Forecasting with Deep Pooled Panel Neural Networks

Forecasting with Deep Pooled Panel Neural Networks*

Ilias Chronopoulos[a], Katerina Chrysikou[b], George Kapetanios[c], James Mitchell[d], and Aristeidis Raftapostolos[c]

[a]Essex Business School, University of Essex; [b]School of Business, University of Leicester; [c]King's Business School, King's College London; [d]Federal Reserve Bank of Cleveland

June 9, 2026

Abstract

In this paper, we propose a deep pooled estimator, motivated by the universal approximation property of neural networks, to capture nonlinear relationships between predictors and targets when modeling and forecasting with panel data. The approach is flexible, accommodating different penalty functions and potentially high-dimensional predictors. It allows for nonlinear cross-sectional dependencies. To evidence the utility of the proposed estimator when forecasting, we apply it in two different applications. First, we forecast the progression of new COVID-19 cases across G7 countries. Second, we forecast inflation in the G7. In both applications, our method delivers significant forecasting gains over both linear panel and nonlinear time-series (unit-specific) models that do not pool data across countries. These results highlight the importance when forecasting of pooling cross-country information via a flexible nonlinear model. Examining partial derivatives from our model provides interpretable insights: school closures and workplace restrictions show declining effectiveness as COVID-19 immunity strengthened, while the inflation-unemployment relationship proves highly unstable across both countries and time periods, particularly during the post-pandemic inflation surge.

JEL codes: C33, C45.

Keywords: Machine learning, neural networks, panel data, nonlinearity, forecasting, COVID-19, inflation, Phillips curve.

*We thank the editor and two anonymous referees for helpful comments. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Cleveland or the Federal Reserve System. The authors confirm that the data supporting the findings of this study are available within the article's online supplementary materials.

Email address for correspondence: profjamesmitchell@gmail.com

1 Introduction

This paper considers forecasting with nonlinear panel data models. We propose the use of a novel machine learning (ML) panel data estimator based on neural networks. In the last decade, ML methods have been incorporated, in various forms, across the natural, social, medical, and economic sciences. There are two main reasons for this. First, ML methods and specifically neural networks, our focus, have been found to forecast well, specifically with high-dimensional data sets. Second, they have great capacity to uncover potentially unknown and highly complicated, possibly nonlinear, relationships in the data.

Studies have shown that feed-forward neural networks can approximate any continuous function of several real variables arbitrarily well; see, for example, Hornik (1991), Hornik et al. (1989), Gallant and White (1992), and Park and Sandberg (1991). Other nonparametric approaches, for example, splines, wavelets, the Fourier basis, as well as simple polynomial approximations, have the universal approximation property, based on the Stone–Weierstrass theorem. However, it has been convincingly argued that neural networks outperform them in prediction (see, for example, Kapetanios and Blake (2010)). More recent work by Liang and Srikant (2016) and Yarotsky (2017, 2018) considers feed-forward neural networks as approximations for complex functions that accommodate multiple layers, provided sufficiently many hidden neurons and layers are available. Other works, like Bartlett et al. (2019), provide the theoretical framework for neural network estimation; while Schmidt-Hieber (2020) focuses on the adaptation property of neural networks, showing that they strictly improve on classical methods. If the unknown target function is a composition of simpler functions, then the composition-based deep net estimator is superior to estimators that do not use compositions. Lastly, recent work of Farrell et al. (2021), building on the work of Yarotsky (2017) and Bartlett et al. (2019), studies deep neural networks and considers their use for semi-parametric inference.

In this paper we focus on forecasting with nonlinear panel data models, where the source of nonlinearity lies in the conditional mean. Our contribution is to propose an ML estimator of the conditional mean, $E(y_{it}|\boldsymbol{x}_{it})$, based on (deep) neural networks, to explore the effect of potential nonlinearities in panel data models. These are introduced by allowing the conditional mean to have a panel/common nonlinear component. We explore the utility of the new model via two forecasting applications – one forecasting COVID-19 across the G7, the other inflation in the G7 countries.

Compared to existing studies forecasting COVID-19 infections and inflation, which we now review in turn, the value-added of our model is to allow for both nonlinearities and commonalities across the units in the panel. A wide set of papers have used neural networks to forecast the COVID-19 pandemic. But they do not directly exploit information across countries (or regions). For example, Ahmadini et al. (2025) propose a neural network framework that combines epidemiological, mobility, vaccination, and environmental data to predict case trends. Their model improves forecast accuracy relative to linear models. Tamang et al. (2020) use neural network–based curve fitting to predict cases and deaths across countries, calibrating to China and South Korea as reference epidemics. Jiang et al. (2024) employ an

artificial neural network to relate reported infections, testing data, and true infection counts. Namasudra et al. (2023) propose a nonlinear autoregressive neural network time-series model to forecast confirmed, recovered, and death cases. They find that their specification outforecasts alternatives. While these studies demonstrate the versatility of neural networks, they do not exploit panel data. Our contribution is to develop a model to show that pooling across countries, in a nonlinear framework, yields substantial gains when forecasting COVID-19 infections.

Other papers forecasting COVID-19 infections, notably Liu et al. (2021), instead pool data across countries but maintain a linear modeling structure. Like ours, their dynamic panel data model is reduced-form. This contrasts Susceptible-Infected-Recovered (SIR) models, widely used in epidemiology, which are "structural." To enable causal inference, SIR models impose strong assumptions. Our reduced-form focus lets the data "speak" and enables us, like Liu et al. (2021), to pool data across countries/regions. We find that pooling across countries improves predictive accuracy relative to both deep time-series models (which ignore cross-sectional dependencies) and linear panel VARs. The nonlinear features of the new model let it capture pandemic features, such as the sharp run-ups in infection rates in early waves of the pandemic. While our models are reduced-form, we do undertake out-of-sample Granger causality tests to assess the impact of non-pharmaceutical interventions.

Turning to inflation, in an important paper Medeiros et al. (2021) demonstrate that, despite earlier skepticism, ML models using a large set of covariates consistently outpredict traditional benchmarks, with the random forest model showing the strongest performance. Previous work, such as Ülke et al. (2018), found mixed evidence on the relative utility of ML and time series models when forecasting inflation. As in the COVID-19 application, our contribution is again to build on earlier studies to consider whether additional forecast gains accrue when we extend the modeling framework to the panel data setting. We find that, except in Japan which distinguished itself by not experiencing high inflation post-pandemic, pooling information across G7 countries – using our deep pooled model – delivers forecast accuracy gains.

While neural networks have great capacity to approximate complicated nonlinear functions and forecast well, they are frequently criticized as being non-interpretable – of being a "black box." This is because they do not offer simple summaries of relationships in the data. Recently, there have been a number of papers that try to make ML output (more) interpretable; see, for example, Athey and Imbens (2017), Wager and Athey (2018), Belloni et al. (2014), Joseph (2019), Chronopoulos et al. (2024), and Kapetanios and Kempf (2022).

In this paper we explore how partial derivatives, calculated from the output of our proposed neural network, can help a modeler understand what is driving their forecasts. We use these derivatives in our two applications to examine the effectiveness of containment policies at lowering new COVID-19 cases and to examine the Phillips curve-type relationship between inflation and unemployment. We find that some, but not all, containment policies were effective at lowering new COVID-19 cases. These policies tended to be more effective two to three weeks after the policy change. There is also considerable heterogeneity across countries in the effectiveness of these policies. Turning to the inflation-unemployment link, we also see a lot of variation across countries and time, consistent with an unstable Phillips curve.

The remainder of the paper proceeds as follows. In Section 2, we introduce the deep pooled panel estimator. In Section 3, we discuss both methodological and implementation aspects of the proposed methodology. In Section 4, we test the new model via the two G7 forecasting applications. We first forecast new COVID-19 cases and assess the effectiveness of containment policies. Then we forecast inflation. Both applications demonstrate the value when forecasting of pooling information across countries within a flexible nonlinear framework. We also find evidence of the *double descent phenomenon*, whereby complex models can perform well without the need for explicit regularization, see, for example, Hastie et al. (2022) and Kelly et al. (2022) (and Remark 2, below). Section 5 concludes. We relegate to an online appendix additional forecasting results, data summaries, and further discussion of the forecast evaluation tests.

2 Setup

Let y_{it} be the observation for the i^{th} cross-sectional unit at time t generated by the following pooled panel data model:

$$E\left(y_{it}|\boldsymbol{x}_{it}\right)=\widetilde{h}\left(\boldsymbol{x}_{it}\right), \quad i=1,\ldots,N, \quad t=1,\ldots,T, \tag{1}$$

where $\{\boldsymbol{x}_{it}\}=\{(x_{t,1},\ldots,x_{t,p})'\}$ is a p-dimensional vector of regressors, belonging to unit i, and $\widetilde{h}(\cdot)$ is the unknown nonlinear function that will be approximated with neural networks. Throughout, we abstract from unconditional mean considerations, by assuming that $E(y_{it})=0$. This is achieved via simple unit-by-unit demeaning of the dependent variable.

To be specific, we consider the following population model that generates the data:

$$y_{it}=\widetilde{h}\left(\boldsymbol{x}_{it}\right)+\varepsilon_{it}, \quad i=1,\ldots,N, \quad t=1,\ldots,T, \tag{2}$$

where ε_{it} is an error term. Standard econometric analysis assumes a linear relationship, where $\widetilde{h}\left(\boldsymbol{x}_{it}\right)=\boldsymbol{x}_{it}'\boldsymbol{\beta}$, which is readily estimable by OLS. However, this assumption can be overly restrictive. Instead, we will proceed to model the nonlinear function, stated in (1), using feed-forward neural networks. This allows us to consider $\widetilde{h}\left(\boldsymbol{x}_{it}\right)$ to be some unknown, (potentially highly) nonlinear function. Specifically, we assume that there exists a neural network with an overall function $g\left(\boldsymbol{x}_{it};\boldsymbol{\theta}\right)$, to be defined below, that can approximate $\widetilde{h}\left(\boldsymbol{x}_{it}\right)$ well using standard universal approximation theorems, that will also be discussed below. Before we discuss the implementation of our neural network pooled panel methodology, we explain how neural networks approximate $\widetilde{h}\left(\boldsymbol{x}_{it}\right)$.

2.1 Neural Networks

We focus on the construction of the *feed-forward neural network* functional parameterization, $g\left(\boldsymbol{x}_{it};\boldsymbol{\theta}\right)$, used to approximate $\widetilde{h}\left(\boldsymbol{x}_{it}\right)$. The feed-forward architecture consists of: an input layer, where the covariates are introduced given an initial set of weights to the inner (hidden)

part of the network; the hidden layers, where a number of computational nodes/neurons are collected in each hidden layer and nonlinear transformations on the (weighted) covariates occur; and the output layer that gives the final predictions and a choice for the activation function $\sigma(x) : \mathbb{R} \to \mathbb{R}$ that is applied element-wise. The architecture is feed-forward, since in each of the hidden layers there exist several interconnected neurons that allow information to flow from one layer to the other, but only in one direction. The connections between layers correspond to weights.

We use L to define the total number of hidden layers and $M^{(l)}$, $l = 1, \ldots, L$, to define the total number of neurons at the l^{th} layer. L and $M^{(l)}$ are measures for the depth and width of the neural network, respectively. We use the rectified linear unit (ReLU) activation function, $\sigma_l(\boldsymbol{X}_t) := \max(\boldsymbol{X}_t, 0)$, where $\boldsymbol{X}_t$ is a $N \times p$ matrix of characteristics for $t = 1, \ldots, T$; $l = 1, \ldots, L-1$ and a linear activation function for $l = L$. The activation functions are applied elementwise. To explain the exact computation of the outcome of the *feed-forward neural network*, we focus on the architecture of the pooled-type estimator $g(\boldsymbol{x}_{it}; \boldsymbol{\theta})$. We assume that the widths (the number of neurons), $M^{(l)}$, and depth (the number of hidden layers), L, of the network are constant positive numbers.

Each of the neurons undergoes a computation similar to the linear combination received in each hidden layer l: $\boldsymbol{g}^{(l)} = \sigma_l(\boldsymbol{g}^{(l-1)}\boldsymbol{W}^{(l)'} + \boldsymbol{b}^{(l)'})$, while the final output of the network is $\boldsymbol{g}^{(L)} = \boldsymbol{g}^{(L-1)}\boldsymbol{W}^{(L)'} + \boldsymbol{b}^{(L)'}$ and $\boldsymbol{g}^{(0)} = \boldsymbol{X}_t$. We can then define for some $t = 1, \ldots, T$, $g(\boldsymbol{X}_t; \boldsymbol{\theta})$ as:

$$g(\boldsymbol{X}_t; \boldsymbol{\theta}) = \left(\sigma_L \cdots \sigma_2 \left(\sigma_1 \left(\boldsymbol{X}_t \boldsymbol{W}^{(1)'} + \boldsymbol{b}^{(1)'}\right) \boldsymbol{W}^{(2)'} + \boldsymbol{b}^{(2)'}\right) \cdots\right) \boldsymbol{W}^{(L)'} + \boldsymbol{b}^{(L)'}, \tag{3}$$

where $\boldsymbol{W}^{(l)}$ is a $M^{(l)} \times M^{(l-1)}$ matrix of weights, $\boldsymbol{b}^{(l)}$ is a $M^{(l)} \times N$ matrix of biases at layer l, with $\boldsymbol{b}^{(1)} = \boldsymbol{0}$. Notice that at $l = 1$, the dimensions of $\boldsymbol{W}^{(1)}$ are $M^{(1)} \times p$ and of $\boldsymbol{b}^{(1)}$ are $M^{(1)} \times N$. At the final layer, that is, at $l = L$, the dimensions of $\boldsymbol{W}^{(L)}$ are $1 \times M^{(L-1)}$ and for $\boldsymbol{b}^{(L)}$ are $1 \times N$.

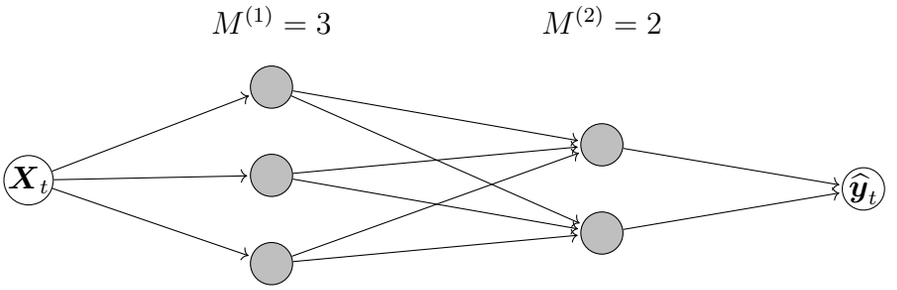
As an illustration, we present an example of a *feed-forward neural network*, based on (3), in Figure 1. The neural network in Figure 1 consists of a matrix input $\boldsymbol{X}_t \in \mathbb{R}^{N \times p}$ and one output vector $\widehat{\boldsymbol{y}}_t \in \mathbb{R}^{N \times 1}$. Between the input and output layers, there are $M = 5$ hidden computational nodes/neurons, in total. The neurons are connected directly forming an acyclic graph that specifies a fixed architecture. Note that the illustration in Figure 1 can correspond to a nonlinear pooled-type estimator of $g(\boldsymbol{X}_t; \boldsymbol{\theta})$.

Note that, throughout the paper, we use $\boldsymbol{\theta}$ to denote a stacked vector containing all ancillary trainable parameters affiliated with the network estimation, as defined below:

$$\boldsymbol{\theta} = \left(\text{vec}\left(\boldsymbol{W}^{(1)'}\right), \ldots, \text{vec}\left(\boldsymbol{W}^{(L)'}\right), \text{vec}\left(\boldsymbol{b}^{(1)'}\right), \text{vec}\left(\boldsymbol{b}^{(2)'}\right), \ldots, \boldsymbol{b}^{(L)'}\right)'. \tag{4}$$

We define the overall number of parameters as $d = |\boldsymbol{\theta}|$. The optimization of the neural network proceeds in a forward fashion (from the input layer, that is, $l = 1$, to the output $l = L$) and layer-by-layer through an optimizer, for example, a version of stochastic gradient

Figure 1: Illustration of a *feed-forward neural network*, based on (3) with $L = 2$ hidden layers, a total of $M = 5$ neurons, depicted by gray circles, across the two hidden layers ($M^{(1)} = 3$ and $M^{(2)} = 2$), yielding a total of $W = 11$ connections, illustrated by solid black arrows. The neural network receives an input matrix $\boldsymbol{X}_t \in \mathbb{R}^{N\times p}$, with vector output $\widehat{\boldsymbol{y}}_t \in \mathbb{R}^{N\times 1}$.



descent (SGD), where the gradients of the parameters $(\boldsymbol{W}^{(l)}, \boldsymbol{b}^{(l)})$ are calculated through back-propagation (using the chain-rule) to train the network. We relegate to online appendix E discussion of adaptive moment estimation (ADAM) for SGD.

Remark 1 *The exact (composition) structure described in* (3) *holds for a subclass of* feed-forward neural networks*, specifically that one that refers to fully connected layers (the one being consecutive to the other) but has no other connections. Each layer has a number of hidden units that are of the same order of magnitude. This architecture is the most commonly used in empirical research, and is often referred to as a* Multi-layer Perceptron *(MLP). Furthermore, the exact structure in* (3) *does not hold generally for any* feed-forward neural network.

The specific choice of the network architecture is crucial and affects the complexity and the approximating power of $g(\boldsymbol{x}_{it}, \boldsymbol{\theta})$ in (3).

Our analysis utilizes standard results from universal approximation theorems, that are applicable in feed-forward neural networks and can be used with panel data. Specifically, according to various universal approximation theorems (see, for example, the theoretical results in Hornik (1991), Hornik et al. (1989), Gallant and White (1992), Kapetanios and Blake (2010), Liang and Srikant (2016), Hanin (2019), and Yarotsky (2017, 2018)), $g(\boldsymbol{x}_{it}, \theta)$ can approximate any continuous function $\tilde{h}(\boldsymbol{x}_{it})$ arbitrarily well, such that, for any $\epsilon > 0$:

$$\sup_{\mathcal{X}} \left| g(\boldsymbol{x}_{it}, \theta) - \tilde{h}(\boldsymbol{x}_{it}) \right| < \epsilon, \tag{5}$$

where we denote by $\mathcal{X} \subset \mathbb{R}^p$ the support of the regressors $\boldsymbol{x}_{it}$, i.e., the set of all possible realizations of $\boldsymbol{x}_{it}$ across units $i = 1, \ldots, N$ and time periods $t = 1, \ldots, T$. The theorem requires $\mathcal{X}$ to be compact (closed and bounded). This assumption is standard in approximation theory and is satisfied in empirical practice by either normalization or standardization of the regressors.

More generally, the compactness assumption can be relaxed. If regressors are unbounded but have finite moments (e.g., $E\|\boldsymbol{x}_{it}\|^k < \infty$ for some $k > 0$), then approximation results continue to hold on large bounded subsets of $\mathbb{R}^p$ with high probability, with approximation error controlled by the tails of the distribution. This relaxation is particularly relevant for economic data, which are often theoretically unbounded but typically have finite second or higher moments (see Yarotsky (2017, 2018) and Schmidt-Hieber (2020)).

The (ϵ)-approximation in (5) can be seen as a sieve-type non-parametric estimation bound of the following form:

$$\widetilde{h}\left(\boldsymbol{x}_{it}\right) = g(\boldsymbol{x}_{it}, \boldsymbol{\theta}) + O(\epsilon),$$

where ϵ can be made arbitrarily small by increasing the complexity of the neural network. It is important to note that the increase in complexity can occur either by letting $L \to \infty$, which stands for deep learning, or by letting $M^{(l)} \to \infty$. While, asymptotically both ways deliver the same results (see, e.g., Farrell et al. (2021) and references therein), the approximation error has been shown to decline exponentially with L (see, for example, Babii et al. (2020)) but only polynomially with $M^{(l)}$, providing some evidence for the prevalent use of deep learning. Notice that there also exists an alternative approximation theory for sparse deep learning; see, for example, the work of Schmidt-Hieber (2020). In addition, while the universal approximation theorem(s) imply that a large feed-forward neural network will be able to approximate/represent the nonlinear function, they do not guarantee that the neural network will be able to learn it. The main reasons for failures lie in overfitting and optimization algorithms failing to find the correct hyperparameters in the neural network's architecture. To avoid these problems we both use cross-validation to select all relevant parameters and hyperparameters and reguralize. We discuss these concepts further in Section 3.2.

2.2 Nonlinear Deep Pooled Panel Model

Our nonlinear pooled panel methodology involves the approximation of the nonlinear function $\widetilde{h}\left(\boldsymbol{x}_{it}\right)$ with neural network functional parameterizations, given by $g\left(\boldsymbol{x}_{\boldsymbol{it}}; \boldsymbol{\theta}\right)$. Abstracting from the approximation error[1] in (5) allows us to re-write the model as:

$$y_{it} = g(\boldsymbol{x}_{it}; \boldsymbol{\theta}) + \varepsilon_{it}, \quad t = 1, \ldots, T, \quad i = 1, \ldots, N, \tag{6}$$

where ε_{it} is an error term and $\boldsymbol{\theta}$ denote the values of the learnable parameters that need to be estimated in order to find the best feasible approximation to $\widetilde{h}\left(\boldsymbol{x}_{it}\right)$, in a sense to be defined below. Notice that, in (6), the functional form is known up to the parameter vector $\boldsymbol{\theta}$, which is a vector of ancillary parameters, such as network weights and biases and, as discussed in the previous section, these boil down to the choice of the researcher.

[1] We note that this approximation error can be non-negligible if the specification of the neural network is not flexible enough.

Under the assumption of linearity, (6) is simplified to:

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}, \tag{7}$$

which is readily estimable by OLS. It is useful to draw some parallels between (6) and (7). We note that most multi-layer neural network architectures have a final linear layer given by:

$$g\left(\boldsymbol{x}_{it};\boldsymbol{\theta}\right) = \boldsymbol{\theta}_L'\boldsymbol{f}\left(\boldsymbol{x}_{it}\right),$$

where $\boldsymbol{f}$ is a vector of known functions that form part of the neural network architecture and L denotes the number of network layers. Then, it follows that we have a linear representation, in $\boldsymbol{f}$, of the following form:

$$y_{it} = \boldsymbol{\theta}_L'\boldsymbol{f}\left(\boldsymbol{x}_{it}\right) + \varepsilon_{it}, \tag{8}$$

which is reminiscent of (7) and thus provides a clear rationale for our nonlinear extension. Furthermore, it provides a rationale for thinking that $\boldsymbol{\theta}$ plays a similar role to the pooled regression coefficients, $\boldsymbol{\beta}$, of the linear model. The model above encompasses a variety of nonlinear specifications. It is also worth emphasizing that the dimension of the regressor vector could be very large. So, it is conceivable that each $\boldsymbol{x}_{it}$ contains regressors from other cross-sectional units, allowing for complex nonlinear interactions across units. In the limit, each unit could have $(\boldsymbol{x}_{1t}, \ldots, \boldsymbol{x}_{Nt})$ as the regressor vector.

To obtain an estimate for the pooled-type nonlinear estimator, $g(\boldsymbol{x}_{it};\widehat{\boldsymbol{\theta}})$, we need an estimate for $\widehat{\boldsymbol{\theta}}$. This requires us to select both the architecture of the neural network, that was discussed above, and a loss function. We discuss the minimization problem and the choice of the loss function next.

3 Implementation Considerations

In this section we provide further details on the implementation of the proposed nonlinear estimator. First, we illustrate how regularization can be applied in the context of the proposed estimators. Then we discuss both the cross-validation exercise used to select the different parameters and hyperparameters of the corresponding network and the optimization algorithm.

3.1 Implementation and Regularization

We focus our discussion on the following panel estimator, $g(\boldsymbol{x}_{it};\widehat{\boldsymbol{\theta}})$, obtained from the optimization of the following problem:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min}\ \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - g\left(\boldsymbol{x}_{it};\boldsymbol{\theta}\right)\right)^2, \tag{9}$$

where we use the mean squared error (MSE) loss function.

This nonlinear panel estimator, and generally neural network estimators, confer many important advantages over traditional panel models, mainly because of their capacity to approximate highly nonlinear and complicated associations between variables that can improve forecasting performance; see, for example, the discussion in Goodfellow et al. (2016) and Gu et al. (2020, 2021). In order to minimize (9) and obtain a feasible estimate for the panel estimator $g(\boldsymbol{x}_{it}; \widehat{\boldsymbol{\theta}})$, we need to choose the overall architecture of the neural network. Following the discussion in Section 2.1, this reduces to choices for the total number of layers L, total number of neurons $M^{(l)}$, at each $l = 1, \ldots, L$ layers, a loss function, which in this paper is taken to be MSE loss, an updating rule for the weights (learning rate, γ) during optimization, and the optimization algorithm itself, typically taken to be some variant of SGD.

However, neural networks tend to overfit, which can lead to a severe deterioration in their (forecasting i.e., out-of-sample) performance. A common empirical solution is to impose a penalty on the trainable parameters of the neural network, $\boldsymbol{\theta}$. The penalized estimator based on LASSO, $g(\boldsymbol{x}_{it}; \widehat{\boldsymbol{\theta}}_{\ell_1})$, is obtained as the solution to the following minimization problem:

$$\widehat{\boldsymbol{\theta}}_{\ell_1} = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\arg\min} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{it} - g(\boldsymbol{x}_{it}; \boldsymbol{\theta}) \right)^2 + \lambda \left\| \boldsymbol{\theta} \right\|_1 ,$$

where λ is the regularization parameter. Note that while explicit regularization improves empirical solutions of neural networks estimators under low signal-to-noise ratios, its role is not clear theoretically, since there are cases where simpler SGD solutions present similar solutions; see, for example, Zhang et al. (2021). Other commonly used regularization techniques frequently employed empirically to assist in the estimation of neural networks, relate to batch normalization, early stopping, and dropout. We succinctly discuss batch normalization next, given its importance because of the cross-sectional aspect of our estimator. We refer the reader to Gu et al. (2020) for a detailed discussion of early stopping and dropout.

Batch normalization, proposed by Ioffe and Szegedy (2015), is a technique used to control the variability of the covariates across different regions of the network and datasets. It is used to address the issue of internal covariate shift, where inputs of hidden layers may follow different distributions than their counterparts in the validation sample. This is a prevalent issue when fitting, in particular, deep neural networks. Effectively, batch normalization cross-sectionally demeans and standardizes the variance of the batch inputs.

Remark 2 *In this paper, we consider both penalized and non-penalized estimation. While using the latter might seem problematic due to the large number of parameters that needs to be estimated, we find, in our empirical work, that this is not necessarily the case (see online appendix B.4). This is not surprising. Recent work in the statistical and ML literature highlights what is known as the "double descent phenomenon." For linear regressions, this relates to the use of generalized inverses to construct least squares estimators, when the number of variables, p, exceeds the number of observations, T. Such estimators work better either when p is small*

(and standard matrix inversion can be used) or when p is much larger than T. Then the quality of the performance of an estimator is implicitly measured in terms of the "bias-variance trade-off," where an optimal performance resides at the lowest reported bias and variance of the corresponding model (either linear or nonlinear). While it is widely accepted that the "bias-variance trade-off" function resembles a U-shaped curve, it has been observed, for example see Belkin et al. (2019) and Hastie et al. (2022), that beyond the interpolation limit the test loss descends again, hence the name "double-descent." To understand why, note that such estimators implicitly impose penalization by using generalized inverses and so choose the parameter vector with the smallest norm, among all admissible vectors; see Hastie et al. (2022). So, once p is much larger than T, such a selection becomes more consequential, as many more candidate vectors are admissible. This linear effect is also present for neural network estimation given the connection to linear models highlighted in Section 2, as discussed in detail in Hastie et al. (2022) and Kelly et al. (2022).

3.2 Cross-Validation

We use cross-validation (CV) to calibrate all hyperparameters of the neural network and aim to maximize its out-of-sample forecasting performance. This procedure ensures that a suboptimal model is not selected and that the architecture is entirely data-driven.

The CV scheme involves choices on: (i) the total number of layers (L) and neurons (M), (ii) the learning rate (γ) for SGD, (iii) the batch size, (iv) the level of regularization (λ) used for LASSO penalization, (v) the dropout rate, and (vi) the activation functions.

Regarding activation functions, we use ReLU for hidden layers and a linear function for the output layer.[2] The learning rate of the optimizer, γ, is tuned from a discrete grid of values $0.01, 0.001$. The depth and width of the network are tuned using the grids $[1, 3, 5, 10, 15]$ and $[5, 10, 15, 20, 30]$, respectively, allowing the choice between deep or shallow architectures to be fully data-driven. We fix the batch size to 14 and use dropout regularization with a probability of up to 10 percent; cf. Gu et al. (2020).

The regularization parameter, λ, is tuned from the grid $c\sqrt{\log p/NT}$, where $c = [0.001, 0.01, 0.1, 0.5, 1, 5, 10]$. To select the trainable parameters, $\boldsymbol{\theta}$, and tune the hyperparameters, we divide the sample into three disjoint time periods that preserve the temporal ordering of the data: the *training* sub-sample, used to estimate $\boldsymbol{\theta}$ given a specific set of hyperparameters; the *validation* sub-sample, used to tune hyperparameters based on $\widehat{\boldsymbol{\theta}}$ estimated from the training sample[3]; and the *testing* sub-sample, which is truly out-of-sample and used to evaluate forecasting performance. The forecasting exercise is recursive, based on an expanding window. At each window, we perform the train-validation split to estimate parameters and tune hyperparameters. Let T^* denote the total sample size for a given window. The *training* sub-sample consists of $\lfloor 0.8T^* \rfloor$ observations, the validation sub-sample consists of $\lfloor 0.2T^* \rfloor - c$ observations, and the testing sub-sample consists of h observations (7, 14, or 21) depending on the forecast

[2] Recall, $\text{ReLU}(x) = \max(0, x)$.

[3] While $\widehat{\boldsymbol{\theta}}$ is used in hyperparameter tuning, it is only estimated on the training sub-sample.

horizon, h. The constant c ensures that the testing sub-sample always contains h observations, and $\lfloor \cdot \rfloor$ denotes the floor function. The network parameters are estimated using SGD, with the ADAM optimizer to improve convergence and reduce noise in gradient evaluation; cf. Kingma and Ba (2014). CV is repeated across all hyperparameter combinations, and the set minimizing the validation loss is selected. The final evaluation of the network's out-of-sample forecasting ability is performed on the testing sub-sample, without further training or tuning.

3.3 Optimization

The estimation of neural networks is, in general, a computationally cumbersome optimization problem, due to nonlinearities and non-convexities. The most commonly used solution utilizes SGD to train a neural network. SGD uses a batch of a specific size, that is, a small subset of the data at each epoch (iteration) of the optimization to evaluate the gradient, to alleviate the computation hurdle. The step of the derivative at each epoch is controlled by the learning rate, γ. We use ADAM, proposed by Kingma and Ba (2014)[4], which is a more efficient version of SGD. Finally, we set the number of epochs to $5,000$ and use early stopping, following Gu et al. (2020), to mitigate potential overfitting.

4 Empirical Applications to COVID-19 and Inflation in the G7

To test the practical usefulness of the proposed deep pooled panel estimator, we consider two different forecasting exercises: predicting new COVID-19 cases during the pandemic and forecasting inflation. Both applications focus on the G7 countries.

To draw out if and how our deep neural network panel data model confers forecasting gains, we compare it against three benchmarks. These switch off, first, panel (cross-country) interactions, second, nonlinear effects, and third nonlinear and cross-country interactions. We do so by estimating: (i) a "deep time-series" model that is identical to our deep neural network panel data model but is estimated separately for each country; (ii) a panel VAR (PVAR) model that allows for cross-country interactions, but assumes linearity between x_{it} and y_{it+h} and, for parsimony, imposes homogeneity restrictions on the nature of the cross-country dynamic interactions; (iii) and a country-specific AR model.[5] Testing our model against these three

[4]ADAM is using estimates for the first and second moments of the gradient to calculate the learning rate.

[5]Specifically, where $z_{it} = (y_{it}, \mathbf{x}'_{it})'$, we recursively estimate PVAR models of the form $z_{it} = \mu_i + A_1 z_{it-1} + ... + A_q z_{it-q} \epsilon_{it}$ using the BIC to select the optimal lag length, q. We allow for one through 28 lags in the COVID-19 application, and one through four lags in the inflation application. We compute h-step-ahead forecasts of y_{it+h} from the PVAR and AR models via iteration, although it is an empirical question whether such indirect forecasts outperform direct forecast alternatives (e.g., see Marcellino et al. (2006)). For the AR model, we focus on the AR(1), given the widespread use of this model as a univariate forecasting benchmark. The forecasts from the deep models are direct rather than indirect or iterated, and involve relating y_{it} to $\mathbf{x}_{it-h}$ and then using this estimated relationship and the known time t values of $\mathbf{x}_{it}$ to forecast y_{it+h}. As discussed in Remark 2 above, we present results in the main paper using only non-penalized variants of the two deep models. In online appendix B.4, we show that penalizing unambiguously leads to less accurate forecasts.

special cases therefore isolates whether it is allowing for cross-country interactions and/or for nonlinearities that is advantageous when forecasting.[6] In both applications, we also show how partial derivatives can provide interpretable insights into the new model's predictions.

4.1 Forecasting New COVID-19 Cases

4.1.1 COVID-19 Data and the Oxford Stringency Index

We use the four models to forecast, at a daily frequency, t, reports of new COVID-19 cases per 100K of the population, y_{it+h}, from April 2020 through December 2022 for the G7 countries.[7] We source these data from the World Health Organization coronavirus dashboard.

We start by forecasting y_{it+h} using a set of 7 indicators, $\boldsymbol{x}_{it}$, and their lags, previously used to forecast new COVID-19 cases (e.g., see Knutson et al. (2023), Mathieu et al. (2021), and Caporale et al. (2022)). These 7 variables (all reported per 100K of the population) are: new deaths, the reproduction rate, new tests, the share of COVID-19 tests that are positive measured as a rolling 7-day average (this is the inverse of tests per case), the number of people vaccinated, the number of people fully vaccinated, and the number of total boosters. For parsimony, we confine attention to lags at 7, 14, and 21 days.

To assess the out-of-sample Granger causality of pandemic-induced lockdown policies on the spread of COVID-19, we then compare the forecasting performance of our models when we add to $\boldsymbol{x}_{it}$ measures of the stringency of government-imposed containment and lockdown policies. Such (non-pharmaceutical) policies were differentially adopted by many countries from March 2020, including the G7, to reduce the spread of COVID-19.

Specifically, we consider the government response stringency index, as compiled by the Oxford Coronavirus Government Response Tracker (OxCGRT). This index is a composite measure based on 9 response indicators, namely: school closures, workplace closures, the cancellation of public events, restriction on gatherings, public transport closing, requirements to stay at home, movement restriction, restrictions on international travel, and public information campaigns. Throughout the pandemic the Oxford stringency index was a widely consulted measure of policy. Since the Oxford index is an aggregation of 9 indicators, with the weights subjectively chosen by Oxford, we also experiment with forecasting when the underlying 9 disaggregates enter individually into our models, so that, in effect, we objectively use the data to weight the disaggregates. Note that we always consider the lagged effects of policy changes on new COVID-19 cases, mitigating endogeneity concerns that, for example, stricter lockdown policies follow increases in new COVID-19 cases. This means that the cross-sectional dimension of our panel is $p = 36$ when we consider the aggregate stringency index (as published by Oxford) and $p = 68$ when we consider the disaggregated stringency index.

[6] When evaluating the forecasts, we treat them as "primitives," i.e., we ignore model parameter estimation errors.

[7] We take a trailing seven-day rolling average. It is common to smooth daily infection data; cf. Liu et al. (2021).

Following the literature (see, for example, Gu et al. (2021)), prior to model estimation (and forecast evaluation) we rank-normalize all data into the $[0, 1]$ interval as follows:

$$\widetilde{\boldsymbol{z}}_i = \frac{\boldsymbol{z}_i - \min(\boldsymbol{z}_i)}{\max(\boldsymbol{z}_i) - \min(\boldsymbol{z}_i)}, \tag{10}$$

where $\boldsymbol{z}_i$ separately stacks each variable in $\boldsymbol{z}_{it} = (y_{it}, \mathbf{x}'_{it})'$ over time. This normalization minimizes the influence of severely outlying observations stemming from data distributions that may have significant departures from normality, a common feature of COVID-19 data, especially at the beginning of the pandemic.

The online Data Appendix provides additional data details. Figure 2 presents the aggregate stringency index and plots new COVID-19 cases per 100K of the population (so not yet rank-normalized) through our sample period. This figure shows that there are apparent commonalities across countries, both in the stringency of policy and the evolution of new COVID-19 cases. But there are differences too, with Japan standing out as having looser containment policies than the other countries during mid-2020 and then experiencing a later spike in new COVID-19 cases in summer 2022. Thus, it remains an empirical question whether forecasting new COVID-19 cases is improved by pooling information across countries.

4.1.2 Out-of-Sample Forecasting Design

We recursively produce forecasts of y_{it+h} – new COVID-19 cases $h = 7$, $h = 14$, and $h = 21$ days-ahead – by estimating our set of models using expanding estimation windows. Note that for each estimation window, we recursively rank-normalize the data, including the outturns against which we compare the forecasts.[8] To ease the computational burden, given that we re-estimate the model and use CV (as discussed in Section 3), at each estimation window we increase the size of the window in increments of 7 days. We evaluate forecasts over the out-of-sample period February 20, 2021 through December 24, 2022.

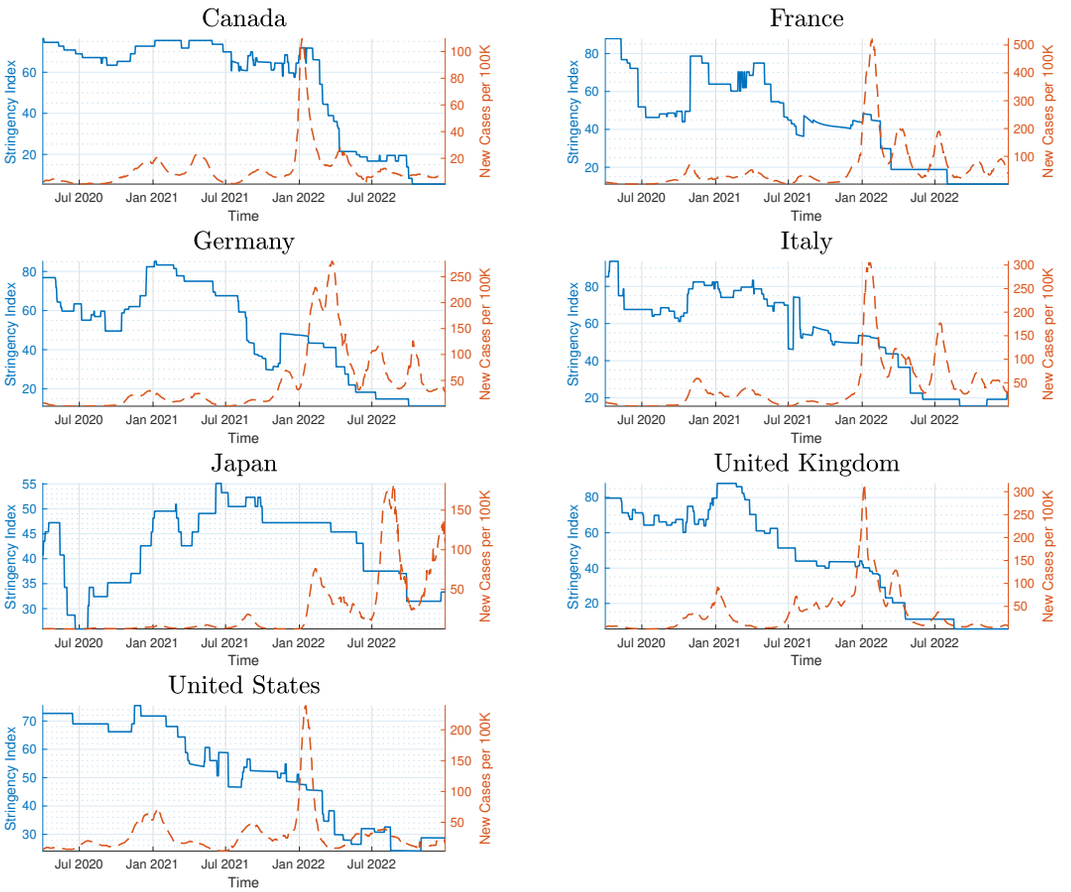
We do not consider forecasting earlier than 7-days-ahead, given that the incubation period of COVID-19 is typically around one week, so that we should not expect policy changes to have effects within one week. During the first wave of the pandemic, many governments updated their policy measures, to restrict the virus, once a week. This further helps to rationalize our choice of forecast horizons.

4.1.3 Forecast Evaluation

In this section we evaluate the forecasting performance of the proposed nonlinear panel estimator relative to the three benchmark models. We then examine whether the inclusion of policy related variables affects forecast accuracy. Specifically, to test for out-of-sample Granger causality of the policy measures adopted by governments to contain the spread of COVID-19, we compare

[8] When rank-normalizing the outturns, we only use information through the point in time the forecast was made.

Figure 2: The Oxford stringency index and new COVID-19 cases per 100K of the population in each of the G7 countries



the forecast accuracy of all of our models with and without the aggregate and disaggregate Oxford stringency indexes.

We evaluate accuracy using the root mean squared forecast error (RMSE) and use the Diebold and Mariano (1995) (DM) test to test whether differences in forecast accuracy across models are statistically significant.

Table 1 compares the accuracy of our deep pooled model against the three benchmarks when we do not include the stringency-based measures of policy and instead focus on using lags of new COVID cases and the other 7 COVID-related measures. The results are striking. The deep pooled model provides large forecasting gains over both the linear PVAR and AR models and the deep time-series neural network across all three forecast horizons. The gains relative to the linear PVAR and AR models increase with the forecast horizon.[9] This evidences the importance of accommodating both the panel dimension and nonlinearites when forecasting

[9]Table 1 tests whether forecasting gains are statistically significant relative to the AR model. Tables B.1-B.3 in the online appendix show that forecasting gains are also statistically significant when compared against the deep time-series model.

the daily path of new COVID-19 cases across the G7 countries. We illustrate in Figure 3 how closely the deep pooled forecasts, at $h = 7$, track subsequent new COVID-19 cases.

Table 1: RMSE statistics for the 7, 14, and 21 days-ahead forecasts of new COVID-19 cases from the 4 models without policy-related variables over the sample February 20, 2021 through December 24, 2022.

	Canada	France	Germany	Italy	Japan	UK	US
			$h = 7$				
AR(1)	0.110	0.139	0.117	0.114	0.166	0.099	0.088
Deep pooled	0.666	0.683	0.971	0.659	0.821	0.886	0.738
Deep time-series	2.798***	2.353***	3.149***	2.771***	2.571***	3.215***	2.819***
PVAR	0.814	0.839	0.811**	0.855	0.663***	1.034	0.935
			$h = 14$				
AR(1)	0.240	0.292	0.211	0.235	0.317	0.181	0.175
Deep pooled	0.380	0.406*	0.557***	0.379	0.427**	0.542	0.427
Deep time-series	1.250	1.027	1.657***	1.263	1.225	1.681***	1.428*
PVAR	0.735	0.816	0.836	0.816	0.558**	1.096	0.927
			$h = 21$				
AR(1)	0.428	0.466	0.328	0.413	0.535	0.262	0.294
Deep pooled	0.269	0.283	0.401**	0.257	0.281**	0.432	0.323
Deep time-series	0.693	0.618	1.041	0.702	0.703	1.122	0.865
PVAR	0.784	0.865	0.918	0.885	0.479**	1.334*	1.081

Notes: The reported models are: Deep pooled: $g\left(\boldsymbol{x}_{it+h}; \boldsymbol{\theta}^{\star}\right)$; Deep time-series: $g\left(\boldsymbol{x}_{t+h}; \boldsymbol{\theta}^{\star}_{TS}\right)$; the PVAR model, and the AR(1) model. $\boldsymbol{\theta}^{\star}$, and $\boldsymbol{\theta}^{\star}_{TS}$ are obtained via out-of-sample CV. Ratios are reported relative to the AR(1) model, whose RMSE is presented in absolute terms in the first gray shaded row. Ratios < 1 indicate superior predictive ability relative to the AR(1). $*$, $**$, and $***$ denote rejection of the null hypothesis of equality of forecast mean squared errors at the 10%, 5%, and 1% levels of significance, respectively, using the Diebold and Mariano (1995) test.

We next test whether the containment or lockdown policies, imposed at the national-level, improve the forecasts. If the policies were effective, conditioning on them should deliver more accurate forecasts. Table 2 presents the relative RMSE ratios for each of the three multivariate forecasting models (so we drop the univariate AR(1) model) when estimating including and excluding the aggregate stringency index. Focusing on the deep pooled model, given its higher accuracy as seen in Table 1, we see that at 7 days policy was only effective in France and Japan. In the other 5 countries, the RMSE ratios are greater than unity, indicating that better one-week-ahead forecasts of new COVID-19 cases are made without the stringency index. But, as we look further ahead to $h = 14$ and $h = 21$, we begin to also see forecast accuracy gains for some other countries, specifically Canada, Germany, and the UK, when using the deep pooled model. For the less accurate deep time-series and PVAR models, policy appears to have been effective in more countries.

Table 3 then tests whether the Oxford stringency index has more value-added when forecasting if we let the data decide how much weight to attach to each of the 9 components (policy levers) in the aggregate index. The fact that in Table 3 the RMSE ratios, for the preferred deep pooled models, are now less than unity across all 7 countries indicates that policy was effective after all – but it is important to let the data determine what policies matter in which country. Table 3 indicates that at $h = 7$ days, policy was least effective in Canada and Italy. While policy interventions still affect new COVID-19 cases, unlike in the other G7 countries, these effects are not statistically significant in these two countries. However, again demonstrating that policy changes took time to have impact, in Canada policy has a larger effect after another week (at $h = 14$ days), as its relative RMSE ratios are lower at 14 days than at 7 days.

Table 2: RMSE ratios, comparing the forecast accuracy of each respective model with and without the aggregate Oxford stringency index at 7, 14, and 21 days-ahead.

	Canada	France	Germany	Italy	Japan	UK	US
			$h = 7$				
Deep pooled	1.073	0.946	1.033	1.233	0.856	1.060	1.309
Deep time-series	0.883	0.936	1.064	0.845	1.045	0.928	1.039
PVAR	1.023***	0.999	1.004	0.995	1.021	0.994	0.974
			$h = 14$				
Deep pooled	1.007	0.920	0.900	1.160	0.834	0.958	1.222
Deep time-series	0.880	0.917	1.060	0.888	1.052	0.912	0.976
PVAR	1.023***	0.997	1.004	0.994	1.036**	0.996	0.971
			$h = 21$				
Deep pooled	0.971	0.887*	0.868*	1.123*	0.846*	0.971	1.089
Deep time-series	0.894	0.872	1.055	0.907	1.055	0.911	0.949
PVAR	1.013**	0.994*	1.004	0.992	1.032**	0.999	0.976**

Notes: Ratios < 1 indicate superior predictive ability for the model with the stringency index. $*$, $**$, and $***$ denote rejection of the null hypothesis of equality of forecast mean squared errors with and without the aggregate Oxford stringency index at the 10%, 5%, and 1% levels of significance, respectively, using the Diebold and Mariano (1995) test.

In the online appendix, we provide additional checks on the forecasting performance of our models. We show that the forecasting gains from the models conditioning on the disaggregated stringency index are often stronger in the first half of our out-of-sample window, when in absolute terms the forecasting errors were higher as COVID-19 infection rates were higher and more volatile. Analysis also indicates that the gains of our deep pooled panel model, over the linear PVAR model, were higher during these earlier waves of COVID-19. This is consistent with the pandemic exhibiting highly nonlinear features in its earlier waves, before vaccinations and other immunities helped restrain the spread of COVID-19. The fluctuation test of Giacomini and Rossi (2010) is used to show that policy in Italy and Japan proved to be effective later

Table 3: RMSE ratios, comparing the forecast accuracy of each respective model with and without the disaggregated Oxford stringency index at 7, 14, and 21 days-ahead.

	Canada	France	Germany	Italy	Japan	UK	US
				$h = 7$			
Deep pooled	0.934	0.846**	0.851**	0.827	0.845*	0.854*	0.847*
Deep time-series	0.936	1.073	1.053	0.902	1.058	0.905	1.080
PVAR	1.030	0.967**	1.011	1.020	1.008	1.001	1.002
				$h = 14$			
Deep pooled	0.918	0.899*	0.879*	0.847**	0.848*	0.917	0.901*
Deep time-series	0.913	1.086	1.043	0.911	1.098	0.899	1.031
PVAR	1.031	0.967**	0.996	1.009	1.019	1.011	1.001
				$h = 21$			
Deep pooled	0.913	0.885*	0.856**	0.832***	0.845**	0.966	0.890**
Deep time-series	0.918	1.087	1.100	0.922	1.108	0.908	1.019
PVAR	0.965	0.955**	0.952	0.952	1.026	0.985	0.949**

Notes: Ratios < 1 indicate superior predictive ability for the model with the disaggregated stringency index. $*$, $**$, and $***$ denote rejection of the null hypothesis of equality of forecast mean squared errors with and without the disaggregated Oxford stringency index at the 10%, 5%, and 1% levels of significance, respectively, using the Diebold and Mariano (1995) test.

than in the other G7 countries: It is only by the fall of 2022 that we see policy having a marked effect on forecast accuracy. We also present results with the LASSO penalization and discuss the observed double descent pattern that our deep models forecast better without any penalty.

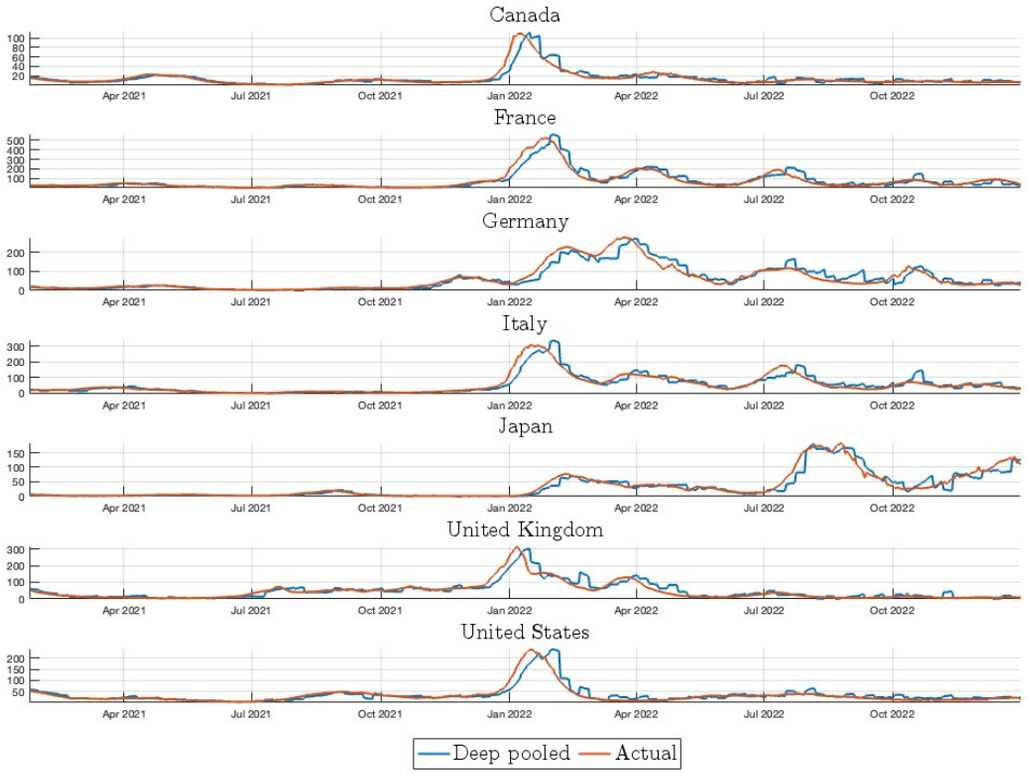
4.1.4 Policy Effectiveness

A common critique of ML algorithms is their putative trade-off between accuracy and interpretability. The output of a highly complicated ML model, such as a deep neural network of the sort we consider, may fit the data well in-sample and even, as we find, out-of-sample. But the model itself is often hard to interpret. In this section, we illustrate how the use of partial derivatives provides one way to assess the impact of covariates. We focus on examination of the effects of changes in policy, as measured by the aggregate and the disaggregated stringency indexes, on the transmission of new COVID-19 cases.

The use of partial derivatives to interpret model output is, of course, common practice in econometrics, ranging from the simple linear regression model to impulse response analysis. Here we show how partial derivatives can be used in deep neural networks to interpret highly nonlinear relationships between covariates and the dependent variable.[10]

[10]We prefer the use of partial derivatives over Shapley additive explanation values, as proposed by Lundberg and Lee (2017), since derivatives tend to be less noisy (see, for example, Chronopoulos et al. (2024)) and computationally less expensive to compute. Perhaps, though, the biggest disadvantage is the set of implicit assumptions, used in the operational construction of Shapley values. A major one is the assumption that inputs are statistically independent. This is discussed in Aas et al. (2021), who also discuss solutions. However, these

Figure 3: COVID-19 new cases per 100K deep pooled forecasts: $h = 7$ forecasts vs outturns



Notes: Forecasts and are presented in original units, i.e., as new cases per 100K.

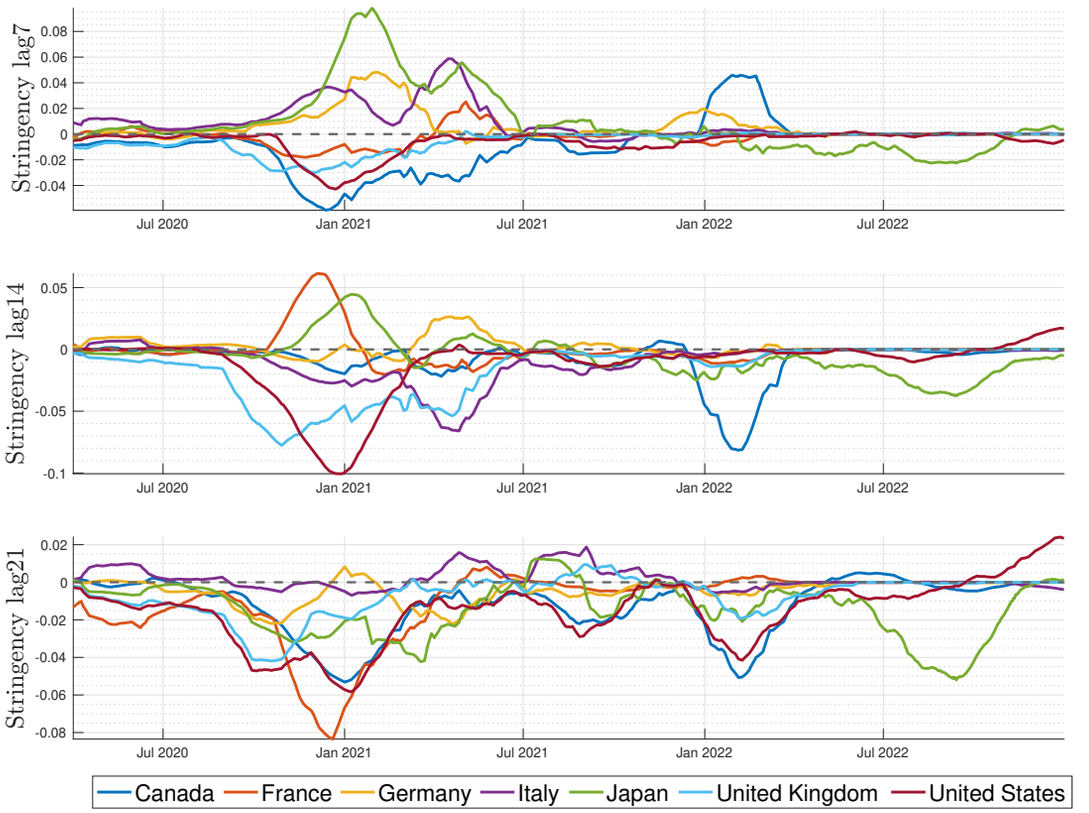
While our deep neural networks are highly nonlinear, their solution/output via SGD optimization methods, can be treated as differentiable functions, as the majority of activation functions are differentiable. In this paper, we consider the case of ReLU, which is not differentiable at zero, whereas it is at every other point in $\mathbb{R}$. From a computational standpoint, the gradient descent, heuristically, works well enough to treat it as a differentiable function. Furthermore, Goodfellow et al. (2016) argue that this issue is negligible and ML software is prone to rounding errors, making them very unlikely to compute the gradient at a singularity point. Note that, even in this extreme case, both SGD and ADAM will use the right sub-gradient at zero.

Let the matrix of characteristics be denoted $\boldsymbol{X}_t \in \mathbb{R}^{N \times p}$, where $\boldsymbol{X}_t = (\boldsymbol{x}_t^{(1)}, \ldots, \boldsymbol{x}_t^{(p)})$. Then for some $i = 1, \ldots, N$, $j = 1, \ldots, p$ and $t = 1, \ldots, T$, the partial derivatives of $g(\boldsymbol{X}_t; \widehat{\theta})$ with respect to the j^{th} characteristic in $\boldsymbol{X}_t$ are:

$$d_{i\,j,\,t} = \frac{\partial g\left(\boldsymbol{X}_t; \widehat{\boldsymbol{\theta}}\right)}{\partial x_{i,\,j,\,t-h}}, \tag{11}$$

are computationally intensive, potentially still quite poor approximations, and not appropriate for large sets of inputs. Although partial derivatives (as well as coefficients in linear models) have similar issues, as discussed in Pesaran and Smith (2014), these issues are both more transparent in nature and much easier to address.

Figure 4: Partial derivatives: The effects of policy (as measured by the Oxford stringency index) on new COVID-19 cases 7, 14 and 21 days after the policy change.



Notes: The partial derivatives, (11), are computed and presented in rank-normalized units, see (10).

where $g(\boldsymbol{X}_t; \widehat{\boldsymbol{\theta}})$ is the function (see Section 2) that approximates the number of new cases per-100K across the i different countries, in our case the G7 countries.

We assess the partial derivatives across time since, following Kapetanios (2007), we expect them to vary due to the inherent nonlinearity of the neural network. Specifically, for each expanding window used in the out-of-sample exercise, we compute the partial derivative with respect to the relevant covariate, given the estimated model parameters at that point in time. Thereby the derivatives evolve over time, as new observations are incorporated into the estimation sample, enabling us to capture any time-varying effects of a given covariate on the dependent variable.

We present point estimates of the partial derivatives, defined in (11), and do not add confidence bands around them. This is because there is currently no rigorous technology available to produce these, especially in the case of penalized estimation. However, recent work by Kapetanios and Kempf (2022) uses a bootstrap approach to construct confidence bands around partial derivatives. A full modification of this work for use in panel models is an interesting and promising avenue to proceed, but is left for future research.

In Figure 4, we present the partial derivatives with respect to the aggregate stringency index at horizons $h \in \{7, 14, 21\}$. Thereby we evaluate the dynamic effectiveness of the stringency

policies adopted across the G7 countries.[11] There are three features that we draw out from Figure 4. First, policy is more effective at containing the spread of COVID-19 after 7 days. Stronger and more negative effects of increases in stringency are seen after 7 days. Secondly, with the exception of in Japan, policy was most effective in the late fall of 2021 and in early 2022, at the time of the highly contagious Omicron variant. The dynamic effects of policy are, on average, much weaker in the second half of our sample. This is consistent with higher vaccination rates meaning that from mid-2021 (non-immunization) policies became less effective at restraining the spread of new COVID-19 cases. Thirdly, there is considerable cross-country variation in the effectiveness of policy. As referenced above when summarizing the Giacomini and Rossi (2010) fluctuation tests reported in the online appendix, policy in Japan is again seen in Figure 4 to have been most effective in late-summer 2022, consistent with COVID-19 cases peaking later in Japan than in the other countries (see Figure 2). Containment policies in Italy tended, relative to the other countries, to have a more muted effect.

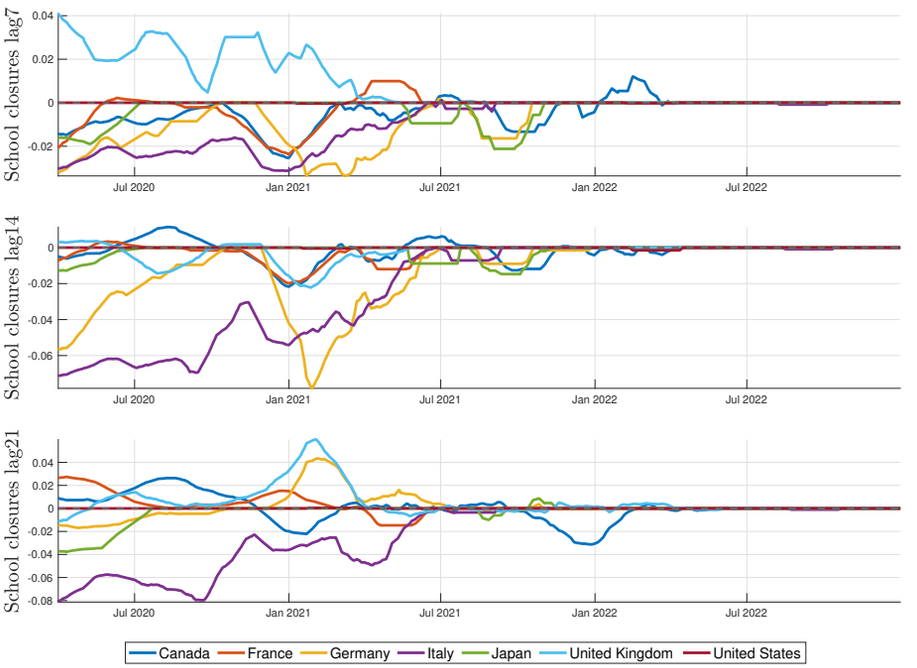
Given the evidence from Table 3 that the disaggregated stringency index confers additional forecasting gains relative to the aggregate index, we next look at the partial derivatives with respect to the 9 components of the Oxford index. This way we aim to shed light on the effectiveness of specific policy measures. We focus on the effects of school and university closings and of workplace closings, since of the 9 components of the Oxford stringency index these tend to be the specific policies associated with the largest marginal effects. Results for the other policy measures are provided in the online appendix. Given the high degree of correlation between the different policy measures (see online Tables A.2-A.3), we should in any case not over interpret these partial derivatives.

Figure 5 shows that over time (as h increases from 7 to 21 days) the effects of school and university closings had an increasingly strong effect. For most countries, as expected, these effects are negative: the closures lead to a fall in new COVID-19 cases. These negative effects are especially strong in Italy. But in the UK, the effects are not so clean cut, with the closings appearing to have a positive effect during the early stages of COVID-19. As in Figure 4, we again see evidence across countries that the effects of school and university closures were far more effective prior to January 2022. Thereafter, the effects are much more modest.

Turning to Figure 6, we see that while workplace closures tended to have a negative effect on COVID-19 soon after the policy change, in particular in Canada and the UK after seven days, thereafter the effects are more uncertain and variable across countries. This can be attributed not just to difficulties in isolating the direct effects of one policy change versus another (related) one, but because in the intervening period there were likely additional and perhaps offsetting changes.

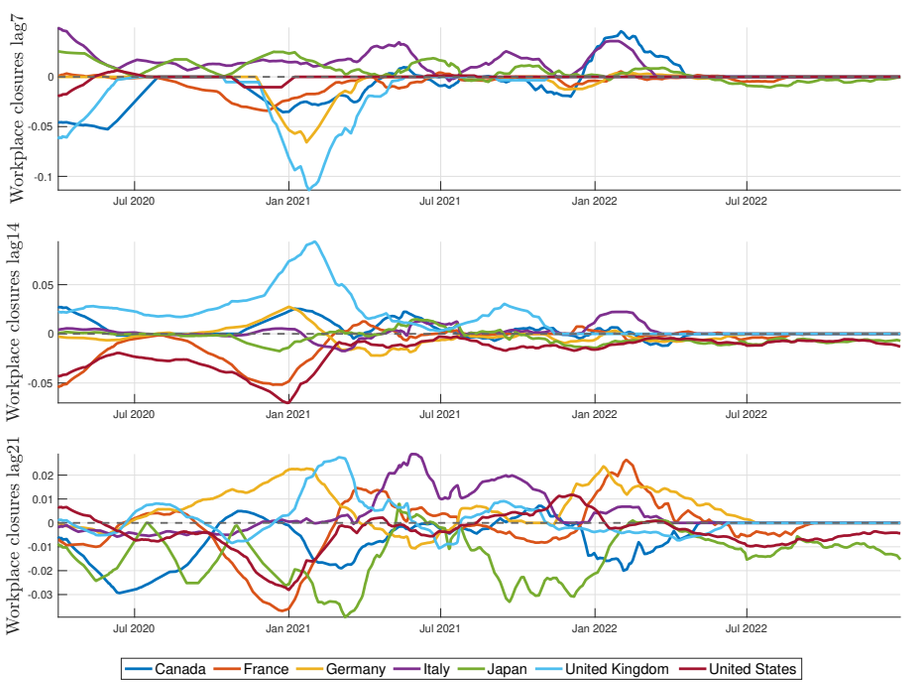
[11] We present the partial derivatives as sixty-day moving averages to smooth out noise.

Figure 5: Partial derivatives: The effects of school and university closures on new COVID-19 cases 7, 14, and 21 days after the policy change.



Notes: The partial derivatives, (11), are computed and presented in rank-normalized units, see (10).

Figure 6: Partial derivatives: The effects of workplace closures on new COVID-19 cases 7, 14, and 21 days after the policy change.



Notes: The partial derivatives, (11), are computed and presented in rank-normalized units, see (10).

4.2 Forecasting Inflation

In this section we turn to assessing how well the proposed nonlinear model can forecast quarterly inflation in the G7 countries.

4.2.1 Inflation Data

We produce forecasts for the quarter-on-quarter headline CPI inflation rate for each of the G7 economies. As indicators, with a Phillips curve framework in mind, we let $\boldsymbol{x}_{it}$ include the (first difference of the) unemployment rate, core CPI inflation, and the energy CPI inflation rate. The inflation data come from the World Bank and the unemployment rate from the OECD. Data are standardized (recursively through the out-of-sample period) prior to model estimation and forecast evaluation.[12]

4.2.2 Phillips Curve-Type Forecasting

Motivated by the Phillips curve, we consider inflation forecasting models of the form:

$$\pi_{i,t+h} = f\left(\pi_{i,t}, \pi^*_{i,t}, u_{i,t}, p^{oil}_{i,t}\right) + \epsilon_{i,t+h}, \tag{12}$$

where $\pi_{i,t+h}$, headline inflation in country $i = 1, \ldots, 7$ at time $t + h$, is determined by lagged inflation, the unemployment rate, $u_{i,t}$, and energy/oil price inflation, $p^{oil}_{i,t}$. As a proxy for inflation expectations, as a measure for trend inflation, we also add core inflation, $\pi^*_{i,t}$, to the set of indicators. Core inflation measures are well-known to be useful indicators for forecasting inflation (see McCracken and Ngan (2023)). For the linear PVAR and AR models, as in the COVID-19 application, we again produce forecasts of $\pi_{i,t+h}$ indirectly, by iterating models relating time t to time $(t-1)$ values of the variables in (12).

The indicator variables used in this formulation thus reflect a conventional (expectations-augmented) Phillips curve-type relationship and align with the empirical models used in papers like Blanchard et al. (2015) and López-Salido and Loria (2024). Importantly, given the evidence from papers like these, our deep pooled model allows for a flexible nonlinear relationship between inflation and its determinants. A wider literature evidences a nonlinear Phillips curve (e.g., see Benigno and Eggertsson (2024)). Our deep pooled model also allows for cross-country linkages, consistent with evidence, such as Auer et al. (2025), that inflation in different countries has common factors.

4.2.3 Out-of-Sample Forecasting Design

We recursively produce and then evaluate $h-$quarters-ahead, $(h = 1, 2, 4, 8)$, forecasts of inflation, $\widehat{\pi}_{i,t+h}$, over the out-of-sample period 2007Q1 through 2024Q4. Forecasts are produced by estimating the deep pooled and benchmark models over expanding samples starting in

[12] Data transformation choices have been found to improve numerical stability when estimating ML models. In this macroeconomic application, we follow neural network papers like Coulombe (2025) and standardize our data.

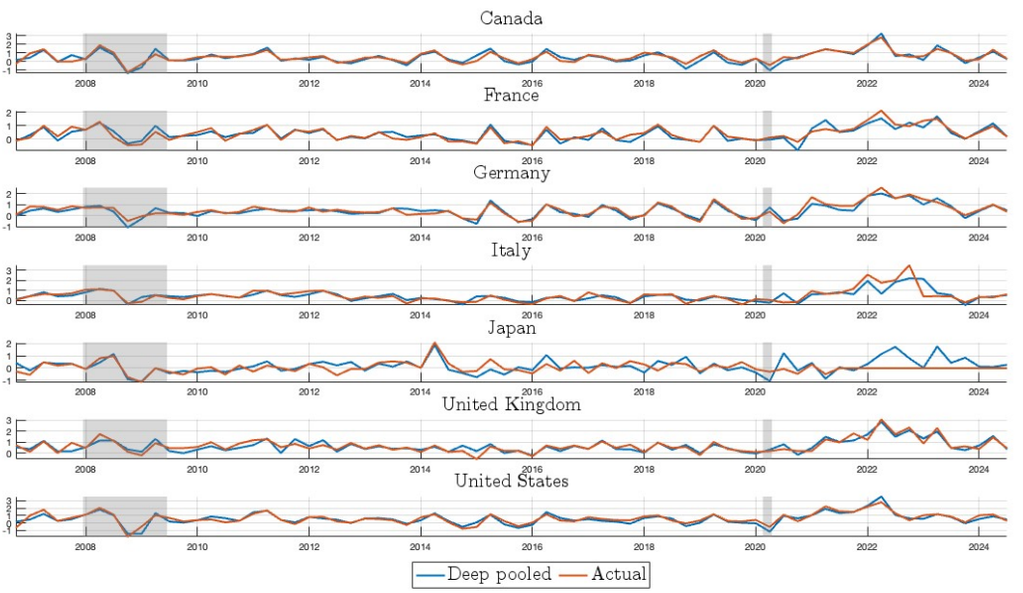
1995Q1. Aware of "temporal instabilities" (cf. Rossi (2021)), we do not consider earlier data. Our sample period includes both the aftermath of the global financial crisis and the run-up of inflation post-pandemic. Deep model parameter estimates are obtained via out-of-sample CV.

Table 4 reports the absolute forecast accuracy of the AR(1) benchmark and then, relative to this, RMSE ratios for the deep pooled model and the benchmark deep time-series and linear PVAR models. Ratios less than unity indicate superior predictive performance relative to the AR(1). We see that both deep learning specifications deliver substantial gains in predictive accuracy in all countries and at all horizons, with the exception of Japan. There are clear gains over both the AR and PVAR models, indicating the benefits of allowing for nonlinearities. Of the two deep models, again for all countries except Japan, the pooled variant generally achieves the lowest RMSE ratios, with improvements often exceeding 50% relative to the AR(1) benchmark.[13] This shows that exploiting cross-country information is in general helpful when forecasting, consistent with inflation having common global determinants.

The fact that Japan is the outlier in Table 4, with the single-country models being harder to beat, is consistent with inflation dynamics in Japan being less similar to other G7 countries. In particular, for Japan, we see no gains in Table 4 to nonlinear pooling, reflecting the (relative to other G7 countries) idiosyncratic nature of Japanese inflation dynamics. Using a dynamic factor model, Auer et al. (2025) also find that Japanese inflation is less affected by common factors than other G7 countries. While the better forecasts for Japan are obtained when cross-country information is not used, we do still see in Table 4 gains to modeling nonlinearities in Japan, with the deep time-series offering the best forecasts. Interestingly, in the online appendix (Table D.1) we show that if we end the out-of-sample evaluation before the inflation surge experienced in the aftermath of COVID-19, then the deep pooled model for Japan becomes the preferred model. This is explained by the fact that inflation did not rise by as much in Japan as in the other G7 countries post-pandemic. Hence, pooling information across countries "forces" the deep pooled forecasts for Japan higher than the subsequent outturns; see Figure 7.

[13] In Table 4, excluding Japan, deep pooled offers more accurate forecasts than deep time-series on 19 out of 24 occasions, so 79% of the time, although these gains are not generally statistically significant.

Figure 7: Inflation forecasts and subsequent outturns ("actual")



Notes: NBER recession bands for US in gray bars. One-quarter-ahead inflation forecasts (from the deep pooled model) and outturns are presented in original units, so as quarter-on-quarter percentage changes.

Table 4: RMSE ratios for the 1-, 2-, 4-, and 8-quarters-ahead forecasts of inflation for the deep pooled, deep time-series, and PVAR models.

	Canada	France	Germany	Italy	Japan	UK	US
				$h = 1$			
AR(1)	1.177	1.302	1.470	1.395	0.930	1.366	1.291
Deep pooled	0.451***	0.461***	0.438***	0.650***	1.075	0.470**	0.437***
Deep time-series	0.719**	0.549***	0.522***	0.611**	0.874	0.513**	0.648
PVAR	1.030**	0.952***	0.969	1.109**	0.978	0.912*	1.050*
				$h = 2$			
AR(1)	1.185	1.304	1.472	1.538	0.887	1.270	1.320
Deep pooled	0.507***	0.567***	0.507***	0.546***	1.218	0.617*	0.429***
Deep time-series	0.565***	0.475***	0.523***	0.612***	0.860	0.660**	0.661
PVAR	1.007*	0.990	0.994	1.022	0.984*	0.990	1.003
				$h = 4$			
AR(1)	1.117	1.290	1.463	1.583	0.863	1.271	1.292
Deep pooled	0.448***	0.553***	0.453***	0.573***	1.236	0.544*	0.483***
Deep time-series	0.766**	0.722***	0.465***	0.501**	0.762	0.617**	0.631***
PVAR	1.001	1.000	1.000	0.994	0.998	1.002	1.002***
				$h = 8$			
AR(1)	1.138	1.257	1.426	1.522	0.843	1.244	1.259
Deep pooled	0.536***	0.442***	0.541***	0.650***	1.244	0.544*	0.460***
Deep time-series	0.742**	0.584***	0.512***	0.498**	0.828	0.653**	0.747
PVAR	1.002**	1.001	1.002	1.010	1.000	0.999	1.002**

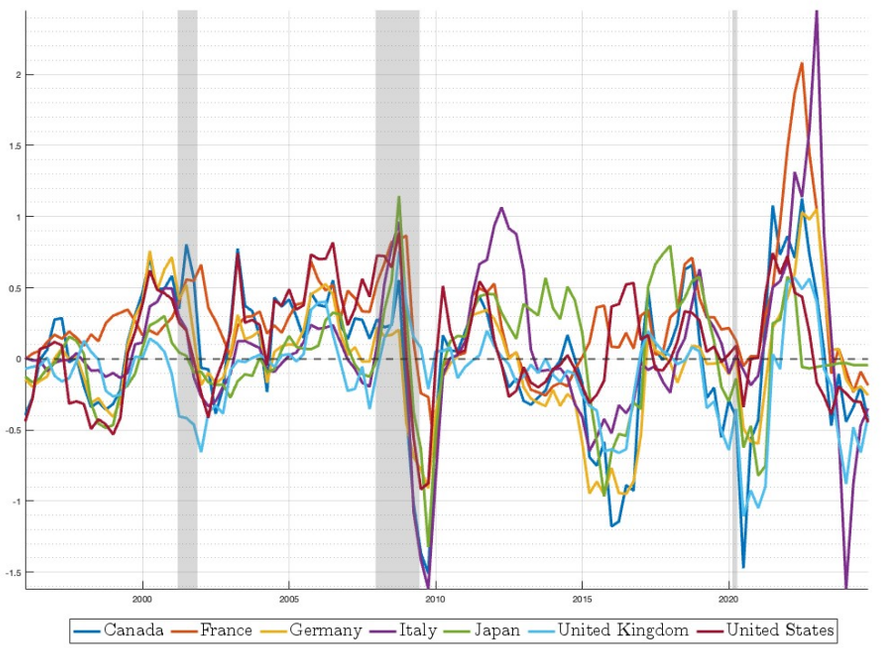
Notes: Ratios are reported relative to the AR(1) model, whose RMSE is presented in absolute terms in the first gray shaded row. Ratios < 1 indicate superior predictive ability relative to the AR(1). $*$, $**$, and $***$ denote rejection of the null hypothesis of equality of forecast mean squared errors at the 10%, 5%, and 1% levels of significance, respectively, using the Diebold and Mariano (1995) test.

4.2.4 The Phillips Curve

We again use partial derivatives, (11), to understand the drivers of the forecasts. Our focus is examining the Phillips curve-type relationship between the unemployment rate and the inflation forecasts. We present results here for the one-quarter-ahead forecasts only, as the derivatives further ahead, as shown in online appendix D, are similar.

Figure 8 depicts, by country, the estimated first derivative of the one-quarter-ahead inflation forecast with respect to the unemployment rate. We see considerable heterogeneity across countries and variation across time, both in terms of the size and sign of the effect. While our analysis is reduced-form, this instability is consistent with the macroeconomic literature evidencing how the slope of the Phillips curve is both hard to identify and time-varying (e.g., see McLeay and Tenreyro (2020)). In Figure 8 we see that, across countries, rises in the unemployment rate were strongly disinflationary in the aftermath of the global financial crisis. They were also often disinflationary in the low-inflation period between the global financial crisis and the COVID-19 pandemic. However, in the early stages of the global financial crisis and the first couple of years post-pandemic, the sign is reversed, and rises in unemployment (other things equal) are estimated to have been inflationary. Given the sharp increase in inflation during this period, this suggests that other factors, like supply shocks, played a larger role in driving inflation dynamics than unemployment alone. It is also consistent with the growing consensus, certainly for the US, that a better measure of slack to use in the Phillips curve when forecasting inflation is the ratio of job openings to vacancies; e.g., see Ball et al. (2025). Movements in the unemployment rate alone may not be helpful or reliable predictors for inflation.

Figure 8: Partial derivatives: The effects of the unemployment rate on the one-quarter-ahead inflation forecasts.



Notes: The partial derivatives, (11), are computed and presented in standardized units. NBER recession bands for US in gray bars.

5 Conclusion

In this paper we propose a nonlinear panel data estimator of the conditional mean based on (deep) neural networks. Utilizing the universal approximation theorem of neural networks, our deep pooled estimator can accommodate nonlinearities and highly complex cross-sectional interactions that neither traditional linear panel methods nor country-specific time-series models can easily capture.

We illustrate the utility of our estimator via two forecasting applications, both using data for G7 countries. First we show that, when forecasting new COVID-19 cases, the deep pooled panel model generates substantial improvements in forecast accuracy over both linear panel data models and deep time series networks that do not exploit the panel dimension. We find that incorporating information about national containment and lockdown policies improves out-of-sample forecasts.

Second, when forecasting inflation we find that our deep pooled estimator consistently outperforms standard benchmarks, except in Japan, emphasizing how its inflation dynamics are distinct from other G7 economies. Our results also underscore the benefits of using a nonlinear forecasting model, so that the model can adapt to changing macroeconomic environments.

In both applications, we explore the use of partial derivatives to show how – on top of producing "good" forecasts – deep learning models can also yield interpretable insights. In the COVID-19 application, we use partial derivatives to show how different containment policies affected the pandemic's trajectory. In the inflation application, we look at the much-debated relationship between unemployment and future inflation. We view these exercises as a first-step toward better integrating ML methods into empirical macroeconomics and policy evaluation, with potential applications well beyond the two studied here.

This paper focuses on the deep pooled estimator, $E[y_{it}|\boldsymbol{x}_{it}] = g(\boldsymbol{x}_{it};\boldsymbol{\theta})$. A natural extension for future research is to allow for heterogeneous nonlinear components across units, $E[y_{it}|\boldsymbol{x}_{it}] = g(\boldsymbol{x}_{it};\boldsymbol{\theta}) + g_i(\boldsymbol{x}_{it};\boldsymbol{\theta}_i)$, where $g(\boldsymbol{x}_{it},\boldsymbol{\theta})$ captures the common structure and $g_i(\boldsymbol{x}_{it},\boldsymbol{\theta}_i)$ represents unit-specific deviations. Developing estimators for this specification would require additional regularization and identification conditions. But this would allow for a unified framework, linking common panel nonlinearities with heterogeneous responses, ranging from heterogeneous nonlinear panel estimators to nonlinear mean group estimators using neural networks.

References

Aas, Kjersti, Martin Jullum, and Anders Løland (2021). "Explaining individual predictions when features are dependent: More accurate approximations to Shapley values." *Artificial Intelligence*, 298, pp. 103–502. doi:10.1016/j.artint.2021.103502.

Ahmadini, Abdullah Ali H, Yashpal Singh Raghav, Ali M Mahnashi, Khalid Ul Islam Rather, and Irfan Ali (2025). "Neural networks to model COVID-19 dynamics and allocate healthcare resources." *Scientific Reports*, 15(1), p. 15,326. doi:10.1038/s41598-025-00153-9.

Athey, Susan and Guido W. Imbens (2017). "The state of applied econometrics: Causality and policy evaluation." *Journal of Economic Perspectives*, 31(2), pp. 3–32. doi:10.1257/jep.31.2.3.

Auer, Raphael, Mathieu Pedemonte, and Raphael Schoenle (2025). "Sixty Years of Global Inflation: A Post-GFC Update." In Guido Ascari and Riccardo Trezzi, editors, *Research Handbook on Inflation*. Edward Elgar Publishing. doi:https://doi.org/10.4337/9781035327768.00031.

Babii, Andrii, Xi Chen, Eric Ghysels, and Rohit Kumar (2020). "Binary choice with asymmetric loss in a data-rich environment: Theory and an application to racial justice." *arXiv preprint arXiv:201008463*. doi:10.48550/arXiv.2010.08463.

Ball, Laurence, Daniel Leigh, and Prachi Mishra (2025). "The Rise and Retreat of US Inflation." *IMF Working Papers*, 2025(094), p. 1. doi:10.5089/9798229007719.001.

Bartlett, Peter L., Nick Harvey, Christopher Liaw, and Abbas Mehrabian (2019). "Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks." *The Journal of Machine Learning Research*, 20(1), pp. 2285–2301. URL `http://jmlr.org/papers/v20/17-612.html`.

Belkin, Mikhail, Daniel Hsu, Siyuan Ma, and Soumik Mandal (2019). "Reconciling modern machine-learning practice and the classical bias–variance trade-off." *Proceedings of the National Academy of Sciences*, 116(32), pp. 15,849–15,854. doi:10.1073/pnas.1903070116.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (2014). "Inference on treatment effects after selection among high-dimensional controls." *The Review of Economic Studies*, 81(2), pp. 608–650. doi:10.1093/restud/rdt044.

Benigno, Pierpaolo and Gauti B. Eggertsson (2024). "The Slanted-L Phillips Curve." *AEA Papers and Proceedings*, 114, p. 84–89. doi:10.1257/pandp.20241051.

Blanchard, Olivier, Eugenio Cerutti, and Lawrence Summers (2015). "Inflation and activity–two explorations and their monetary policy implications." Technical report, National Bureau of Economic Research. doi:10.3386/w21726.

Caporale, Guglielmo Maria, Woo-Young Kang, Fabio Spagnolo, and Nicola Spagnolo (2022). "The COVID-19 pandemic, policy responses and stock markets in the G20." *International Economics*, 172, pp. 77–90. doi:10.1016/j.inteco.2022.09.001.

Chronopoulos, Ilias, Aristeidis Raftapostolos, and George Kapetanios (2024). "Forecasting Value-at-Risk using deep neural network quantile regression." *Journal of Financial Econometrics*, 22(3), pp. 636–669. doi:10.1093/jjfinec/nbad014.

Coulombe, Philippe Goulet (2025). "A neural Phillips curve and a deep output gap." *Journal of Business & Economic Statistics*, 43(3), pp. 669–683. doi:10.1080/07350015.2024.2421279.

Diebold, Francis X. and Robert S. Mariano (1995). "Comparing predictive accuracy." *Journal of Business and Economic Statistics*, 13(3), pp. 253–263. doi:10.1198/073500102753410444.

Farrell, Max H., Tengyuan Liang, and Sanjog Misra (2021). "Deep neural networks for estimation and inference." *Econometrica*, 89(1), pp. 181–213. doi:10.3982/ECTA16901.

Gallant, A. Ronald and Halbert White (1992). "On learning the derivatives of an unknown mapping with multilayer feedforward networks." *Neural Networks*, 5(1), pp. 129–138. doi:/10.1016/S0893-6080(05)80011-5.

Giacomini, Raffaella and Barbara Rossi (2010). "Forecast comparisons in unstable environments." *Journal of Applied Econometrics*, 25(4), pp. 595–620. doi:10.1002/jae.1177.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. MIT Press. URL www.deeplearningbook.org.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu (2020). "Empirical asset pricing via machine learning." *The Review of Financial Studies*, 33(5), pp. 2223–2273. doi:10.1093/rfs/hhaa009.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu (2021). "Autoencoder asset pricing models." *Journal of Econometrics*, 222(1), pp. 429–450. doi:10.1016/j.jeconom.2020.07.009.

Hale, Thomas, Noam Angrist, Rafael Goldszmidt, Beatriz Kira, Anna Petherick, Toby Phillips, Samuel Webster, Emily Cameron-Blake, Laura Hallas, Saptarshi Majumdar, and Helen Tatlow (2021). "A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker)." *Nature Human Behaviour*, 5(4), pp. 529–538. doi:10.1038/s41562-021-01079-8.

Hanin, Boris (2019). "Universal function approximation by deep neural nets with bounded width and relu activations." *Mathematics*, 7(10), p. 992. doi:10.3390/math7100992.

Hastie, Trevor, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani (2022). "Surprises in high-dimensional ridgeless least squares interpolation." *The Annals of Statistics*, 50(2), pp. 949–986. doi:10.1214/21-AOS2133.

Hornik, Kurt (1991). "Approximation capabilities of multilayer feedforward networks." *Neural Networks*, 4(2), pp. 251–257. doi:10.1016/0893-6080(91)90009-T.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White (1989). "Multilayer feedforward networks are universal approximators." *Neural Networks*, 2(5), pp. 359–366. doi:10.1016/0893-6080(89)90020-8.

Ioffe, Sergey and Christian Szegedy (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift." In *International Conference on Machine Learning*, pp. 448–456. JMLR.org. URL https://dl.acm.org/doi/10.5555/3045118.3045167.

Jiang, Ning, Charles Kolozsvary, and Yao Li (2024). "Artificial neural network prediction of covid-19 daily infection count." *Bulletin of Mathematical Biology*, 86(5), p. 49. doi:10.1007/s11538-024-01275-3.

Joseph, Andreas (2019). "Parametric inference with universal function approximators." Bank of England working papers 784, Bank of England. URL https://ideas.repec.org/p/boe/boeewp/0784.html.

Kapetanios, George (2007). "Measuring conditional persistence in nonlinear time series." *Oxford Bulletin of Economics and Statistics*, 69(3), pp. 363–386. doi:10.1111/j.1468-0084.2006.00437.x.

Kapetanios, George and Andrew P. Blake (2010). "Tests of the martingale difference hypothesis using boosting and RBF neural network approximations." *Econometric Theory*, 26(5), pp. 1363–1397. doi:10.1017/S0266466609990612.

Kapetanios, George and Felix Kempf (2022). "Interpretable machine learning for asset pricing." *King's Business School Working Paper No 2022/1*. URL https://www.kcl.ac.uk/business/assets/pdf/dafm-working-papers/2022-papers/interpretable-machine-learning-modelling-for-asset-pricing.pdf.

Kelly, Bryan T., Semyon Malamud, and Kangying Zhou (2022). "The virtue of complexity everywhere." Swiss Finance Institute Research Paper Series 22-57, Swiss Finance Institute. URL https://ideas.repec.org/p/chf/rpseri/rp2257.html.

Kingma, Diederik P. and Jimmy Ba (2014). "Adam: A method for stochastic optimization." *arXiv preprint arXiv:14126980*. doi:10.48550/arXiv.1412.6980.

Knutson, Victoria, Serge Aleshin-Guendel, Ariel Karlinsky, William Msemburi, and Jon Wakefield (2023). "Estimating global and country-specific excess mortality during the COVID-19 pandemic." *The Annals of Applied Statistics*, 17(2), pp. 1353 – 1374. doi:10.1214/22-AOAS1673.

Liang, Shiyu and Rayadurgam Srikant (2016). "Why deep neural networks for function approximation?" *arXiv preprint arXiv:161004161*. doi:10.48550/arXiv.1610.04161.

Liu, Laura, Hyungsik Roger Moon, and Frank Schorfheide (2021). "Panel forecasts of country-level covid-19 infections." *Journal of Econometrics*, 220(1), pp. 2–22. doi:10.1016/j.jeconom.2020.08.010.

Lundberg, Scott M. and Su-In Lee (2017). "A unified approach to interpreting model predictions." In *Advances in Neural Information Processing Systems*, pp. 4765–4774. URL https://dl.acm.org/doi/10.5555/3295222.3295230.

López-Salido, David and Francesca Loria (2024). "Inflation at risk." *Journal of Monetary Economics*, 145, p. 103,570. doi:10.1016/j.jmoneco.2024.103570.

Marcellino, Massimiliano, James H. Stock, and Mark W. Watson (2006). "A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series." *Journal of Econometrics*, 135(1), pp. 499–526. doi:10.1016/j.jeconom.2005.07.020.

Mathieu, Edouard, Hannah Ritchie, Esteban Ortiz-Ospina, Max Roser, Joe Hasell, Cameron Appel, Daniel Gavrilov, Charlie Giattino, and Lucas Rodés-Guirao (2021). "A global database of COVID-19 vaccinations." *Nature Human Behaviour*, 5(7), pp. 947–953. doi:10.1038/s41562-021-01122-8.

Mathieu, Edouard, Hannah Ritchie, Lucas Rodés-Guirao, Cameron Appel, Charlie Giattino, Joe Hasell, Bobbie Macdonald, Saloni Dattani, Diana Beltekian, Esteban Ortiz-Ospina, and Max Roser (2020). "Coronavirus pandemic (COVID-19)." *Our World in Data*. URL https://ourworldindata.org/coronavirus.

McCracken, Michael W. and Tran Khanh Ngan (2023). "Using core inflation to predict headline inflation." *Federal Reserve Bank of St Louis: On the Economy Blog*. URL https://www.stlouisfed.org/on-the-economy/2023/nov/using-core-inflation-predict-headline-inflation.

McLeay, Michael and Silvana Tenreyro (2020). "Optimal inflation and the identification of the Phillips curve." *NBER Macroeconomics Annual*, 34, pp. 199–255. doi:10.1086/707181.

Medeiros, Marcelo C., Gabriel F. R. Vasconcelos, Álvaro Veiga, and Eduardo Zilberman (2021). "Forecasting inflation in a data-rich environment: The benefits of machine learning methods." *Journal of Business & Economic Statistics*, 39(1), pp. 98–119. doi:10.1080/07350015.2019.1637745.

Namasudra, Suyel, S Dhamodharavadhani, and R Rathipriya (2023). "Nonlinear neural network based forecasting model for predicting covid-19 cases." *Neural processing letters*, 55(1), pp. 171–191. doi:10.1007/s11063-021-10495-w.

Park, Jooyoung and Irwin W. Sandberg (1991). "Universal approximation using radial-basis-function networks." *Neural Computation*, 3(4), pp. 246–257. doi:10.1162/neco.1991.3.2.246.

Pesaran, M. Hashem and Ron P. Smith (2014). "Signs of impact effects in time series regression models." *Economics Letters*, 122(2), pp. 150–153. doi:10.1016/j.econlet.2013.11.015.

Rossi, Barbara (2021). "Forecasting in the presence of instabilities: How we know whether models predict well and how to improve them." *Journal of Economic Literature*, 59(4), p. 1135–90. doi:10.1257/jel.20201479.

Schmidt-Hieber, Johannes (2020). "Nonparametric regression using deep neural networks with ReLU activation function." *The Annals of Statistics*, 48(4), pp. 1875 – 1897. doi:10.1214/19-AOS1875.

Tamang, SK, PD Singh, and B Datta (2020). "Forecasting of Covid-19 cases based on prediction using artificial neural network curve fitting technique." *Global Journal of Environmental Science and Management*. doi:10.22034/GJESM.2019.06.SI.06.

Ülke, Volkan, Afsin Sahin, and Abdulhamit Subasi (2018). "A comparison of time series and machine learning models for inflation forecasting: empirical evidence from the USA." *Neural Computing and Applications*, 30(5), pp. 1519–1527. doi:10.1007/s00521-016-2766-x.

Wager, Stefan and Susan Athey (2018). "Estimation and inference of heterogeneous treatment effects using random forests." *Journal of the American Statistical Association*, 113(523), pp. 1228–1242. doi:10.1080/01621459.2017.1319839.

Yarotsky, Dmitry (2017). "Error bounds for approximations with deep ReLU networks." *Neural Networks*, 94, pp. 103–114. doi:10.1016/j.neunet.2017.07.002.

Yarotsky, Dmitry (2018). "Optimal approximation of continuous functions by very deep ReLU networks." *Proceedings of Machine Learning Research*, 75, pp. 1–11. URL http://proceedings.mlr.press/v75/yarotsky18a/yarotsky18a.pdf.

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals (2021). "Understanding deep learning (still) requires rethinking generalization." *Communications of the ACM*, 64(3), pp. 107–115. doi:10.1145/3446776.

Online Appendix for: "Forecasting with Deep Pooled Panel Neural Networks" by Chronopoulos, Chrysikou, Kapetanios, Mitchell, and Raftapostolos

A Data Appendix

This appendix provides additional details on the COVID-19 dataset used in the empirical analysis in Section 4.1 of the main paper. Specifically, we present the main variables for each country, i, that constitute the design matrix $\boldsymbol{X}$; and we provide summary statistics for each variable considered. We assemble our data set from four publicly available different sources. The stringency index is obtained from the Oxford Coronavirus Government Response Tracker (OxCGRT) (these data can be found at https://www.bsg.ox.ac.uk/research/covid-19-government-response-tracker). The daily confirmed COVID-19 cases, which is the "raw data" version of our response, as well as the daily confirmed deaths, were collected from the World Health Organization Coronavirus Dashboard (available at https://covid19.who.int/?mapFilter=cas). The official numbers and metrics from governments and health ministries, worldwide, regarding vaccinations were collected from Mathieu et al. (2021). Lastly, the testing and virus passivity-rates data are from Mathieu et al. (2020).

The rapid spread of COVID-19 led countries to take drastic measures to contain the virus and protect their health systems. The OxCGRT data set gathers together a set of longitudinal measures of government responses from January 1, 2020. These measures include school closings, national/international travel restrictions, bans on public gatherings, emergency investments in healthcare facilities, new forms of social welfare provision, contact tracing, among others; see Hale et al. (2021) for more details. These different measures are then aggregated into one unified measure – the stringency index – that records the strictness of policies that primarily restricted people's behavior, such as via lockdowns. The index is calculated using all ordinal containment and closure policy indicators, including an indicator recording public information campaigns. The higher the value of this index, the stricter the policies adopted. Table A.1 presents the nine different response indicators underlying the aggregate stringency index.

Figure A.1 presents the correlation matrix across new deaths, the reproduction rate, new tests, the share of COVID-19 tests that are positive measured as a rolling 7-day average (this is the inverse of tests per case), the number of people vaccinated, the number of people fully vaccinated, the number of total boosters, and new cases per-100K. In Figures A.2 – A.3 we present the correlation matrix of the different variables for each G7 country.

In Figure A.1 a high negative correlation between the new cases per-100K and the stringency index and its nine components is observed. This of course makes sense, as it implies that when more cases emerge, the stricter the containment policies adopted. Furthermore, there exist a positive correlation between the stringency index and its nine constituent components.

ID	Name	Description	Coding
C1	c1m_school_closing	Record closings of schools and universities	0 - no measures 1 - recommend closing or all schools open with alterations resulting in significant differences compared to non-COVID-19 operations 2 - require closing (only some levels or categories, eg just high school, or just public schools) 3 - require closing all levels Blank - no data
C2	c2m_workplace_closing	Record closings of workplaces	0 - no measures 1 - recommend closing (or recommend work from home) or all businesses open with alterations resulting in significant differences compared to non-Covid-19 operation 2 - require closing (or work from home) for some sectors or categories of workers 3 - require closing (or work from home) for all-but-essential workplaces (eg grocery stores, doctors) Blank - no data
C3	c3m_cancel_public_events	Record cancelling public events	0 - no measures 1 - recommend cancelling 2 - require cancelling Blank - no data
C4	c4m_restrictions_on_gatherings	Record limits on gatherings	0 - no restrictions 1 - restrictions on very large gatherings (the limit is above 1000 people) 2 - restrictions on gatherings between 101-1000 people 3 - restrictions on gatherings between 11-100 people 4 - restrictions on gatherings of 10 people or less Blank - no data
C5	c5m_close_public_transport	Record closing of public transport	0 - no measures 1 - recommend closing (or significantly reduce volume/route/means of transport available) 2 - require closing (or prohibit most citizens from using it) Blank - no data
C6	c6m_stay_at_home_requirements	Record orders to "shelter-in-place" and otherwise confine to the home	0 - no measures 1 - recommend not leaving house 2 - require not leaving house with exceptions for daily exercise, grocery shopping, and 'essential' trips 3 - require not leaving house with minimal exceptions (eg allowed to leave once a week, or only one person can leave at a time, etc) Blank - no data
C7	c7m_movementrestrictions	Record restrictions on internal movement between cities/regions	0 - no measures 1 - recommend not to travel between regions/cities 2 - internal movement restrictions in place Blank - no data
C8	c8ev_internationaltravel	Record restrictions on international travel. Note: this records policy for foreign travellers, not citizens.	0 - no restrictions 1 - screening arrivals 2 - quarantine arrivals from some or all regions 3 - ban arrivals from some regions 4 - ban on all regions or total border closure Blank - no data
H1	h1_public_information_campaigns	Record presence of public info campaigns. Note no differentiated policies reported in this indicator.	0 - no Covid-19 public information campaign 1 - public officials urging caution about Covid-19 2- coordinated public information campaign (eg across traditional and social media) Blank - no data

Table A.1: Mnemonics for the 9 components of the Oxford stringency index

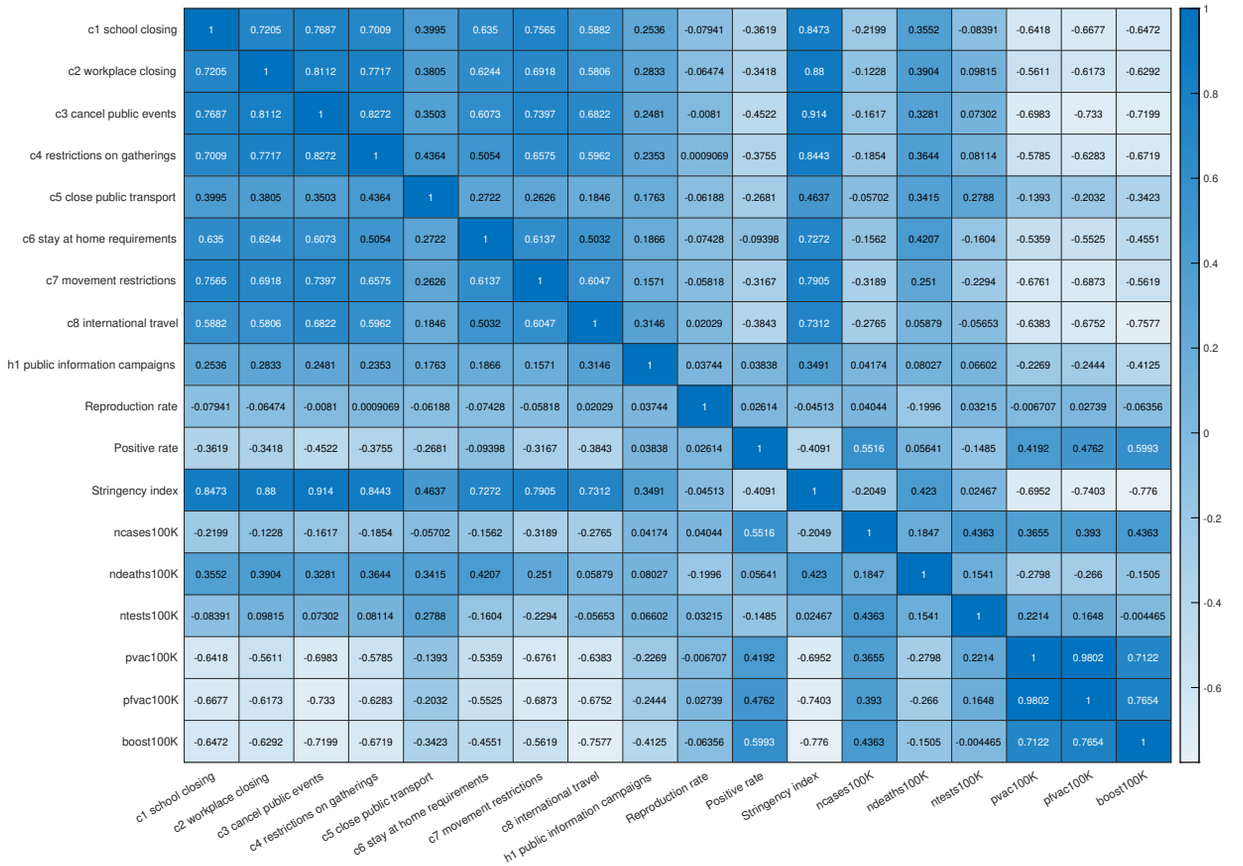


Figure A.1: Correlation matrix (pooled across the G7 countries) between the COVID-19 variables and the Oxford stringency variables

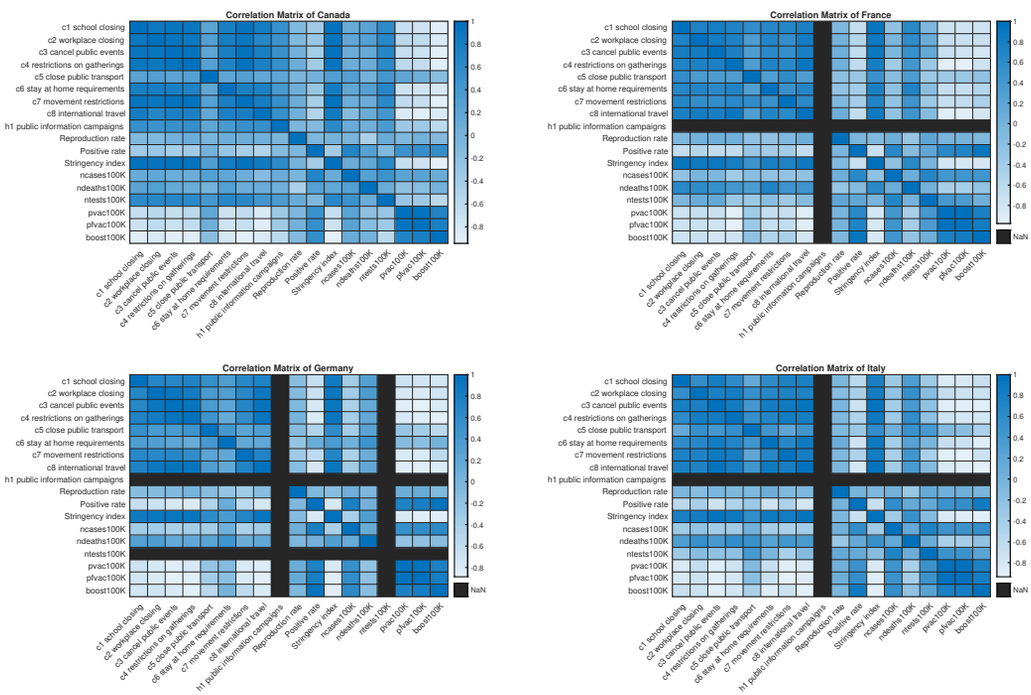


Figure A.2: Correlation matrix by country between the COVID-19 variables and the Oxford stringency variables

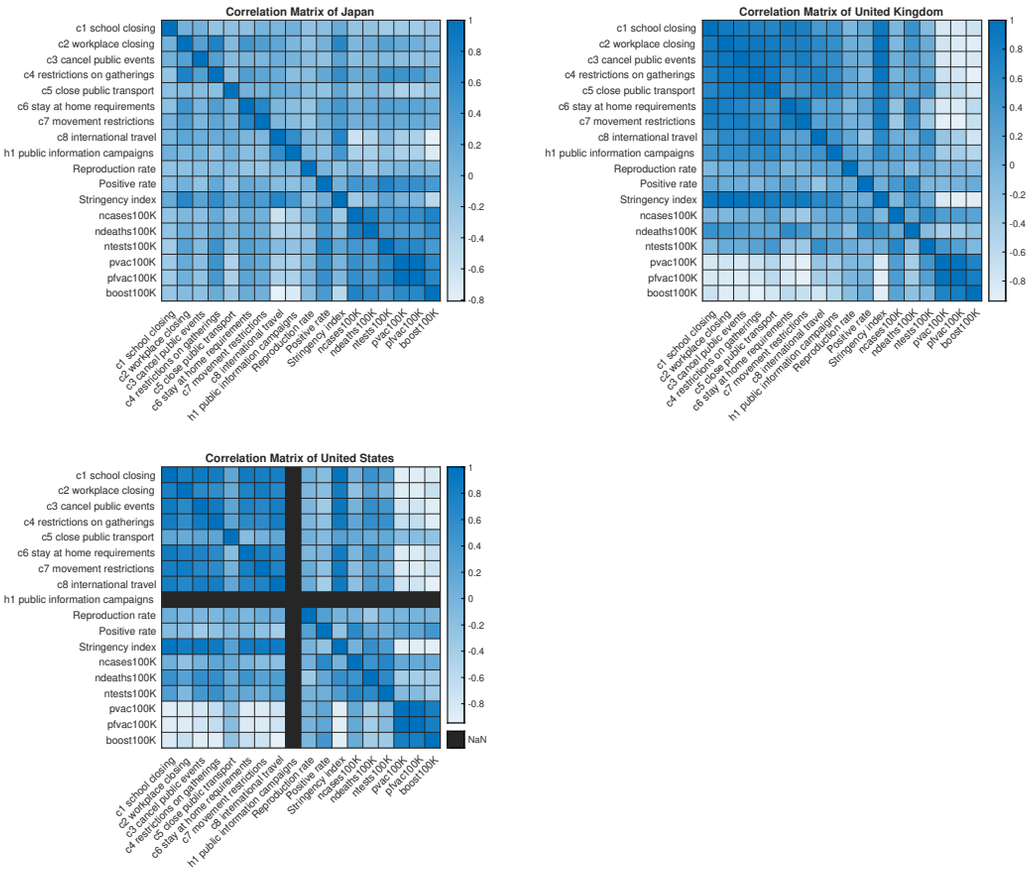


Figure A.3: Correlation matrix by country between the COVID-19 variables and the Oxford stringency variables (cont.)

B Additional empirical results for the COVID-19 application

In this section, we present supplementary results, as referenced in the main paper, when forecasting new COVID-19 cases.

B.1 Diebold-Mariano tests for equal forecast performance against the deep time-series model

We present relative RMSE ratios, in order to compare the forecasting accuracy of each model against the deep time series model, described in more detail in Section 4.1. Similarly to the main paper, we examine the forecasting ability of models without policy variables, specifically the stringency index (see Table B.1), including the aggregated stringency index (see Table B.2), and including the dissagregated stringency index (see Table B.3).

Table B.1: Forecasting results without the stringency index

	Canada	France	Germany	Italy	Japan	UK	US
				$h = 7$			
Deep pooled	0.227***	0.293***	0.307***	0.239***	0.322***	0.284***	0.252***
PVAR	0.278***	0.363***	0.262***	0.322***	0.227***	0.325***	1.102***
				$h = 14$			
Deep pooled	0.304***	0.395***	0.336***	0.300***	0.349***	0.323***	0.299***
PVAR	0.588**	0.795*	0.505***	0.646***	0.456***	0.652***	0.649***
				$h = 21$			
Deep pooled	0.389***	0.458***	0.386***	0.366***	0.400***	0.384***	0.374***
PVAR	1.131	1.399*	0.883	1.260	0.681***	1.189	1.249

Notes: RMSE ratios, comparing the accuracy of each model against the deep time series model. Ratios < 1 indicate superior predictive ability of the respective model relative to the deep time-series model. $*$, $**$, and $***$ denote rejection of the null hypothesis of equality of forecast mean squared errors at the 10%, 5%, and 1% levels of significance, respectively, using the Diebold and Mariano (1995) test.

Table B.2: Forecasting results with the aggregate Oxford stringency index

	Canada	France	Germany	Italy	Japan	UK	US
				$h = 7$			
Deep pooled	0.290***	0.293***	0.299***	0.347***	0.261***	0.315***	0.330***
PVAR	0.337***	0.381***	0.243***	0.363***	0.252***	0.345***	0.311***
				$h = 14$			
Deep pooled	0.348***	0.397***	0.285***	0.392***	0.277***	0.339***	0.375***
PVAR	0.683***	0.864	0.478***	0.723***	0.449***	0.712**	0.646***
				$h = 21$			
Deep pooled	0.423***	0.465***	0.318***	0.452***	0.321***	0.410***	0.429***
PVAR	1.282	1.594**	0.840	1.378	0.667***	1.303	1.283

Notes: RMSE ratios, comparing the accuracy of each model against the deep time series model. Ratios < 1 indicate superior predictive ability of the respective model relative to the deep time-series model. For a description of the forecasting models, see the notes to Table 1. See further notes in Table B.1

B.2 Forecast evaluation in sub-samples

In this section we present supplementary forecasting results as referenced in the main paper. Specifically, we evaluate the forecasting performance of the proposed nonlinear panel estimator(s) relative to the two benchmark models, described in Section 4.1 of the main paper, over two distinct sub-periods within our overall out-of-sample window. The first sub-sample covers the period from February 20, 2021 to April 30, 2022, when COVID-19 was at its worst except in Japan, while the second is from May 1, 2022 through December 24, 2022. Our aim is to examine whether policy mattered more in this earlier period, before immunity within each of the countries strengthened and COVID infection rates declined.

Table B.3: Forecasting results with the disaggregated Oxford stringency index

	Canada	France	Germany	Italy	Japan	UK	US
				$h = 7$			
Deep pooled	0.237***	0.229***	0.249***	0.218***	0.255***	0.260***	0.206***
PVAR	0.319***	0.321***	0.247***	0.349***	0.246***	0.356***	0.308***
				$h = 14$			
Deep pooled	0.306***	0.327***	0.283***	0.279***	0.270***	0.329***	0.262***
PVAR	0.664***	0.708*	0.482***	0.716***	0.423***	0.733**	0.630***
				$h = 21$			
Deep pooled	0.387***	0.373***	0.300***	0.330***	0.305***	0.409***	0.327***
PVAR	1.189	1.230	0.764*	1.300	0.631***	1.289	1.163***

Note: RMSE ratios, comparing the accuracy of each model against the deep time series model. Ratios < 1 indicate superior predictive ability of the respective model relative to the deep time-series model. For a description of the forecasting models, see the notes to Table 1. See further notes in Table B.1

In Tables B.4–B.5 we compare RMSE statistics across different models and countries for these first and second sub-periods. We do not include the stringency-based measures of policy and instead focus on predicting new COVID-19 cases using lags of new cases and the other seven COVID-related measures. We find across the forecasting horizons, $h \in \{7, 14, 21\}$ days, that the deep models yield significant forecasting gains over both the linear PVAR model and the deep time-series neural network. Similarly to the analysis in Section 4.1, this shows the importance of both the panel dimension and of modeling nonlinearites when forecasting the daily path of new COVID-19 cases across the G7 countries. As anticipated, the RMSE values are smaller in the second sub-period, indicative of the lower COVID-19 transmission rates seen in Figure 2 from May 2022.

In Tables B.6–B.7 we present RMSE ratios, comparing the predictive ability of each model with and without the aggregate Oxford stringency index over the two sub-samples. We see that policy as measured by the aggregate stringency index is more effective, with more RMSE ratios less than unity, in the latter sub-sample. But turning to the disaggregate stringency index, we see from Tables B.8–B.9 that policy was then more effective even over the first sub-period. It is important to let the models choose how to weights the 9 components of the Oxford stringency index.

Table B.4: RMSE statistics for the 7, 14, and 21 days-ahead forecasts of new COVID-19 cases from the 4 models without policy-related variables over the sample February 20, 2021 to April 30, 2022.

	Canada	France	Germany	Italy	Japan	UK	US
			$h = 7$				
Deep pooled	0.088	0.108	0.122	0.067	0.115	0.104	0.074
Deep time-series	0.376	0.390	0.433	0.339	0.402	0.376	0.295
PVAR	0.111	0.144	0.114	0.117	0.110	0.126	0.102
AR(1)	0.137	0.171	0.139	0.135	0.180	0.123	0.109
			$h = 14$				
Deep pooled	0.112	0.141	0.129	0.091	0.129	0.118	0.088
Deep time-series	0.368	0.353	0.406	0.305	0.356	0.363	0.301
PVAR	0.220	0.294	0.213	0.233	0.189	0.247	0.3203
AR(1)	0.300	0.361	0.252	0.283	0.358	0.226	0.219
			$h = 21$				
Deep pooled	0.144	0.159	0.150	0.116	0.149	0.139	0.116
Deep time-series	0.366	0.338	0.397	0.297	0.343	0.351	0.304
PVAR	0.423	0.501	0.367	0.453	0.281	0.440	0.401
AR(1)	0.539	0.582	0.387	0.509	0.620	0.328	0.371

Table B.5: RMSE statistics for the 7, 14, and 21 days-ahead forecasts of new COVID-19 cases from the 4 models without policy-related variables over the sample May 1, 2022 to December 24, 2022.

	Canada	France	Germany	Italy	Japan	UK	US
			$h = 7$				
Deep pooled	0.033	0.067	0.098	0.088	0.168	0.043	0.044
Deep time-series	0.105	0.166	0.213	0.269	0.465	0.166	0.120
PVAR	0.017	0.033	0.046	0.045	0.110	0.025	0.010
AR(1)	0.015	0.043	0.059	0.059	0.136	0.013	0.013
			$h = 14$				
Deep pooled	0.032	0.065	0.095	0.086	0.147	0.044	0.042
Deep time-series	0.103	0.172	0.220	0.283	0.440	0.153	0.119
PVAR	0.028	0.075	0.080	0.083	0.154	0.040	0.020
AR(1)	0.028	0.084	0.105	0.109	0.230	0.029	0.021
			$h = 21$				
Deep pooled	0.032	0.065	0.094	0.086	0.153	0.046	0.041
Deep time-series	0.103	0.175	0.216	0.278	0.425	0.162	0.136
PVAR	0.038	0.120	0.130	0.118	0.208	0.065	0.037
AR(1)	0.040	0.116	0.191	0.147	0.346	0.065	0.028

Table B.6: RMSE ratios, comparing the forecast accuracy of each respective model with and without the aggregate Oxford stringency index at 7, 14, and 21 days-ahead over the sample February 20, 2021 to April 30, 2022

	Canada	France	Germany	Italy	Japan	UK	US
			$h = 7$				
Deep pooled	1.096	0.934	1.092	1.466	1.013	1.064	1.387
Deep time-series	0.835	0.940	1.096	0.909	1.138	0.950	1.026
PVAR	1.021***	0.999	1.006	0.996	1.030*	0.993	0.974
			$h = 14$				
Deep pooled	1.014	0.907*	0.900	1.248	0.931	0.947	1.261
Deep time-series	0.899	0.910	1.116	0.997	1.165	0.913	0.936
PVAR	1.021	1.019	1.011	0.997	1.040**	0.998	1.017
			$h = 21$				
Deep pooled	0.954	0.876*	0.850*	1.166	0.915	0.953	1.090
Deep time-series	0.944	0.865	1.117	0.993	1.141	0.938	0.934
PVAR	1.026	1.0224	1.039	1.006	0.842	0.922	1.022

Notes: Ratios < 1 indicate superior predictive ability for the model with the stringency index. $*$, $**$, and $***$ denote rejection of the null hypothesis of equality of forecast mean squared errors with and without the aggregate Oxford stringency index at the 10%, 5%, and 1% levels of significance, respectively, using the Diebold and Mariano (1995) test.

Table B.7: RMSE ratios, comparing the forecast accuracy of each respective model with and without the aggregate Oxford stringency index at 7, 14, and 21 days-ahead over the sample May 1, 2022 to December 24, 2022

	Canada	France	Germany	Italy	Japan	UK	US
			$h = 7$				
Deep pooled	0.916	0.900	0.853	0.911	0.741	0.764	0.896
Deep time-series	1.103	1.005	0.920	0.793	0.902	0.948	1.193
PVAR	1.171***	0.981	0.978**	0.981	0.995	1.041	1.026
			$h = 14$				
Deep pooled	0.954	0.906	0.880	0.935	0.742	1.029	0.883
Deep time-series	1.138	0.954	0.877	0.736	0.873	0.985	1.212
PVAR	1.181***	0.977**	0.969**	0.978	0.998	1.035	1.003
			$h = 21$				
Deep pooled	0.926	0.917	0.928	0.947	0.727	1.022	0.894
Deep time-series	1.093	0.953	0.945	0.762	0.871	0.922	1.062
PVAR	1.134*	0.974**	0.971**	0.974	0.994	0.985	0.971

Notes: Ratios < 1 indicate superior predictive ability for the model with the stringency index. $*$, $**$, and $***$ denote rejection of the null hypothesis of equality of forecast mean squared errors with and without the aggregate Oxford stringency index at the 10%, 5%, and 1% levels of significance, respectively, using the Diebold and Mariano (1995) test.

Table B.8: RMSE ratios, comparing the forecast accuracy of each respective model with and without the disaggregate Oxford stringency index at 7, 14, and 21 days-ahead over the sample February 20, 2021 to April 30, 2022

	Canada	France	Germany	Italy	Japan	UK	US
			$h = 7$				
Deep pooled	0.934	0.847**	0.880*	1.027	0.963	0.852*	0.867
Deep time-series	0.860	1.053	1.041	0.928	1.087	0.864	1.074
PVAR	1.017	0.970**	0.993	1.012	1.036**	1.001	0.991
			$h = 14$				
Deep pooled	0.915	0.895*	0.889*	0.932	0.901	0.894*	0.929
Deep time-series	0.886	1.079	1.034	0.970	1.108	0.838	1.040
PVAR	1.588	1.564	1.165	1.080	1.054	0.995	1.551
			$h = 21$				
Deep pooled	0.897	0.886*	0.859**	0.878**	0.880*	0.939	0.898**
Deep time-series	0.951	1.096	1.027	0.969	1.095	0.860	1.009
PVAR	1.767	1.748	1.734	1.456	0.871	1.400	1.728

Notes: Ratios < 1 indicate superior predictive ability for the model with the stringency index. $*$, $**$, and $***$ denote rejection of the null hypothesis of equality of forecast mean squared errors with and without the aggregate Oxford stringency index at the 10%, 5%, and 1% levels of significance, respectively, using the Diebold and Mariano (1995) test.

Table B.9: RMSE ratios, comparing the forecast accuracy of each respective model with and without the disaggregate Oxford stringency index at 7, 14, and 21 days-ahead over the sample May 1, 2022 to December 24, 2022.

	Canada	France	Germany	Italy	Japan	UK	US
			$h = 7$				
Deep pooled	1.020	0.771	0.769	0.600*	0.743*	0.734*	0.719**
Deep time-series	1.712	1.150	1.272	0.854	1.039	1.192	1.081
PVAR	0.977	0.935**	1.044	0.954	0.960	1.106***	1.669***
			$h = 14$				
Deep pooled	1.027	0.870	0.795	0.627*	0.796	0.797	0.717**
Deep time-series	1.698	1.084	1.243	0.799	1.081	1.309	1.097
PVAR	0.957	0.945	1.00	0.960	0.956	1.114***	1.743***
			$h = 21$				
Deep pooled	1.007	0.872	0.781	0.624*	0.766	0.764*	0.716**
Deep time-series	1.660	1.047	1.541	0.809	1.068	1.227	1.015
PVAR	0.973	0.957	1.064	0.965	0.961	1.062*	1.549***

Notes: Ratios < 1 indicate superior predictive ability for the model with the stringency index. For a description of the 4 forecasting models, see the notes to Table 1. $*$, $**$, and $***$ denote rejection of the null hypothesis of equality of forecast mean squared errors with and without the aggregate Oxford stringency index at the 10%, 5%, and 1% levels of significance, respectively, using the Diebold and Mariano (1995) test.

B.3 Temporal instabilities in forecast performance: the fluctuation test

To compare the predictive performance of competing models in unstable environments, Giacomini and Rossi (2010) propose the fluctuation test. It utilizes the test statistic of Diebold and Mariano (1995) computed over rolling out-of-sample windows of size m. Given the evidence in Table B.8 that the disaggregate Oxford stringency index improves the forecasts from the deep pooled panel model – on average over the period February 20, 2021 through December 24, 2022 – Figure B.1 uses the fluctuation test to test the null hypothesis that the local RMSE equals zero at each point in time. When the test statistic (the solid blue line) crosses the critical values (the dashed red line) equal forecast performance is rejected.

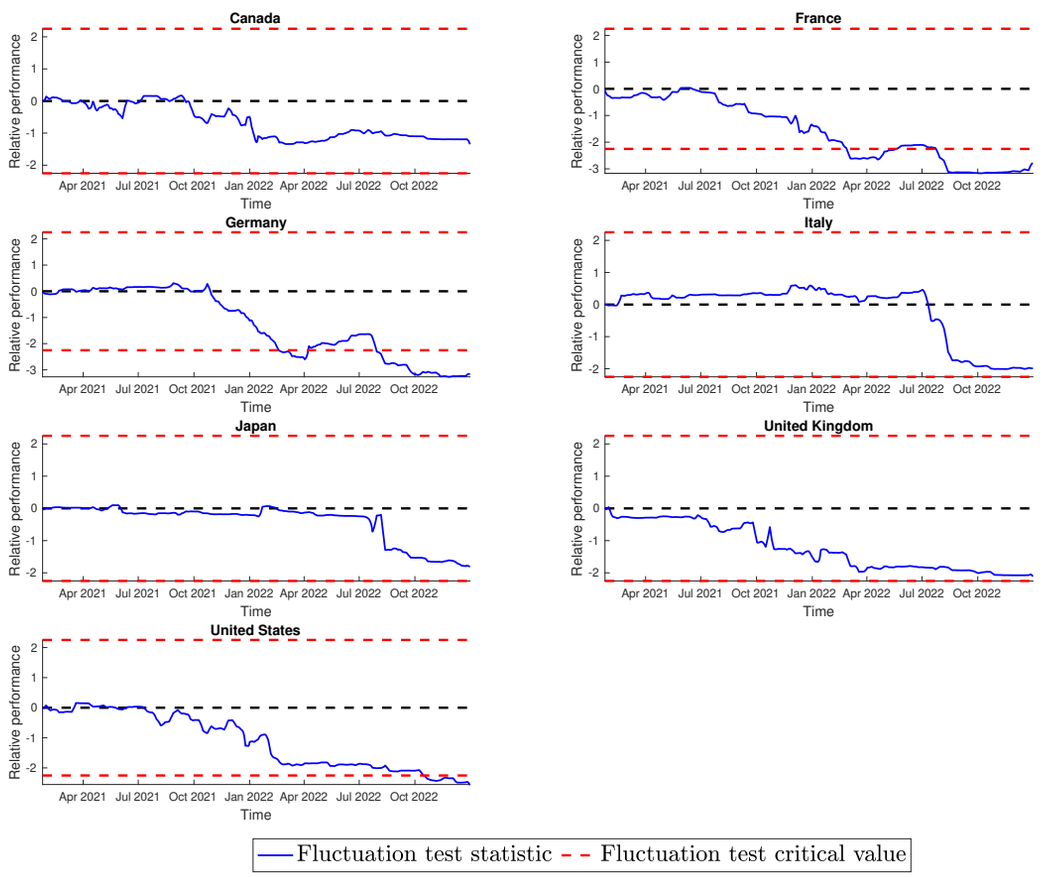


Figure B.1: Giacomini and Rossi's (2010) fluctuation test, obtained as the (standardized) difference between the MSE of the deep pooled panel model with and without the disaggregated Oxford stringency index at $h = 7$. Negative values of the fluctuation statistic imply that the model with the disaggregate stringency index is better. Critical values are at the 10% level of significance.

B.4 Empirical evidence using penalized models

In this section we present forecasting results for the COVID-19 application adding an ℓ_1 penalty on the weights of each corresponding network estimator; see Section 3 of the main paper. We follow the same forecasting design as in Section 4.1.2 of the main paper.

We start by presenting the results from the penalized estimation. In Table B.10 we report the RMSE ratio of each model with the aggregate Oxford stringency index as considered in Table 2, i.e., deep pooled, and deep time series versus the LASSO deep pooled panel model, and deep time series LASSO. Ratios less than one indicate superior predictive ability for the model without the LASSO penalization. In Table B.11 we report the same metrics as in Table B.10, but consider the disaggregated stringency index as considered in Table 3 in the main paper.

In both Tables B.10–B.11 the evidence is compelling. We find that the heavily parameterized models – deep pooled and deep time-series – forecast better without penalization. On the face of it, this seems quite surprising, given that in many other contexts penalized models have been found to forecast well. But this finding can be understood in relation to the recent statistical literature on so called *double descent*; see Hastie et al. (2022) and Kelly et al. (2022). We discuss this issue further in Remark 2 in Section 3 of the main paper.

Table B.10: RMSE ratios, comparing the forecast accuracy of each respective model with the aggregate Oxford stringency index 7 days-ahead when estimated with and without penalization

	Canada	France	Germany	Italy	Japan	UK	US
Deep pooled	0.809	0.740	0.670	0.782	0.620	0.736	0.772
Deep time-series	0.854	1.047	0.989	0.884	0.923	1.057	1.016

Notes: Entries < 1 indicate superior predictive ability of the model with no penalty.

Table B.11: RMSE ratios, comparing the forecast accuracy of each respective model with the disaggregate Oxford stringency index 7 days-ahead when estimated with and without penalization

	Canada	France	Germany	Italy	Japan	UK	US
Deep pooled	0.728	0.600	0.581	0.653	0.541	0.664	0.628
Deep time-series	1.048	1.081	0.984	0.989	1.023	0.978	0.812

Notes: Entries < 1 indicate superior predictive ability of the model with no penalty.

C The effectiveness of policy: Disaggregated partial derivatives

This section presents plots of the partial derivatives, (11), for those disaggregated stringency measures from the Oxford index not shown in the main paper.

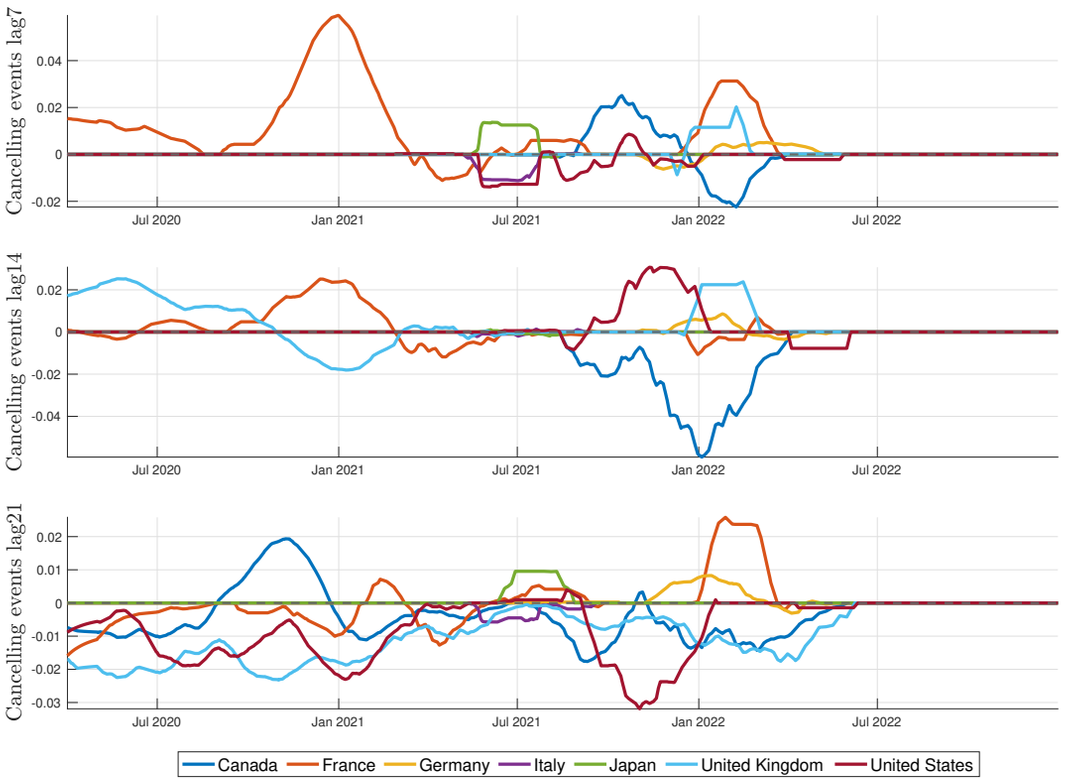


Figure C.1: Partial derivatives: The effects of cancelling public events on new COVID-19 cases 7, 14, and 21 days after the policy change. The partial derivatives, (11), are computed and presented in rank-normalized units, see (10).

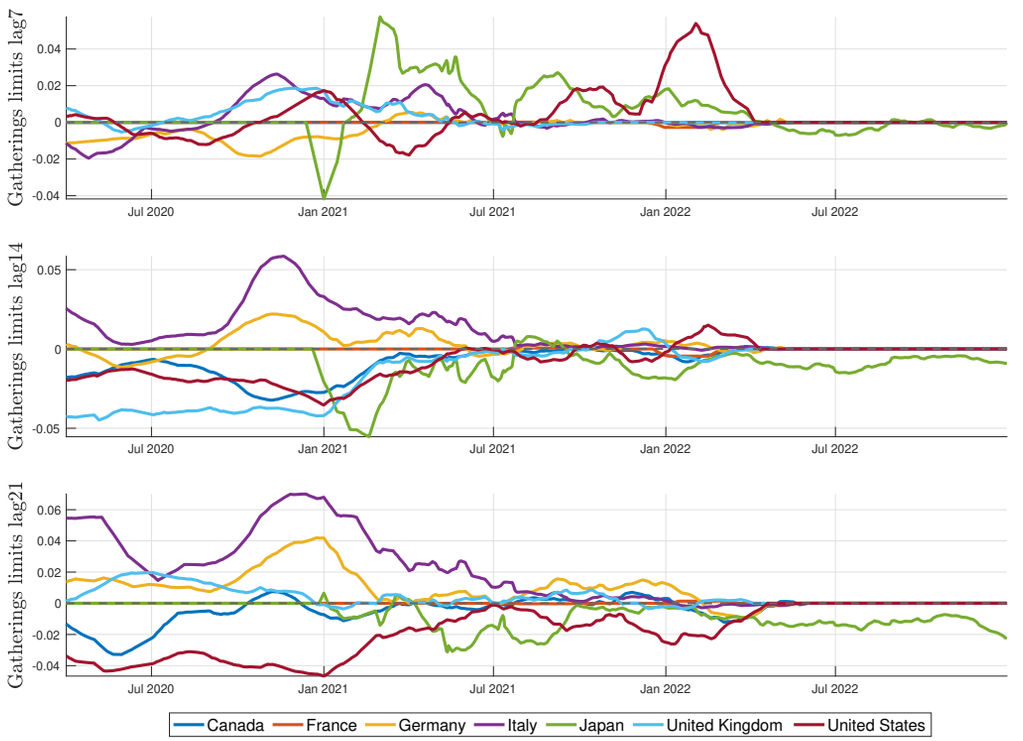


Figure C.2: Partial derivatives: The effects of imposing limits on gatherings on new COVID-19 cases 7, 14, and 21 days after the policy change. The partial derivatives, (11), are computed and presented in rank-normalized units, see (10).

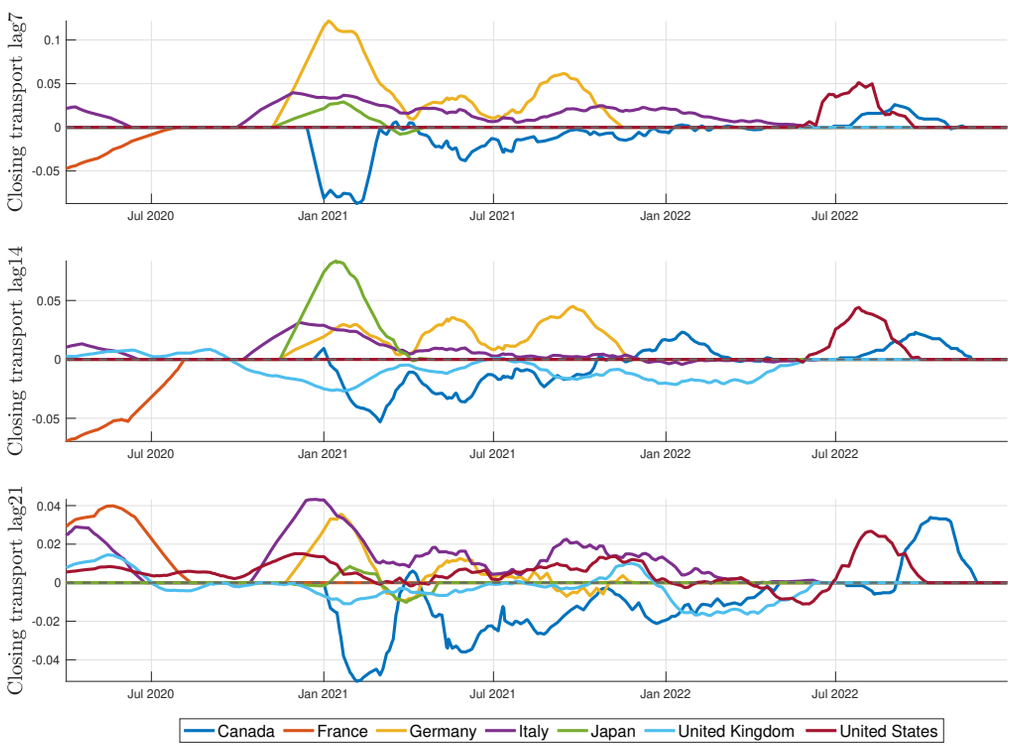


Figure C.3: Partial derivatives: The effects of closing public transport on new COVID-19 cases 7, 14, and 21 days after the policy change. The partial derivatives, (11), are computed and presented in rank-normalized units, see (10).

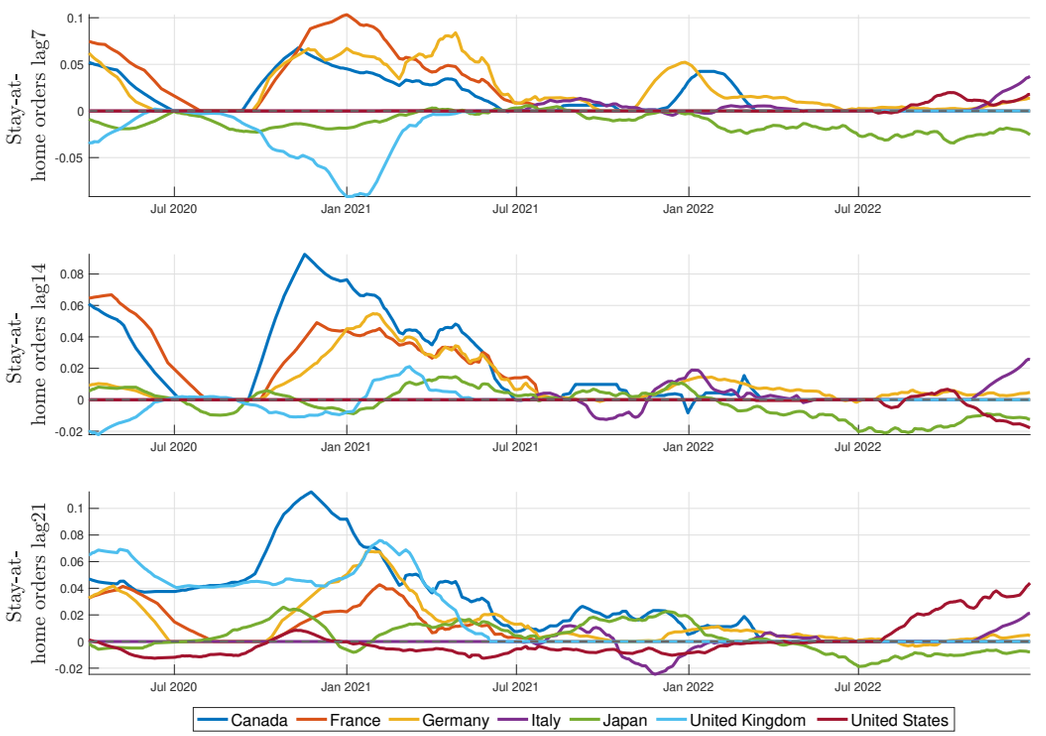


Figure C.4: Partial derivatives: The effects of orders to "shelter-in-place" and other stay-at-home orders on new COVID-19 cases 7, 14, and 21 days after the policy change. The partial derivatives, (11), are computed and presented in rank-normalized units, see (10).

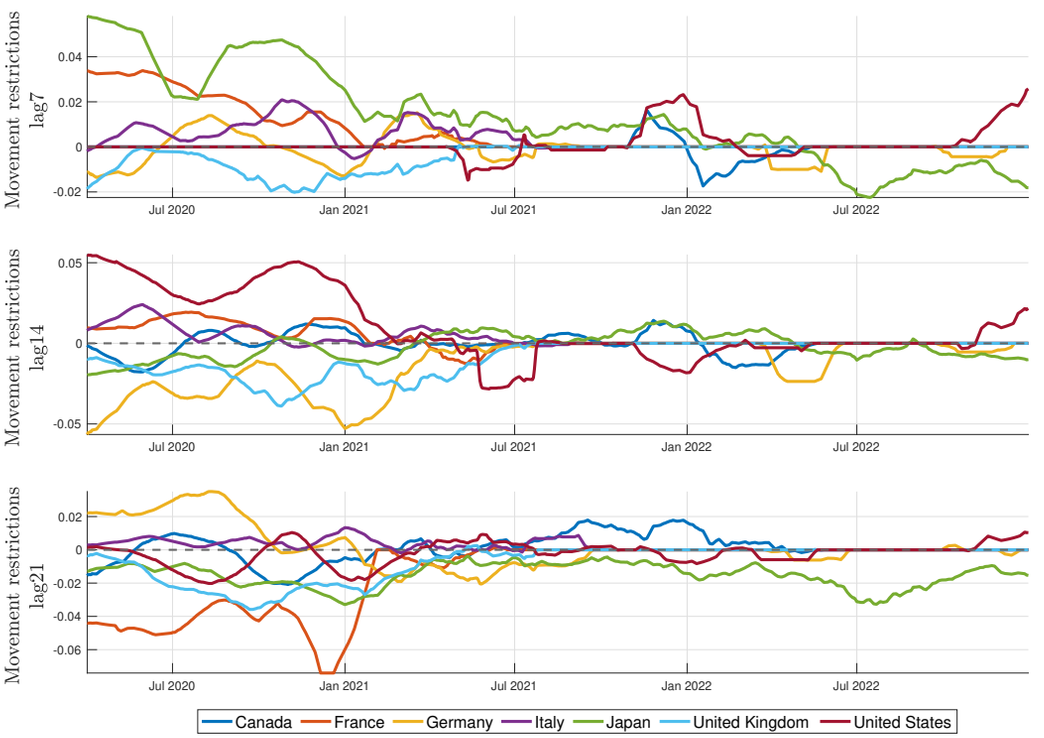


Figure C.5: Partial derivatives: The effects of restrictions on internal movement between cities/regions on new COVID-19 cases 7, 14, and 21 days after the policy change. The partial derivatives, (11), are computed and presented in rank-normalized units, see (10).

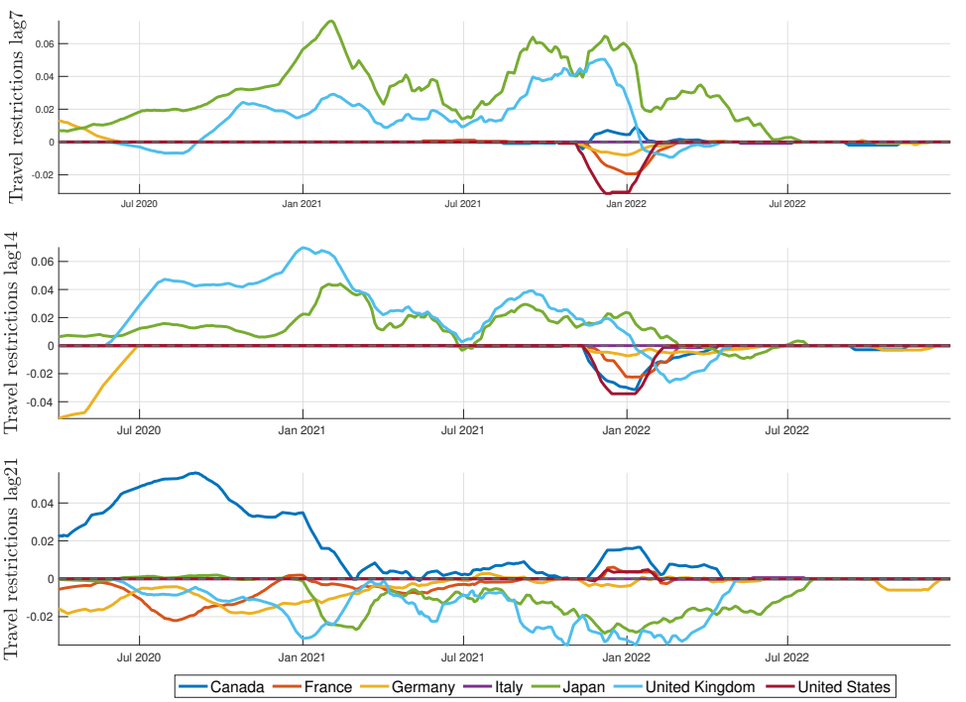


Figure C.6: Partial derivatives: The effects of restrictions on international travel on new COVID-19 cases 7, 14, and 21 days after the policy change. The partial derivatives, (11), are computed and presented in rank-normalized units, see (10). Note: international travel restrictions apply to foreign travellers not citizens.

Figure C.7: Partial derivatives: The effects of public information campaigns on new COVID-19 cases 7, 14, and 21 days after the policy change. The partial derivatives, (11), are computed and presented in rank-normalized units, see (10). Note: no differentiated policies reported in this indicator.

D Additional empirical results for inflation

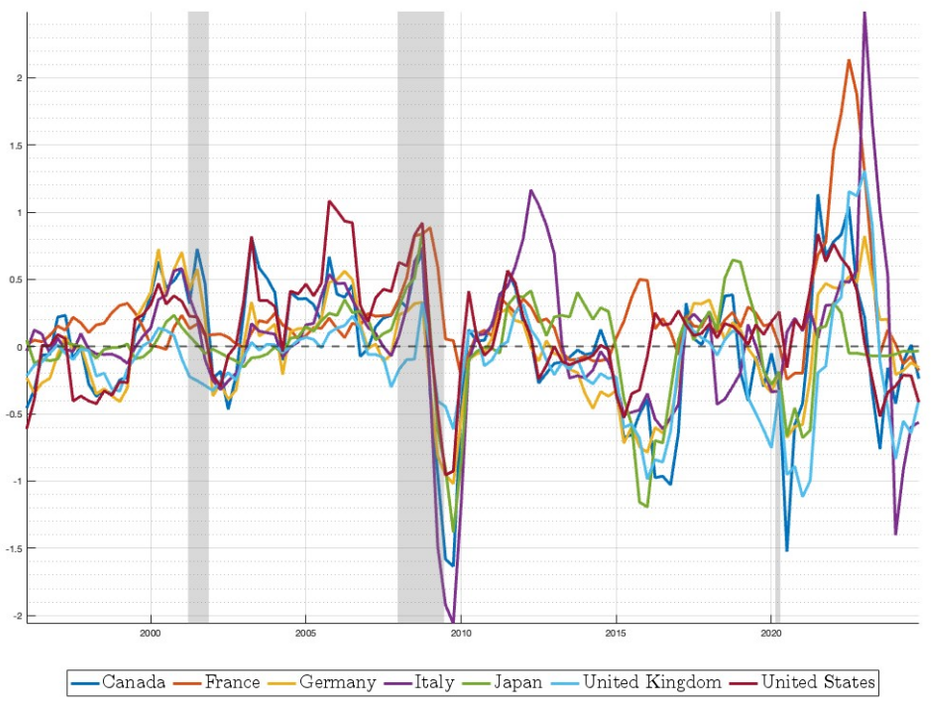
In this section we present supplementary results, as referenced in the main paper, forecasting inflation.

Table D.1: Pre-pandemic results: 2007Q1-2019Q4. RMSE ratios for the 1-, 2-, 4-, and 8-quarters-ahead forecasts of inflation for the deep pooled, deep time-series, and PVAR models.

	Canada	France	Germany	Italy	Japan	UK	US
			$h = 1$				
AR(1)	1.087	1.222	1.319	1.112	1.063	1.014	1.308
Deep pooled	0.505***	0.442***	0.483***	0.604***	0.664	0.636***	0.446**
Deep time-series	0.473***	0.557***	0.487***	0.697***	0.866	0.511***	0.444**
PVAR	1.027**	0.955***	0.965	1.118**	0.966	0.922	1.025*
			$h = 2$				
AR(1)	1.077	1.196	1.268	1.235	1.018	0.931	1.314
Deep pooled	0.527***	0.524***	0.516***	0.520***	0.731	0.743***	0.344**
Deep time-series	0.552**	0.529***	0.476***	0.608***	0.846	0.627***	0.404**
PVAR	1.012*	0.997	0.994	1.035	0.985	1.027	1.001
			$h = 4$				
AR(1)	1.075	1.197	1.268	1.226	1.003	0.944	1.294
Deep pooled	0.457***	0.542***	0.486***	0.573***	0.749	0.733**	0.521**
Deep time-series	0.529***	0.709***	0.429***	0.574***	0.726*	0.605***	0.633**
PVAR	1.000*	1.001	0.999	1.045**	0.999	1.012**	1.002
			$h = 8$				
AR(1)	1.075	1.198	1.268	1.244	1.005	0.955	1.294
Deep pooled	0.554***	0.401***	0.530***	0.628***	0.878	0.778***	0.487**
Deep time-series	0.517**	0.587***	0.461***	0.633***	0.731*	0.693***	0.532**
PVAR	1.001*	1.000	1.000	1.033***	0.998	0.997	1.003*

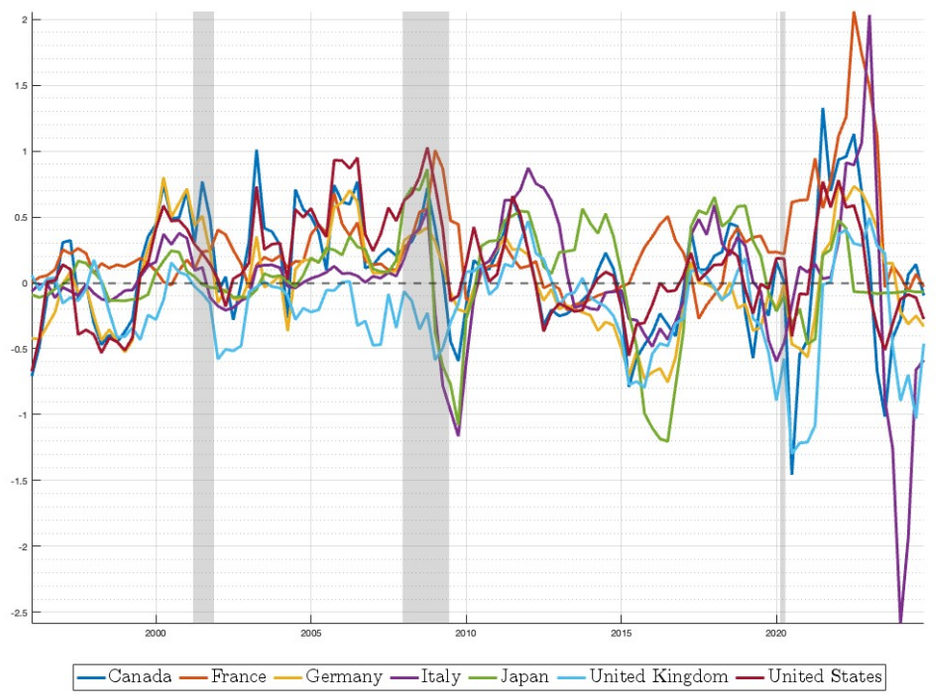
Notes: Ratios are reported relative to the AR(1) model, whose RMSE is presented in absolute terms. Ratios < 1 indicate superior predictive ability relative to the AR(1). $*$, $**$, and $***$ denote rejection of the null hypothesis of equality of forecast mean squared errors at the 10%, 5%, and 1% levels of significance, respectively, using the Diebold and Mariano (1995) test.

Figure D.1: Partial derivatives: The effects of the unemployment rate on the two-quarters-ahead inflation forecasts.



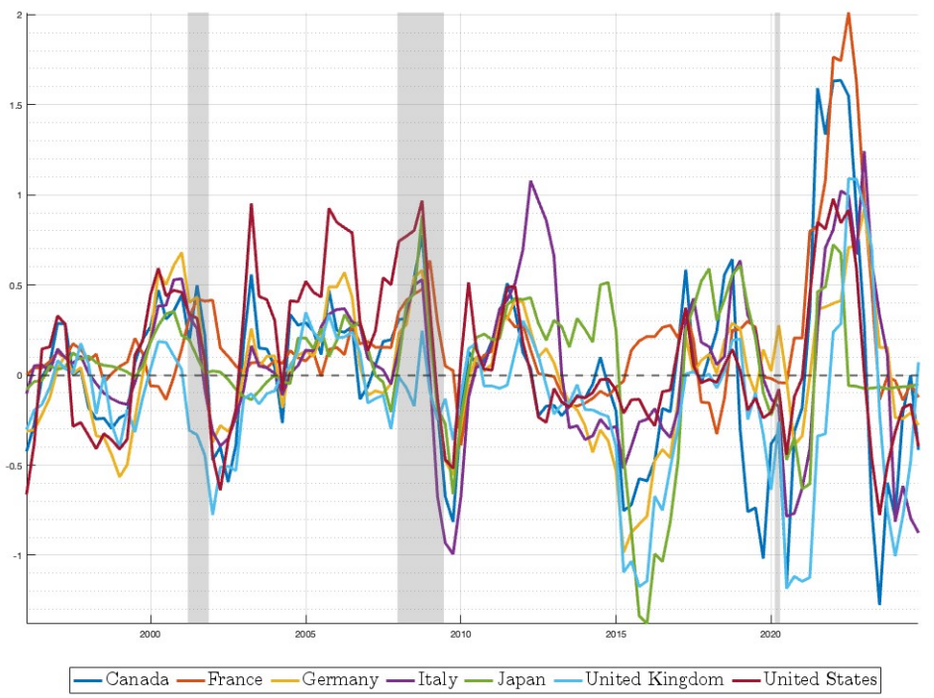
Notes: NBER recession bands for US in gray bars. The partial derivatives, (11), are computed and presented in standardized units.

Figure D.2: Partial derivatives: The effects of the unemployment rate on the four-quarters-ahead inflation forecasts.



Notes: NBER recession bands for US in gray bars. The partial derivatives, (11), are computed and presented in standardized units.

Figure D.3: Partial derivatives: The effects of the unemployment rate on the eight-quarters-ahead inflation forecasts.



Notes: NBER recession bands for US in gray bars. The partial derivatives, (11), are computed and presented in standardized units.

E Neural Network Algorithms

Algorithm 1: ADAM for Stochastic Gradient Descent (SGD)

Result: Final pooled parameter estimate θ_t across all units i

Initialize θ_0, $m_0 = 0$, $v_0 = 0$, $t = 0$;

while θ_t *not converged* **do**

$t \leftarrow t + 1$; $g_t \leftarrow \frac{1}{N} \sum_{i=1}^{N} \nabla_{\theta_{t-1}} L_{it}(\theta_{t-1})$; `// Pooled gradient over all units`

$m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1) g_t$; `// First moment (mean) pooled`

$v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2) g_t \odot g_t$; `// Second moment (variance) pooled`

$\hat{m}_t \leftarrow \frac{m_t}{1-\beta_1^t}$; `// Bias-corrected first moment`

$\hat{v}t \leftarrow \frac{v_t}{1-\beta_2^t}$; `// Bias-corrected second moment`

$\theta_t \leftarrow \theta t - 1 - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}$; `// Update pooled parameters`

end

$\odot$ and $\oslash$ denote element-wise multiplication and division, respectively.