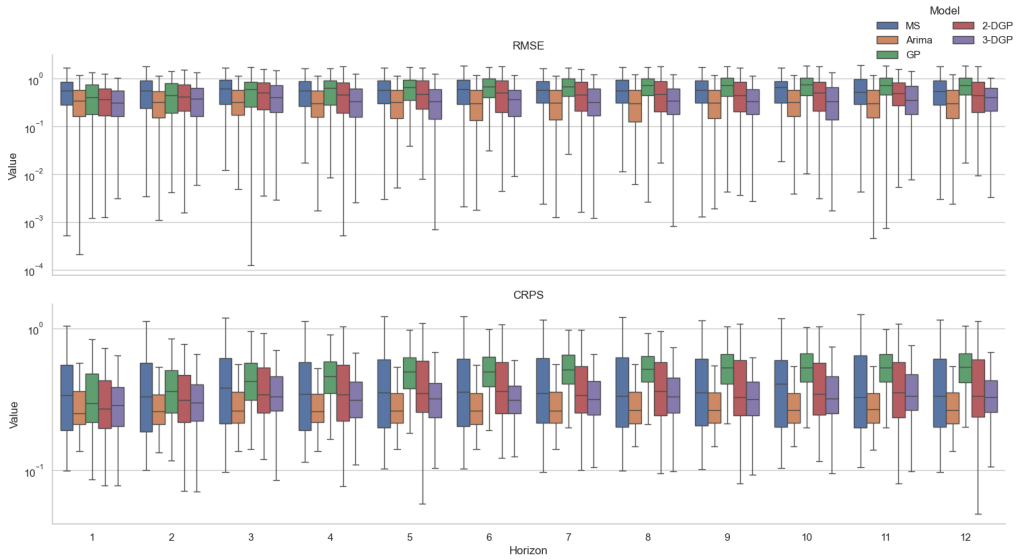


Figure 5.: Boxplot comparison of model performance across forecast horizons (1 to 12 steps ahead) for four architectures: Markov-Switching (MS), standard Gaussian Process (GP), two-layer Deep Gaussian Process (2-DGP), and three-layer Deep Gaussian Process (3-DGP). The top panel reports Root Mean Squared Error (RMSE), and the bottom panel reports Continuous Ranked Probability Score (CRPS). Lower values indicate better performance.

Figure **??** provides a comparative visualization of model performance across multiple forecast horizons, utilizing RMSE and CRPS. Each boxplot summarizes the distribution of forecasting errors at each horizon step (from 1 to 12) for the five models: the Markov-Switching (MS) model, the ARIMA model, the Standard Gaussian Process (GP), the Two-layer DGP (2-DGP), and the Three-layer DGP (3-DGP).

In the upper panel, we observe that both the three-layer DGP and ARIMA models consistently achieve the lowest median RMSE values across all forecast horizons. ARIMA performs particularly well in the short to mid-range horizons (1–6), often outperforming all other models, including the 3-DGP. At longer horizons (7–12), the three-layer DGP tends to slightly edge out ARIMA, suggesting its strength in capturing long-term nonlinear dynamics. Notably, both models also display tight interquartile ranges and relatively few outliers, reflecting high precision and stability. The two-layer DGP ranks close behind, though with somewhat greater variance and slightly elevated errors, indicating that while additional depth helps, a two-layer configuration may not be sufficient to fully model the complexity of the series. In contrast, the standard GP and Markov-Switching models exhibit higher RMSE values and greater dispersion, especially as the forecast horizon increases. These results suggest that both GP and MS struggle to adapt to evolving temporal structures, with GP limited by its stationarity assumptions and MS constrained by its rigid regime-switching framework.

The lower panel reports CRPS, which assesses the models' ability to produce well-calibrated probabilistic forecasts. Here, ARIMA consistently achieves the lowest median CRPS values across all horizons. The three-layer DGP performs almost as well, trailing ARIMA only slightly and maintaining low CRPS with tight spreads, reinforcing its strength in both accuracy and calibration. The two-layer DGP shows moderate performance, often ranking third, but with greater variance and less consistent sharp-

ness. Meanwhile, the standard GP and MS models again exhibit higher and more dispersed CRPS values, indicating poorer calibration and less reliable predictive distributions. Overall, these results reveal that ARIMA and the three-layer DGP are the most effective models, excelling in both point and probabilistic forecasting, while the remaining models offer more limited predictive performance.

In general, the results in Figure **??** reinforce the advantages of incorporating depth and hierarchical latent structure in Gaussian Process models. The strong performance of the three-layer DGP across both point prediction and uncertainty quantification underscores its suitability for complex forecasting tasks characterized by nonstationarity, structural breaks, and long-range dependencies, conditions commonly observed in macroeconomic time series.

These findings offer compelling empirical support for the use of deeper architectures in probabilistic forecasting frameworks. At the same time, ARIMA's consistently strong results highlight its rare combination of simplicity and robustness, delivering reliable and well-calibrated forecasts without the computational and modeling complexity associated with deep learning approaches.

4. Conclusion

This study assessed the performance of Gaussian Process models of varying depth (standard GP, two-layer DGP, and three-layer DGP) in modeling and forecasting U.S. real GNP. A key limitation of the standard GP lies in its reliance on stationary kernels, which assume constant statistical properties over time. While this assumption allows for smooth interpolation within local regions of the training data, it restricts the model's ability to accommodate structural breaks, evolving dynamics, and regime-dependent behavior, common in macroeconomic time series.

Empirical results confirm that the standard GP achieves the lowest RMSE during in-sample evaluation, reflecting a high degree of fit to the training data. However, this performance does not generalize well out-of-sample. The standard GP exhibits poor adaptability in long-horizon forecasts, with narrow credible intervals and frequent miscoverage during periods of transition. This overconfidence stems from its restrictive stationarity assumption, which undermines its flexibility in dynamic environments.

By contrast, deeper architectures such as the three-layer DGP demonstrate substantial gains in both predictive accuracy and uncertainty quantification. The hierarchical composition of latent layers enables the model to capture nonstationary behavior, complex temporal dependencies, and smooth structural shifts without requiring explicit specification of regimes. The three-layer DGP consistently outperforms shallower alternatives across forecast horizons, delivering well-calibrated posterior distributions and robust out-of-sample performance. While the two-layer DGP offers moderate improvements over the standard GP, its performance is less consistent and more sensitive to forecast horizon, suggesting that additional depth is critical in highly nonlinear settings.

Importantly, the results also highlight the strong performance of the ARIMA model. Despite its simplicity, ARIMA proves to be a competitive benchmark, often matching or even surpassing the deep models, especially in probabilistic calibration (CRPS). Compared to the Markov-Switching model, which captures regime-dependent dynamics via discrete state transitions, the three-layer DGP achieves better performance while offering a more flexible and continuous representation of structural change. Whereas the MS model requires pre-specifying the number of regimes and assumes

abrupt transitions, the DGP framework learns gradual, smooth variations directly from the data, enabling more nuanced generalization across time.

In summary, while the standard GP may suffice in stable environments, it fails to address the complexities inherent in real-world macroeconomic forecasting. Deeper DGP architectures, particularly the three-layer variant, provide a principled Bayesian approach to modeling nonstationarity, hidden structure, and forecast uncertainty.

A natural extension of this work is to incorporate additional variables within a Vector Autoregression (VAR) framework. Given the importance of interpretability in economics, a key objective moving forward is to develop a method that not only captures nonlinear dynamics but also enables the identification of regimes and the calculation of regime probabilities in a transparent and interpretable manner.

References

Beveridge, S. and Nelson, C. R. (1981). A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the 'business cycle'. *Journal of Monetary Economics*, 7:151–174.

Bie, S., Diebold, F. X., He, J., and Li, J. (2024). Machine learning and the yield curve: Tree-based macroeconomic regime switching. Working paper, August 2024.

Binois, M., Gramacy, R. B., and Ludkovski, M. (2018). Practical heteroscedastic gaussian process modeling for large simulation experiments. *Journal of Computational and Graphical Statistics*, 27(4):808–821.

Booth, A. S. (2024). *deepgp: Bayesian Deep Gaussian Processes using MCMC*. R package version 1.1.3.

Cai, J. (1994). A markov model of switching-regime arch. *Journal of Business & Economic Statistics*, 12(3):309–316.

Campbell, J. Y. and Mankiw, N. G. (1987). Permanent and transitory components in macroeconomic fluctuations. *American Economic Review Papers and Proceedings*, 77:111–117.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86(2):221–241.

Cosslett, S. R. and Lee, L.-F. (1985). Serial correlation in latent discrete variable models. *Journal of Econometrics*, 27(1):79–97.

Dacco, R. and Satchell, S. (1999). Why do regime-switching models forecast so badly? *Journal of Forecasting*, 18(1):1–16.

Damianou, A. and Lawrence, N. D. (2013). Deep gaussian processes. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 207–215. PMLR.

Diebold, F. X., Lee, J., and Weinbach, G. C. (1994). Regime switching with time-varying transition probabilities. *Discussion Paper Series, Institute for Empirical Macroeconomics, Federal Reserve Bank of Minneapolis*, (83).

Filardo, A. J. (1994). Business-cycle phases and their transitional dynamics. *Journal of Business & Economic Statistics*, 12(3):299–308.

Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2011). Bayesian nonparametric methods for learning markov switching processes. *IEEE Signal Processing Magazine*, 27(6):43–54.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Goldfeld, S. M. and Quandt, R. E. (1973). A markov model for switching regressions. *Journal of Econometrics*, 1(1):3–15.

Gramacy, R. B. and Apley, D. W. (2015). Local gaussian process approximation for large computer experiments. *Journal of Computational and Graphical Statistics*, 24(2):561–578.

Gramacy, R. B. and Lee, H. K. H. (2008). Bayesian treed gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103(483):1119–1130.

Gray, S. F. (1996). Modeling the conditional distribution of interest rates as a regime-switching process. *Journal of Financial Economics*, 42(1):27–62.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384.

Hamilton, J. D. and Susmel, R. (1994). Autoregressive conditional heteroskedasticity and changes in regime. *Journal of Econometrics*, 64(1–2):307–333.

Hauzenberger, N., Huber, F., Marcellino, M., and Petz, N. (2024). Gaussian process vector autoregressions and macroeconomic uncertainty. *Journal of Business & Economic Statistics*, 43(1):27–43.

Heinonen, M., Mannerström, H., Rousu, J., Kaski, S., and Lähdesmäki, H. (2016). Non-stationary gaussian process regression with hamiltonian monte carlo. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol-

ume 51 of *Proceedings of Machine Learning Research*, pages 732–740. PMLR.

Higdon, D., Swall, J., and Kern, J. (1999). Non-stationary spatial modeling. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 6*, volume 6, pages 761–768. Oxford University Press.

Katzfuss, M. (2013). Bayesian nonstationary spatial modeling for very large datasets. *Environmetrics*, 24(3):189–200.

Kim, C.-J. and Nelson, C. R. (1998). *State-Space Models with Regime Switching: Classical and Gibbs-Sampling Approaches with Applications*. MIT Press, Cambridge, MA.

Kim, C.-J., Piger, J., and Startz, R. (2003). Mixed frequency markov-switching vector autoregression. *Studies in Nonlinear Dynamics & Econometrics*, 7(1):1–24.

Kim, H.-M., Mallick, B. K., and Holmes, C. (2005). Analyzing nonstationary spatial data using piecewise gaussian processes. *Journal of the American Statistical Association*, 100(470):653–668.

Krolzig, H.-M. (1997). *Markov-Switching Vector Autoregressions: Modelling, Statistical Inference, and Application to Business Cycle Analysis*, volume 454 of *Lecture Notes in Economics and Mathematical Systems*. Springer, Berlin.

Liu, R. H. and Nguyen, D. (2015). A tree approach to options pricing under regime-switching jump diffusion models. *International Journal of Computer Mathematics*, 92(12):2575–2595.

Neftci, S. N. (1984). Are economic time series asymmetric over the business cycle? *Journal of Political Economy*, 92:307–328.

Nelson, C. R. and Plosser, C. I. (1982). Trends and random walks in macroeconomic time series: Some evidence and implications. *Journal of Monetary Economics*, 10:139–162.

Paciorek, C. J. and Schervish, M. J. (2003). Nonstationary covariance functions for gaussian process regression. In *Proceedings of the 16th International Conference on Neural Information Processing Systems (NIPS)*, pages 273–280, Cambridge, MA, USA. MIT Press.

Potter, S. M. (1999). Nonlinear time series modeling: An introduction. *Journal of Economic Surveys*, 13(5):505–528.

Rasmussen, C. E. and Ghahramani, Z. (2001). Infinite mixtures of gaussian process experts. In *Advances in Neural Information Processing Systems*, volume 14.

Sampson, P. D. and Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87(417):108–119.

Sauer, A., Gramacy, R. B., and Higdon, D. (2023). Active learning for deep gaussian process surrogates. *Technometrics*, 65(1):4–18.

Schmidt, A. M. and O'Hagan, A. (2003). Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):743–758.

Sichel, D. E. (1987). Business cycle asymmetry: A deeper look. Mimeographed, Princeton University.

Tong, H. (1983). *Threshold Models in Non-linear Time Series Analysis*, volume 21 of *Lecture Notes in Statistics*. Springer, New York.

Wu, H., Mardt, A., Pasquali, L., and Noé, F. (2018). Deep generative markov state models. *Advances in Neural Information Processing Systems*, 31.

Appendix A. Forecast Evaluation Metrics

To assess forecast performance, we employ two complementary metrics: the Root Mean Squared Error (RMSE) and the Continuous Ranked Probability Score (CRPS).

The RMSE is a widely used metric, that quantifies accuracy of point estimates. It is defined as:

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

where y_i denotes the observed values and $\hat{y}_i$ the corresponding forecasts. Lower RMSE values indicate higher point forecast accuracy.

In contrast, the CRPS evaluates the quality of a full predictive distribution rather than a single point estimate. Given a cumulative distribution function F associated with a forecast and an observed value y, CRPS is defined as:

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} \left(F(x) - \mathbf{1}\{x \geq y\}\right)^2 dx$$

Here, $\mathbf{1}\{x \geq y\}$ is an indicator function that steps from 0 to 1 at the observed value y. The CRPS can be interpreted as the squared distance between the predictive CDF and a step function representing the actual outcome. It rewards forecasts that assign high probability mass near the true value, thus capturing both calibration and sharpness of the distribution. As with RMSE, lower CRPS values denote better forecasting performance. The CRPS belongs to a broader class of proper scoring rules, which are designed to evaluate the quality of probabilistic forecasts as we can see in Gneiting and Raftery (2007).

Figure A1 provides a visual explanation of the CRPS computation. The area between the forecast cumulative distribution function and the step function defined by the actual observation represents the score. The smaller this area, the more accurate and better calibrated the forecast.

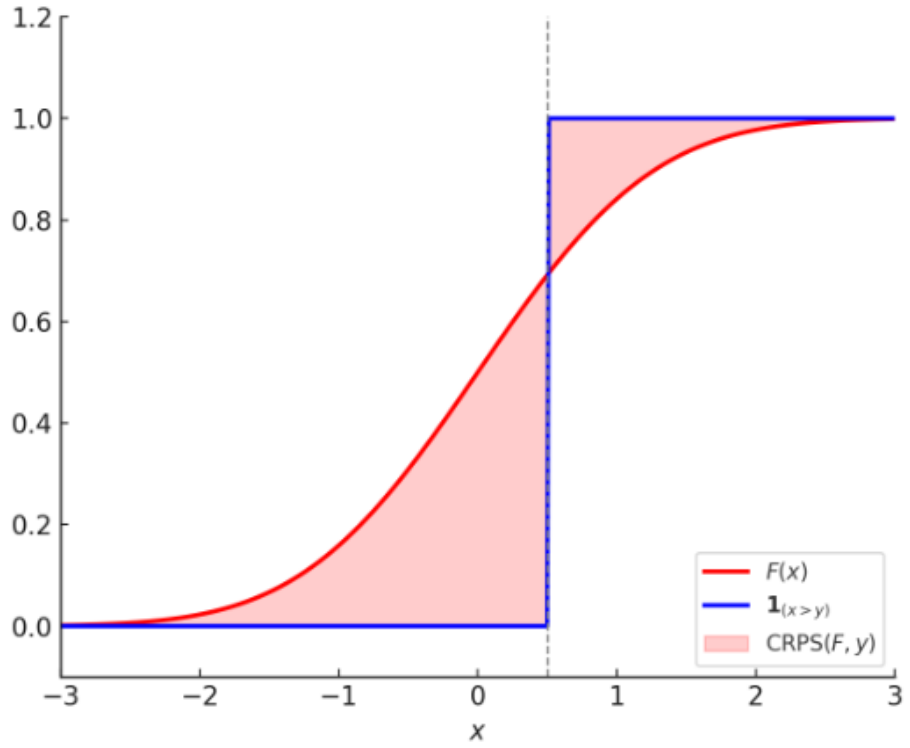


Figure A1.: Illustration of CRPS computation. The red curve represents the predictive cumulative distribution function $F(x)$, while the blue step function corresponds to the observation indicator $\mathbf{1}_{(x>y)}$, where y is the realized value. The shaded red area quantifies the Continuous Ranked Probability Score (CRPS), measuring the squared distance between the predicted distribution and the observed outcome. Lower CRPS values indicate better-calibrated probabilistic forecasts.