Evaluating Forecast Performance with State Dependence*

Florens Odendahl[1], Barbara Rossi[2], and Tatevik Sekhposyan[3]

[1]*Banco de España*†
[2]*Universitat Pompeu Fabra, Barcelona GSE and CREI*‡
[3]*Texas A&M University*§

May 7, 2021

Abstract

We propose a novel forecast evaluation methodology to assess models' absolute and relative forecasting performance when it is a state-dependent function of economic variables. In our framework, the forecasting performance, measured by a forecast error loss function, is modeled via a hard or smooth threshold model with unknown threshold values. Existing tests either assume a constant out-of-sample forecast performance or use non-parametric techniques robust to time-variation; consequently, they may lack power against state-dependent predictability. Our tests can be applied to relative forecast comparisons, forecast encompassing, efficiency, and, more generally, moment-based tests of forecast evaluation. Monte Carlo results suggest that our proposed tests perform well in finite samples and have better power than existing tests in selecting the best forecast or assessing its efficiency in the presence of state dependence. Our tests uncover "pockets of predictability" in U.S. equity premia; although the term spread is not a useful predictor on average over the sample, it forecasts significantly better than the benchmark forecast when real GDP growth is low. In addition, we find that leading indicators, such as measures of vacancy postings and new orders for durable goods, improve the forecasts of the U.S. industrial production when financial conditions are tight.

Keywords: State Dependence, Forecast Evaluation, Predictive Ability Testing, moment-based Tests, Pockets of Predictability.
JEL codes: C52, C53, E17, G17.

*We thank Lukas Hoesch for useful comments and suggestions. The views expressed herein are those of the authors and should not be attributed to the Banco de España or the Eurosystem. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 615608) and the Spanish Ministry of Economy and Competitiveness, Grant ECO2015-68136-P and FEDER, UE. The Barcelona GSE acknowledges financial support from the Spanish ministry of the Economy and Competitiveness through the Severo Ochoa Programme for Centers of Excellence in R&D (SEV-2015-0063).

†Address: Calle de Alcalá 48, 28014 Madrid, Spain. Email: florens.odendahl@bde.es.
‡Address: c/Ramon Trias Fargas 25/27, 08005 Barcelona, Spain. Email: barbara.rossi@upf.edu.
§Address: 4228 TAMU, College Station, TX 77843, USA. Email: tsekhposyan@tamu.edu.

1 Introduction

In practice, decision-makers face an abundance of candidate forecasting models, and, starting with Diebold and Mariano (1995) and West (1996), the literature has proposed a variety of forecast comparison tests to guide forecasters in choosing the model. However, usually, no single model emerges as the best overall; typically, the forecasting performance is prone to instabilities and, therefore, depends on the sample. One possible explanation is that the economic mechanisms that generate the data are time-varying such that a given model is better in some periods and worse in others, resulting in a state-dependent (or more generally, non-linear) forecasting performance. These empirical findings hold when evaluating the forecasting performance in absolute terms as well.

In this paper, we propose a new forecast comparison test as well as, more generally, moment-based forecast evaluation tests (such as rationality, efficiency, and encompassing) that have power against the alternative of state dependence in the forecasting performance. The state dependence is assumed to take the parametric form of a threshold model, i.e. the relative forecasting performance is a non-linear function of an economic observable variable and a respective threshold. We consider both hard threshold models as well as logistic smooth threshold and exponential smooth threshold models. Importantly, we allow the value of the threshold to be unknown and estimate it alongside the testing procedure. Existing tests either focus on constant relative out-of-sample performance (Giacomini and White, 2006) or use non-parametric techniques to detect time-varying deviations from equal performance (Giacomini and Rossi, 2010; Amisano and Giacomini, 2007); as we show the latter approaches may lack power against the alternative of parametric state dependence.

Our paper is the first to model state dependence in the form of a hard or smooth threshold model directly on the forecasting performance. While Hansen's (1996b) test detects non-linearities *in-sample*, our test instead allows forecasters to evaluate the *out-of-sample* predictive ability when it may be state-dependent. Testing in the presence of an unknown threshold requires non-standard statistics since the nuisance parameter (the threshold) is present only under the alternative; therefore, the standard Wald, Likelihood ratio and Lagrange multiplier tests do not have the usual asymptotic chi-square distribution (Davies, 1977, 1987). While in some cases there might be an economic justification for selecting an ad-hoc threshold value and treating it as known, this is not generally the case, and allowing for an unknown threshold makes our approach broadly applicable. There are several differences between our approach and Hansen's (1996b): for example, when evaluating the relative forecasting performance (i) we apply the threshold model directly to the relative predictive performance, measured by the forecast loss differential, and (ii) we test for a zero expected forecast performance differential, while Hansen (1996b) leaves the expected value unspecified under the null hypothesis. Consequently, we jointly test whether the out-of-sample average relative forecasting performance is different from zero as well as whether it is state-dependent.

While our focus is on the loss differential, the leading evaluation approach in applied work, our methodology more generally applies to other forecast evaluation tests that are implemented with moment conditions. For instance, tests of efficiency and encompassing can be formulated in terms of moment conditions summarizing the information in the relevant forecasts and forecast

errors. We investigate the finite sample performance of our tests with Monte Carlo simulations. Our simulations suggest, for example, that our proposed tests are well-sized and have better power against the alternative of an unequal and state-dependent forecasting performance relative to the existing tests. These results continue to hold even when the state variable is observed with measurement error (see in the Online Appendix).

There are several reasons why considering state dependence in the forecast error losses/loss differentials is interesting. First, it allows the forecaster to impose the null hypothesis directly on the object of interest. While it is true that the researcher can potentially consider state dependence in the forecasting models directly, satisfactory in-sample fit does not necessarily translate to satisfactory out-of-sample performance. Therefore, studying the object of interest (i.e. the forecast error loss) directly is useful. Second, studying forecast error losses/loss differentials makes our framework applicable to the case when the forecasting model is known as well as the case when it is not known (as in widely used survey forecasts). Third, in the case of multivariate models or models with many predictors, the researcher faces the problem of selecting the predictors that are state-dependent; this problem does not arise when directly studying the losses with respect to the target variable of interest. Last, parsimonious linear models are used extensively in the forecasting literature. Our tests allow us to evaluate these models against state dependence in a theoretically coherent framework. The test results guide researcher how to modify the original forecasting models and their estimation technique, as well as on how to select models for prediction at a given point in time.

Our paper contributes to the recent literature on forecast evaluation. (Diebold and Mariano, 1995; West, 1996; Clark and McCracken, 2001; Clark and West, 2006, 2007; Giacomini and White, 2006; Giacomini and Rossi, 2010). In particular, Giacomini and White (2006) (GW henceforth) show the validity of the asymptotic Normal distribution for the out-of-sample test of equal predictive ability proposed by Diebold and Mariano (1995) (DM henceforth) when the underlying forecasting models are estimated using a rolling window estimation scheme and the data satisfy certain mixing properties.[1] Following their framework, our testing procedure similarly relies on a rolling window estimation scheme to preserve the parameter estimation error asymptotically. Hence, we compare forecasting methods rather than forecasting models. However, while GW focus on the null hypothesis of an equal out-of-sample predictive ability, on average, our test allows for state dependence on the conditioning variables, i.e. we test for deviations from the null hypothesis in sub-samples identified by state variables. Importantly, we do not require the conditioning variable itself to explain the forecasting performance (although it could) but only to indicate the state, i.e. the magnitude of the predictive ability within a state can be independent of the conditioning variable. Our paper is also related to Giacomini and Rossi (2010), which allow the relative forecasting performance to be prone to instabilities, using a non-parametric time-variation approach based on the rolling/recursive window estimation of a local GW test. As a result, their test has good power against smooth and persistent changes but, as we show, it might lack power against the switches of a state dependent model.

Several papers in the literature (Stock and Watson, 2009; Rapach et al., 2010; Neely et al., 2014; Dotsey et al., 2018; Granziera and Sekhposyan, 2019) evaluate forecast performance in subsamples, where the subsamples are identified conditional on some economic variable being smaller or

[1] Hereafter, we refer to the DM test under the conditions of Giacomini and White (2006) as the GW test.

larger than an ad-hoc threshold. For instance, Stock and Watson (2009) analyze the forecasting ability of Phillips curve models for U.S. inflation and conclude, from plotting the loss differential, that the forecasting performance depends on the unemployment gap: "when the unemployment gap exceeds 1.5 in absolute value, the Phillips curve forecasts improve substantially upon the UC-SV [unobserved component with stochastic volatility] model." (Stock and Watson, 2009). Rapach et al. (2010) investigate equity return predictability during different states of the business cycle by using the"good, normal, and bad times" classification of GDP growth from Liew and Vassalou (2000). Our methodology, instead, systematically tests for potential state dependence without having to know or assume the threshold value.

We demonstrate the usefulness of our methodology in two empirical applications. First, we compare models that predict U.S. equity premia from 1966 to 2011. As noted in Pesaran and Timmermann (1995) and Rapach and Wohar (2006), financial return predictability is typically time-varying and appears only in sub-samples.[2] Instabilities in forecasting performances in other financial variables are widespread as well: Paye and Timmermann (2006), for instance, cannot reject the presence of structural breaks in stock return predictive regressions and Rossi (2006, 2013b) finds similar results for exchange rate returns. As summarized in Timmermann (2008), "... there appear to be pockets in time where there is modest evidence of local predictability; (...) the best forecasting method can be expected to vary over time, and there are likely to be periods of model breakdown where no approach seems to work". In our empirical results, we do find evidence of state-dependent predictive ability. When forecasting stock market returns, we show the usefulness of our test statistic for detecting pockets of predictability. Furthermore, our approach can shed light on which factors create such pockets. More in detail, our benchmark model is an in-sample mean, re-estimated in real-time in rolling windows, whereas the competitor models use the financial variables from Goyal and Welch's (2008) comprehensive dataset of predictors. We find evidence of state dependence in the relative forecasting performance, where the state dependence is a function of the business cycle, measured by the monthly real GDP growth estimate of Koop et al. (2020): in periods of above-average GDP growth, the economic model tends to underperform relative to the benchmark model. However, in periods of low growth, forecasting with the term spread, defined as the difference between the long-term and the T-bill yields, leads to forecast improvements. On the other hand, the GW and Fluctuation tests cannot reject the null hypothesis of equal forecasting ability and fail to uncover such pockets of predictability.[3]

Second, we evaluate for U.S. industrial production forecasts from January 1971 to December 2019. The choice of our state variable, the adjusted National Financial Conditions Index (ANFCI) computed by the Chicago Fed, is motivated by the work of Adrian et al. (2019). They demonstrated the importance of financial conditions for forecasting the distribution of output growth, particularly tail risk, advocating for a non-linear relationship between financial stability and macroeconomic performance. We find that certain leading indicators such as measures of vacancy postings and new orders for durable goods are particularly useful (relative to parsimonious autoregressive benchmarks) for forecasting U.S. industrial production when financial conditions, measured by ANFCI, are tight.

[2] See Goyal and Welch (2003, 2008) for a related discussion.

[3] Note that the forecasting gains using the financial predictors are small and that any large deviations from equal predictive ability in favor of the economic models would imply strong violations of the rational expectations hypothesis.

Our paper further is related to Harvey et al. (2021) and Inoue and Rossi (2015), who also propose methodologies to study time-variation in predictive regressions. There are several differences with their methodologies, however. As indicated before, our tests are designed for out-of-sample forecast evaluation, while Harvey et al. (2021) and Inoue and Rossi (2015) consider testing the significance of a specific predictor in-sample, which does not necessarily imply better out-of-sample performance. Most importantly, our method allows the researcher to shed light on the economic causes behind the changes in the models' predictive ability since the variation in the forecasting performance is linked to the economic variables that define the states. On the other hand, Harvey et al. (2021) and Inoue and Rossi (2015) propose a sequential procedure as opposed to the one-shot evaluation proposed in this paper. The sequential procedure is tailored to monitor the predictive performance in real-time. Our one-shot procedure, on the other hand, evaluates the out-of-sample forecasting performance historically, yet still allows picking the best performing model depending on the identified state at the end of the sample.

The paper is organized as follows. Section 2 formalizes our null hypothesis, introduces our test statistics, and describes the challenges that arise when testing for state dependence in relative forecasting performance. Section 3 evaluates size and power of our proposed procedure in finite samples via Monte Carlo simulations. Section 4 investigates the existence of pockets of predictability in financial data and Section 5 investigates state-dependence in the relative forecast performance of models predicting U.S. industrial production. Section 6 concludes.

2 Testing for state dependence: methodology

We first describe the model and the null hypothesis. We further provide illustrative examples on how state-dependence in forecast losses can arise. Then, we introduce the necessary notation, the technical assumptions, and the test statistic.

2.1 The General Framework

Let $\widehat{f}^{(1)}_{t+h|t}(A_t, A_{t-1}, ..., A_{t-R+1}; \widehat{\beta}^{(1)}_{t,R})$ and $\widehat{f}^{(2)}_{t+h|t}(A_t, A_{t-1}, ..., A_{t-R+1}; \widehat{\beta}^{(2)}_{t,R})$ denote two measurable functions, which provide the forecasts of two competing models, labeled (1) and (2), where t denotes the forecast origin, h denotes the forecast horizon, and the vector of stochastic processes $A_t = (Y_t, Z_t)$ contains the variable of interest Y_t and the column vector of predictors Z_t. In turn, $\widehat{\beta}^{(i)}_{t,R}$ denotes the vector of estimated parameters at time t of model "i"$(i = 1, 2)$ using a rolling window estimation scheme of size $R \leq \bar{R} < \infty$ and data $A_t, ..., A_{t-R+1}$.[4] Henceforth, we simply write $\widehat{f}^{(1)}_{t+h|t}$ and $\widehat{f}^{(2)}_{t+h|t}$. Importantly, note that the function $\widehat{f}^{(i)}_{t+h|t}$ can denote either a point or a density forecast.

Let $L_{t+h|t}\left(Y_{t+h}, \widehat{f}^{(i)}_{t+h|t}\right)$ denote a loss function, which evaluates the prediction $\hat{f}^{(i)}_{t+h|t}$ of Y_{t+h}. The loss functions we allow for are quite general and encompass the quadratic loss (which gives rise to a Mean Squared Forecast Error (MSFE) measure of predictive ability), asymmetric losses (such as the lin-lin loss), as well as the log score and Continuous Rank Probability Score (CRPS) for density forecasts. We define the forecast error loss of interest to the researcher as $\mathcal{L}_{t+h|t}$. Note that $\mathcal{L}_{t+h|t}$ denotes our object of interest, which can be different from the loss function $L_{t+h|t}(\cdot)$ itself.

[4] The window size R is assumed to be the same across the two models for notational convenience only.

For instance, in the case of relative predictive ability, $\mathcal{L}_{t+h|t}$ is the forecast error loss differential, while in the case of forecast efficiency, this is the relevant moment condition; we describe this in more detail below. Note that $\mathcal{L}_{t+h|t}$ is a function of the estimated parameters $\widehat{\beta}_{t,R}^{(i)}$ and the rolling window size R. As we assume that the parameters are estimated over a rolling and finite window size, the loss differential compares forecasting methods rather than forecasting models.

We allow the forecast error loss to evolve over time according to a non-linear model (Teräsvirta, 2006):

$$\mathcal{L}_{t+h|t} = X_t'\mu + X_t'\theta \cdot G(S_t;\varphi) + u_{t+h}, \tag{1}$$

where X_t and S_t are explanatory variables, φ is a vector of parameters, u_{t+h} is an error term and $G(\cdot)$ is allowed to be a non-linear function. In eq. (1), μ and θ denote the parameters of interest, the vector X_t is a k_1 dimensional column vector that denotes economic observables and a constant, S_t denotes the economic observable that introduces the state dependence, φ denotes the unknown threshold, u_{t+h} is the error term. For the remainder, S_t is assumed to be a scalar. In Appendix A.2 we discuss the possibility of several candidate variables for S_t and how to extend the testing procedure to account for that. Potential serial correlation can be accounted for by including lags of $\mathcal{L}_{t+h|t}$, which are allowed, but not required, to also be a function of the threshold indicator. S_t is a stochastic process, assumed to be continuous and allowed to be a subvector of X_t.

The non-linear model in equation (1) allows for a wide range of models which includes time-varying parameter models as long as the time-variation is a parametric function of an observable S_t. Note that our framework does not allow S_t to be unobservable; for example, this rules out the case of Markov switching models which we separately discuss in the Online Appendix. We also show in Monte Carlo simulations that our test has power in cases where only a noisy measure, $\tilde{S}_t$, of S_t is available; for instance, in the case where $\tilde{S}_t$ is an estimate of the true but unobserved variable S_t.

In particular, the model classes we consider encompass several interesting cases for $G(S_t;\varphi)$. The first case is a threshold regression (TR) model:

$$\text{TR:} \quad G(S_t;\varphi) = \mathbb{1}\left(S_t \geq \gamma\right), \quad \text{where} \quad \varphi = \gamma, \tag{2}$$

i.e. the effect of X_t on $\mathcal{L}_{t+h|t}$ changes if S_t is above the threshold γ. The second case is the logistic smooth threshold regression (LSTR) model:

$$\text{LSTR:} \quad G(S_t;\varphi) = \left(1+\exp\{-\tau(S_t-\gamma)\}\right)^{-1}, \quad \text{where} \quad \varphi = (\gamma,\tau), \tag{3}$$

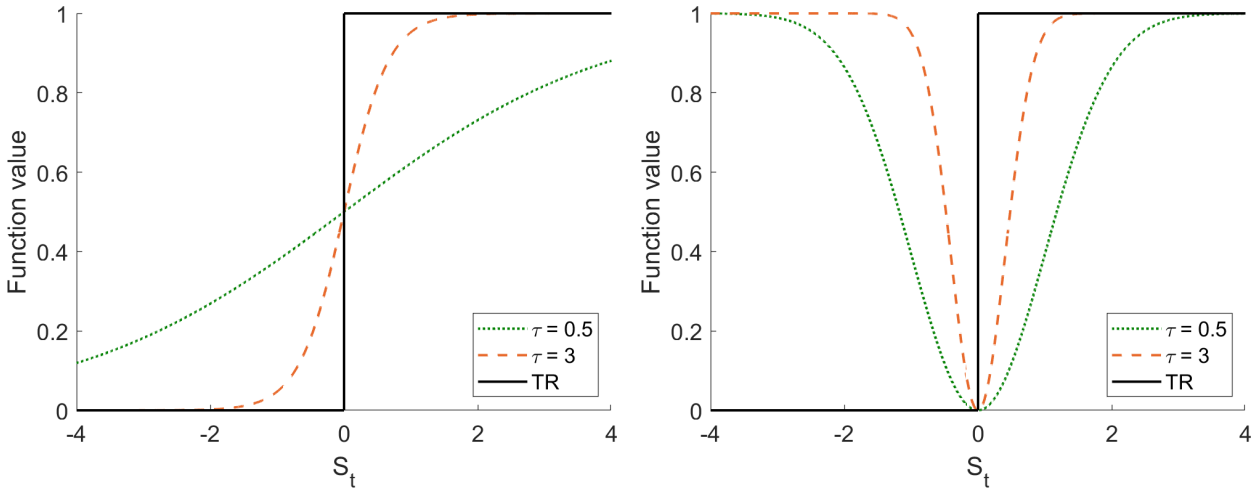
with $0 < \tau < \infty$ and the effect of X_t on $\mathcal{L}_{t+h|t}$ changes smoothly if S_t is either above or below a threshold γ and the smoothness of the function is controlled by τ. The third case is an exponential smooth threshold regression (ESTR) model:

$$\text{ESTR:} \quad G(S_t;\varphi) = 1-\exp\{-\tau(S_t-\gamma)^2\}, \quad \text{where} \quad \varphi = (\gamma,\tau), \tag{4}$$

with $0 < \tau < \infty$ and the effect of X_t on $\mathcal{L}_{t+h|t}$ changes smoothly if S_t is either above or below a threshold γ, where τ controls the smoothness of the change. We give more details below about how to choose the grid for γ. In the following, we refer to a smooth threshold regression model (STR) whenever the function $G(\cdot)$ has no discontinuity.

Figure 1 plots the functional forms of the TR, LSTR and ESTR models. The TR and LSTR are similar, with the difference being that the LSTR is smooth and the TR has a kink at γ. Both models are useful in cases where the delta losses change when the threshold indicator variable is larger (smaller) than γ. For instance, Rapach et al. (2010) find pockets of predictability for U.S. equity premia when U.S. GDP growth is low. The ESTR instead is most useful whenever threshold effects are present for deviations from an "equilibrium condition", independently of the direction. For instance, Stock and Watson (2009) conclude from visual inspection that the Phillips curve model outperforms a benchmark model whenever the unemployment gap deviates strongly from its steady state value of zero, i.e. when the unemployment gap is particularly large *or* small.

Figure 1: Functional forms of TR, LSTR, and ESTR



(a) TR vs LSTR (b) TR vs ESTR

Note: Panel (a) plots the LSTR, eq. (3), for two different values of τ against the TR, eq. (2). Panel (b) plots the ESTR, eq. (4), for two different values of τ against the TR. γ is set equal to zero in both plots. The y-axis denotes the function value, $G(S_t;\varphi)$. The x-axis denotes the value of S_t.

We aim at evaluating forecasting models' predictive performance while being able to detect possible additive non-linearities in the form of a hard or smooth threshold model. Our null hypothesis of equal predictive ability at each point in time is:

$$\mathrm{E}\big(\mathcal{L}_{t+h|t}\big) = 0 \quad \forall t, \tag{5}$$

versus the alternative

$$\mathrm{E}\big(\mathcal{L}_{t+h|t}|X_t, S_t\big) = X_t'\mu + X_t'\theta \cdot G(S_t;\varphi). \tag{6}$$

The null and alternative hypotheses involve μ and θ and become $\mathrm{H}_0 : \mu = \theta = 0$ and $\mathrm{H}_\mathrm{A} : \mu \neq 0,\ \theta \neq 0$ respectively. Note that the null hypothesis defined in eq. (5) holds conditionally on X_t and S_t, and, therefore, by the law of iterated expectations, also *unconditionally*. Our test has power against either μ or θ or both jointly deviating from zero under the alternative, i.e. either a constant non-equal predictive ability or a state-dependent (or non-linear) predictive ability or both.[5] Importantly, we allow the nuisance parameter φ to be unknown. Therefore, testing for the

[5]Note that the case of $\mu = \theta \neq 0$ is a valid alternative and merely represents the joint presence of a non-equal and

null hypothesis described in equation (5) is subject to the problem of a nuisance parameter that is present only under the alternative, which makes standard asymptotic inference invalid (Davies, 1977, 1987; Hansen, 1996b).

Before describing our proposed test statistics, we want to emphasize two points. First, although the assumption of an unknown φ comes at the cost of non-standard inference, it brings the large benefit that it allows the researcher to test over a range of values, instead of having to choose an arbitrary value. This is particularly important in practice because an ad-hoc choice for φ can be detrimental to the power of detecting state dependence. In practice, when working with the TR, LSTR or ESTR model, we recommend to formulate γ in terms of the empirical distribution function $\Xi_n(\cdot)$ of S_t such that the indicator becomes $\mathbb{1}\big(\Xi_n(S_t) \geq \gamma\big)$, with $\gamma \in \Gamma = [0,1]$ and $\Xi_n^{-1}(\gamma)$ provides the threshold in units of S_t (Hansen, 1996b). This is particularly useful when implementing the model in statistical programs, as it allows formulating a unit-free grid for γ. Following Hansen (1996b) and others, we restrict γ to be away from the boundaries and choose, for instance, $\Gamma = [0.15, 0.85]$. Therefore, when we express the bounds of the threshold value in terms of S_t, its lower and upper bounds are $\underline{S}_t = \Xi_n^{-1}(0.15)$ and $\bar{S}_t = \Xi_n^{-1}(0.85)$ respectively. In other words, restricting the values of γ to $\Gamma = [0.15, 0.85]$ refers to the lower 15th and upper 85th percentile of the empirical cdf of S_t; importantly, the restriction of $\Gamma = [0.15, 0.85]$ does not refer to a time index where we have to leave the endpoints out. Thus, a detection of a change of the state is possible in real-time using our method.

Tests of equal predictive ability: Tests of equal predictive ability are implemented by letting

$$\mathcal{L}_{t+h|t} = \Delta L_{t+h|t} \equiv L_{t+h|t}\left(Y_{t+h|t}, \widehat{f}^{(1)}_{t+h|t}\right) - L_{t+h|t}\left(Y_{t+h|t}, \widehat{f}^{(2)}_{t+h|t}\right). \tag{7}$$

The following specification of eq. (1) is of particular interest in the forecast comparison case, as it specifies state dependence that is a function solely of S_t and does not depend on any additional observables X_t:

$$\Delta L_{t+h|t} = \mu + \theta \cdot G(S_t; \varphi) + u_{t+h}. \tag{8}$$

The specification in eq. (8) encompasses the standard Diebold and Mariano (1995) and Giacomini and White (2006) tests for equal predictive ability as special cases, and, unlike the latter, is capable of detecting periods of unequal performance that depend on S_t.

Tests of forecast encompassing: While our leading case is the loss differential, as it is widely used applied work, our methodology can also be applied to the moment conditions of forecast encompassing tests. Let $\epsilon_{t+h|t,1} = Y_{t+h} - \widehat{f}^{(1)}_{t+h|t}$ and $\epsilon_{t+h|t,2} = Y_{t+h} - \widehat{f}^{(2)}_{t+h|t}$. To test whether model (1) encompasses model (2), we define

$$\mathcal{L}_{t+h|t} = ENC_{t+h|t} \equiv \epsilon^2_{t+h|t,1} - \epsilon_{t+h|t,1}\epsilon_{t+h|t,2}. \tag{9}$$

Then, to test our null hypothesis that the forecast of model one encompasses the forecast of model two, the following specification is of particular interest:

$$ENC_{t+h|t} = \mu + \theta \cdot G(S_t; \varphi) + u_{t+h}. \tag{10}$$

non-linear predictive ability.

As in the case of the loss differential in eq. (8), the null and alternative hypothesis in the forecast encompassing test in eq. (10) involve μ and θ and become $\mathrm{H}_0 : \mu = \theta = 0$ and $\mathrm{H}_\mathrm{A} : \mu \neq 0,\ \theta \neq 0$ respectively.

Tests of forecast optimality. Tests of forecast optimality include tests of forecast unbiasedness and efficiency. In those cases, only the forecast error loss of one model is evaluated (e.g. model i). Tests of forecast unbiasedness and forecast efficiency, respectively, can be formulated as a moment-based test by defining

$$\mathcal{L}_{t+h|t} = UB_{t+h|t} \equiv y_{t+h} - \widehat{f}^{(i)}_{t+h|t}, \quad \text{and} \quad \mathcal{L}_{t+h|t} = FE_{t+h|t} \equiv \left(y_{t+h} - \widehat{f}^{(i)}_{t+h|t}\right) \cdot \widehat{f}^{(i)}_{t+h|t} \tag{11}$$

and can be tested using the specification:

$$UB_{t+h|t} = \mu + \theta \cdot G\left(S_t; \varphi\right) + u_{t+h,t} \quad \text{and} \quad FE_{t+h|t} = \mu + \theta \cdot G\left(S_t; \varphi\right) + u_{t+h,t}. \tag{12}$$

In both cases, the null hypothesis of unbiasedness and efficiency, respectively, is implemented by $\mu = \theta = 0$. In the Online Appendix, we show simulation results for forecast efficiency tests.

2.2 Examples of thresholds in the loss differential

In this subsection, we discuss two simple analytical examples of how threshold-type effects in the losses can arise.

A first example is based on a DGP that includes a common component. In particular, two observable variables, y_t and x_t, are driven by a common component, c_t, and unpredictable idiosyncratic components, e_t and η_t respectively:

$$y_{t+1} = \alpha + c_{t+1} + e_{t+1}, \quad x_{t+1} = c_{t+1} + \eta_{t+1}, \quad c_{t+1} = \rho_c c_t + v_{t+1} \tag{13}$$

where α is a constant, $e_{t+1} \sim N(0, \sigma_e^2)$, $\eta_{t+1} \sim N(0, \sigma_\eta^2)$, and e_{t+1}, η_{t+1} and v_{t+1} are mutually independent. Importantly, $v_{t+1} \sim N(0, \sigma^2_{v,t+1})$, i.e. the variance is time-varying with $\sigma^2_{v,t+1} = \sigma_{v,1} + G(S_t; \varphi)\sigma_{v,2}$, where $\sigma_{v,1}, \sigma_{v,2} > 0$. S_t can be interpreted as determining the magnitude of shocks to the common component; another interpretation is that $\sigma^2_{v,t+1}$ is unobservable and S_t is a proxy variable that is indicating the strength of the common component. The two forecasting models are

$$f^{(1)}_{t+1|t} = \alpha \quad \text{and} \quad f^{(2)}_{t+1|t} = \alpha + x_t, \tag{14}$$

i.e. the common component is unknown to the forecasters and we abstract from parameter estimation error for simplicity. The expected value of the loss differential for the one-step-ahead forecast in this case takes the form

$$E[\Delta L_{t+1|t}] = E[\epsilon^2_{t+1|t,1} - \epsilon^2_{t+1|t,2}] = (2\rho_c - 1)\rho_c^2\sigma_c^2 + (2\rho_c - 1)E[\sigma_{v,1} + G(S_t; \varphi)\sigma_{v,2}] + \sigma_\eta^2, \tag{15}$$

i.e. the loss differential is a non-linear function of S_t and φ through the term $E[\sigma_{v,L} + G(S_t; \varphi)\sigma_{v,H}]$.

As a second example, consider the following DGP:

$$y_{t+1} = \alpha + \phi z_t + \beta G(S_t; \varphi) x_t + e_{t+1}, \tag{16}$$

with $e_t \sim N(0, \sigma_e^2)$, $x_t \sim N(0, \sigma_x^2)$, and S_t and φ are the indicator variable and parameters that govern the non-linear function $G(\cdot)$. The two competing forecasting models are:

$$f_{t+1|t}^{(1)} = \alpha + \phi z_t \quad \text{and} \quad f_{t+1|t}^{(2)} = \alpha + \phi z_t + \beta x_t, \tag{17}$$

where we abstracted from parameter estimation error of the coefficients α, ϕ, and β for simplicity. In this framework, the variable x_t enters non-linearly in the DGP, as a function of S_t and φ, but the non-linear relationship is not accounted for in the forecasting model. The expected value of the one-step-ahead forecast loss differential, the squared forecast error of the first model, $\epsilon_{t+1|t,1}^2$, minus the squared forecast error of the second model, $\epsilon_{t+1|t,2}^2$, is:

$$E[\Delta L_{t+1|t}] = E[\epsilon_{t+1|t,1}^2 - \epsilon_{t+1|t,2}^2] = -2\beta^2\sigma_x^2 E[G(S_t;\varphi)] + \beta^2\sigma_x^2, \tag{18}$$

such that $E[\Delta L_{t+1|t}]$ is a function of S_t and φ. For instance, if $G(S_t;\varphi) = \mathbb{1}(S_t \geq \gamma)$, with $\varphi = \gamma$, then

$$E[\Delta L_{t+1|t}|S_t \leq \gamma] = \beta^2\sigma_x^2 \quad \text{and} \quad E[\Delta L_{t+1|t}|S_t > \gamma] = -\beta^2\sigma_x^2, \tag{19}$$

i.e. the loss differential is a threshold function of S_t and γ. In the case of forecast encompassing, the expected loss takes the form:

$$E[\epsilon_{t+1|t,1}^2 - \epsilon_{t+1|t,1}\epsilon_{t+1|t,2}] = \beta^2\sigma_x^2 E[G(S_t;\varphi)], \tag{20}$$

which is again a non-linear function of S_t and φ.

2.3 Test statistics

The parameter vector φ is an element of the compact set Φ which is a bounded subset of R^q. Let $Q_t(\varphi)$ be a k-dimensional column vector that contains the explanatory variables of the threshold model described in eq. (1), i.e. $Q_t(\varphi) = \left[X_t', \left(X_t \cdot G(S_t;\varphi)\right)'\right]'$, and let $Q_t = \sup_{\varphi\in\Phi}|Q_t(\varphi)|$. Let $\widehat{\psi}(\varphi) = \left[\widehat{\mu}(\varphi)', \widehat{\theta}(\varphi)'\right]'$ denote the vector of OLS parameter estimates under the alternative, and let $\widehat{u}_{t+h} = \mathcal{L}_{t+h|t} - Q_t(\varphi)'\widehat{\psi}(\varphi)$ denote the error term under the alternative. The score under the alternative is then given by $\widehat{s}_{t+h}(\varphi) = Q_t(\varphi)\widehat{u}_{t+h}(\varphi)$. Let H_r denote a restriction matrix that corresponds to the null hypothesis defined in eq. (5). For instance, for the model described in eq. (8) without any additional regressors, we have that $H_r = I_2$, where I_2 is a two-dimensional identity matrix.[6] Let T denote the total sample size and $P = T - R - h$ denote the out-of-sample size, i.e. the number of observations of $\mathcal{L}_{t+h|t}$. Let $\widehat{V}_P(\varphi) = \frac{1}{P}\sum_{t=R}^{T-h}\widehat{s}_{t+h}(\varphi)\widehat{s}_{t+h}(\varphi)'$ denote the variance-covariance matrix of the score, let $V(\varphi) = \mathrm{E}\left(s_{t+h}(\varphi)\, s_{t+h}(\varphi)'\right)$ be finite and positive definite for $s_{t+h}(\varphi) = Q_t(\varphi)u_{t+h}$, and let $\widehat{V}_P^*(\varphi) = M_P(\varphi,\varphi)^{-1}\widehat{V}_P(\varphi)M_P(\varphi,\varphi)^{-1}$ be the robust estimator of the variance-covariance matrix of $\widehat{\psi}$, with $M_P(\varphi_1,\varphi_2) = \frac{1}{P}\sum_{t=R-h}^{T} Q_t(\varphi_1)Q_t(\varphi_2)'$, and $M(\varphi_1,\varphi_2) = \mathrm{E}\big(Q_t(\varphi_1)Q_t(\varphi_2)'\big)$; also, let $K_P(\varphi_1,\varphi_2) = \frac{1}{P}\sum_{t=R}^{T-h} s_{t+h}(\varphi_1)s_{t+h}(\varphi_2)'$.

We consider the following test statistics, based on Hansen (1996b) and Andrews and Ploberger

[6] If eq. (1) contains additional control variables, such as lags of $\mathcal{L}_{t+h|t}$, that are not part of the forecast comparison null hypothesis, the restriction matrix will not be equal to an identity matrix.

(1994), which we collectively refer to as the DM^{NL} test:

$$\text{DM}^{\text{NL}}: \; g_{\Phi}(W_P) = \begin{cases} \sup_{\varphi \in \Phi} W_P(\varphi) & (\text{“sup-W”}) \\ \int_{\Phi} W_P(\varphi) \mathrm{d}w(\varphi) & (\text{“ave-W”}) \\ \ln\big(\int_{\Phi} \exp(\frac{1}{2} W_P(\varphi)) \mathrm{d}w(\varphi)\big) & (\text{“exp-W”}) \end{cases} \tag{21}$$

where $w(\varphi)$ is a weighting function[7] over $\varphi \in \Phi$, $\ln(\cdot)$ denotes the natural logarithm and $W_P(\varphi)$ is defined as

$$W_P(\varphi) = P\widehat{\psi}(\varphi)' H_r \big[H_r' \widehat{V}_P^*(\varphi) H_r\big]^{-1} H_r' \widehat{\psi}(\varphi). \tag{22}$$

Henceforth, we let $g_{\Phi}\big(W_P(\varphi)\big)$ denote either of the three above mentioned functions, i.e. sup-W, exp-W, and ave-W.

Establishing the uniform convergence of our test statistic requires an empirical process central limit theorem (CLT) such that: (i) the regression score, $\psi(\cdot)$, is unbounded (since, for instance, u_t is unbounded) and, (ii), functional forms for $G(\cdot;\varphi)$ are non-smooth in φ (since we include a threshold model which is discontinuous around the threshold). As pointed out by Andrews (1993) and Hansen (1996b,c), the work of Doukhan et al. (1995) provides an adequate empirical process CLT. Other work, for instance, Andrews (1991) and Hansen (1996c), do not require strict stationary but impose smoothness conditions on the function $G(\cdot;\varphi)$ that are violated by the discontinuity of the threshold model. In turn, that means that the regularity conditions stated below are somewhat stricter than required for the smooth threshold models since both Andrews (1991) and Hansen (1996c) provide milder assumptions for an empirical process CLT of smooth functions $G(\cdot;\varphi)$. We derive the limiting distribution of DM^{NL} under the following assumptions:

Assumption A1 *(i) (A_t, X_t, S_t) is strictly stationary and absolutely regular with mixing coefficients $\eta(m) = O(m^{-\delta})$ for some $\delta > v/(v-1)$ and $v > 1$. (ii) The estimation window size (R) is finite and the estimation scheme is a rolling window estimation.*

Assumption A2 *For $r > v > 1$, $E|Q_t|^{4r} < \infty$, $E|u_t|^{4r} < \infty$, and $\inf_{\varphi \in \Phi}$ $det\big(M(\varphi,\varphi)\big) > 0$.*

Assumption A3 *(i) Let $\xi_t(\varphi) \equiv X_t' G(S_t;\varphi)$; for some $B < \infty$ and $\lambda > 0$, $\big|\big|\big(\xi_t(\varphi_1) - \xi_t(\varphi_2)\big)u_{t+h}\big|\big| < B\big|\big|\varphi_1 - \varphi_2\big|\big|^{\lambda}$. (ii) $M_P(\varphi_1,\varphi_2)$ and $K_P(\varphi_1,\varphi_2)$ converge almost surely to $M(\varphi_1,\varphi_2)$ and $K(\varphi_1,\varphi_2)$, uniformly in $\varphi_1, \varphi_2 \in \Phi$.*

Assumption A4 *$f^{(i)}_{t+h|t}(.)$ is a measurable function of lags of A_t, for $i = 1, 2$.*

A1 limits the dependence and time-variation allowed in the loss differential under the null. A2 ensures that the explanatory variables in eq. (1) has more than $4r$ finite moments and that the variance-covariance matrix of X_t and S_t is non-singular for all φ. A3 (i) imposes a continuity assumption on the element of the score associacted with the non-linear function $G(\cdot)$, and A3 (ii) ensures uniform convergence of $M_P(\cdot)$ and $K_P(\cdot)$ over $\varphi in \Phi$. A4 is an assumption on the functional form of the point forecast itself, and ensures measurability of $\mathcal{L}_{t+h|h}$. Then, the asymptotic distribution in eq. (21) can be described as follows.

[7] Throughout the paper we use an equal weighting, i.e. $w(\varphi) = \varphi$.

Proposition 1 *Let $g_\Phi(W_P)$ be either $\sup_{\varphi\in\Phi} W_P(\varphi)$, $\int_\Phi W_P(\varphi)dw(\varphi)$ or $ln\left(\int_\Phi exp(\frac{1}{2}W_P(\varphi))dw(\varphi)\right)$, where Φ is compact and $W_P(\varphi) = P\widehat{\psi}(\varphi)'H_r\left[H_r'\widehat{V}_P^*(\varphi)H_r\right]^{-1}H_r'\widehat{\psi}(\varphi)$, and $\widehat{\psi}(\varphi) = \left[\widehat{\mu}(\varphi)', \widehat{\theta}(\varphi)'\right]'$ is estimated from eq. (1). Then, under A1 to A4 and H_0 defined in eq. (5): $E\left(\mathcal{L}_{t+h|t}\right) = 0$ for all $t = R+h, ..., T$ and*

$$\lim_{P\to\infty} g_\Phi\left(W_P(\varphi)\right) \underset{d}{\to} g_\Phi\left(\chi^2(\varphi)\right), \tag{23}$$

where $\chi^2(\varphi)$ is a chi-square distribution with degrees of freedom $rank(H_r)$, and $g_\Phi\left(\chi^2(\varphi)\right)$ can be completely characterized by its covariance kernel $K(\varphi_1, \varphi_2)$. The test rejects H_0 defined in eq. (5) when $g_\Phi\left(W_P(\varphi)\right) > \phi_\alpha$, where ϕ_α is the critical value (for a nominal size of α) that can be simulated according to Algorithm 1 below.

The proof of Proposition 1 is provided in Appendix A.1.

2.4 Practical implementation

The asymptotic distribution in eq. (40) is not nuisance parameter free and cannot be tabulated except for special cases.[8] Therefore, we follow Hansen (1996b) to propose an algorithm that can be used to simulate the critical values and which we report here for the readers' convenience. Loss differentials may exhibit serial correlations since it is a function of forecast errors, which are serially correlated for $h > 1$. To deal with potentially unaccounted serial correlation in the loss differential, i.e. an autocorrelated u_{t+h} in eq. (1), we adjust the original algorithm for simulating the asymptotic distribution proposed by Hansen (1996b). The adjustment is based on a suggestion of Hansen (1996a) in the context of a test for Markov switching models.

Simulation Algorithm 1 Let $\widehat{s}_{t+h}(\varphi)$, $M_P(\varphi,\varphi)$, $\widehat{V}_P^*(\varphi)$, and H_r be as defined in Section 2.3. Let $B = (4(P/100)^{(2/9)} + 1)$ be the bandwidth parameter of the Bartlett kernel used in the simulation algorithm. Then, for each $j = 1, ..., J$ do the following steps:

1. Draw a set of standard Normal random variates $\{v_{t,j}\}_{t=R}^{T-h+B}$:
   (a) Calculate $\widehat{\lambda}_P^j(\varphi) = \frac{1}{\sqrt{P}}\frac{1}{\sqrt{1+B}}\sum_{b=0}^{B}\sum_{t=R}^{T-h}\widehat{s}_{t+h}(\varphi)v_{t+b,j}$;
   (b) Calculate $W_P^j(\varphi) = \widehat{\lambda}_P^j(\varphi)'M_P(\varphi,\varphi)^{-1}H_r\left[H_r'\widehat{V}_P^*(\varphi)H_r\right]^{-1}H_r'M_P(\varphi,\varphi)^{-1}\widehat{\lambda}_P^j(\varphi)$;
   (c) Repeat (a)-(b) for all $\varphi \in \Phi$;
2. Compute $W_P^j = g_\Phi\left(W_P^j(\varphi)\right)$.

For instance, for the case of the threshold model with $G(S_t;\varphi) = \mathbb{1}(S_t \leq \gamma)$ the algorithm iterates over different values of $\gamma \in \Phi$. In the case of a smooth threshold model, the algorithm iterates over different values of the pair $\varphi = (\gamma, \tau) \in \Phi$.

The adjustment for serial correlation in the simulation of the asymptotic distribution does not specify a specific process for the serial correlation and is, therefore, suited for a variety of possible autocorrelations structures in the loss differential. If the researcher has reason to believe that there is no autocorrelation, for instance, in the case of $h = 1$, she can set $B = 0$ which reduces Simulation Algorithm 1 to the original algorithm proposed in Hansen (1996b).

[8] See Hansen (1996b) for a discussion.

2.5 Practical suggestions about what to do after applying our method

Our analysis focuses on a loss differential approach to formally test predictive ability. The reason why we focus on the loss differential is because our goal is to directly evaluate the out-of-sample performance of the forecasts based on the loss function preferred by the researcher. There are several advantages in working directly with the loss function approach to model evaluation, as opposed to testing the model's parameters. One advantage is that in our framework we can directly analyze the predictive performance even when the underlying forecasting models are unknown, for instance, in the case of survey data, Greenbook forecasts, or the forecasts published by institutions. A second advantage is that estimating a threshold regression requires some potentially complicated choices regarding the model specification. For instance, in a model with many predictors, the researcher has to decide which variable is subject to the threshold effect and which one isn't. This problem does not arise when testing for non-linearities directly in the forecast error losses.

On the other hand, when our test rejects the null hypothesis, it is important to analyze why. In fact, when our test finds empirical evidence of threshold effects in the relative forecasting performance, this could be due to either the fact that a model performing worse/better while the performance of the other model remained the same, or that both models forecasting performance deteriorated/improved. To further investigate the reasons behind the time-varying performance, the researcher can compute the MSFE of each model in the different regimes identified via our methodology. Then, he/she can compute the overall ratio of the MSFEs of the two competing models, as well as ratios in the respective sub-samples identified by our test. While not a formal test, this provides evidence to answer the questions posed above: did one model get worse, or both models but one less so, or did one model get better and the other worse. We discuss such an approach in the Monte Carlo simulations in Section 3.

In addition, in a second step, after finding non-linearities, a researcher might want to investigate the reasons why a model forecasts better than a competitor and look for ways to improve his/her model specification and forecasts. In particular, it is plausible that non-linearities in the forecasting ability could be due to the existence of neglected non-linearities in the underlying models. For example, if a researcher, while comparing the forecasting performance of two models with constant parameters, finds non-linearities in the relative forecasting performance, he/she might consider adding threshold parameters or time-varying parameters to the best model in the hope to obtain a model that forecasts consistently better over time. In that case, the researcher could then estimate such a model and use our test again to evaluate whether the newly proposed model does indeed forecast consistently better. For example, if the researcher suspects that the non-linearities are due to omitted changes in parameters in the conditional mean or in the volatility of the disturbances, he/she might include the non-linearity explicitly in the model; then, to verify whether the modified model forecasts better, he/she would apply again our test. Clearly, researchers should be careful when applying our test multiple times, for example, using standard adjustments to the critical values in order to control the size when doing multiple testing (e.g. Bonferroni procedures to control the overall size of the sequence of the tests).

2.6 State dependence via Markov switching

An alternative approach to model state dependence is Markov switching (Hamilton, 1989). Unlike the threshold model, the regime changes in the Markov switching model depend upon an unobserved (latent) Markov chain, S_t. Testing in the presence of Markov switching also requires non-standard statistics as it is subject to two problems. The first problem is again the presence of nuisance parameters that are only identified under the alternative; in this case, the state-to-state transition probabilities and the coefficients that switch. The second problem is that, under the null, the score with respect to the restricted parameters is identically zero, which violates the regularity conditions that are imposed to derive the asymptotic chi-square distribution of the finite dimensional LR (Wald, LM) statistic by a first-order approximation. Therefore, the procedure proposed in Hansen (1996b), which deals with a nuisance parameter present only under the alternative, does not readily apply to the case of Markov switching models. Instead, Hansen (1992); Garcia (1998); Cho and White (2007); Carrasco et al. (2014) and Qu and Zhuo (2020) provide a variety of solutions that address both problems.

We propose a test for predictive ability in the presence of Markov switching based on Carrasco et al. (2014) in Appendix E and investigate its size properties as well. However, the test, like all the tests for Markov switching listed above, relies crucially on a correctly specified distribution under the null hypothesis.[9] A misspecified likelihood under the null will generally lead to size distortions. For instance, consider the case where the true but unknown distribution is a Student's t with no Markov switching. The researcher assumes a Gaussian distribution with no Markov switching under the null and a Markov switching model with regime dependent, conditional Gaussian distributions under the alternative. Then, despite the absence of Markov switching in the data generating process, the mixture property of the Markov switching model under the alternative may approximate the Student's t distribution better than the Gaussian model under the null hypothesis. Unreported results show that this leads to an over-rejection of the null hypothesis of no Markov switching.

While the assumption of Normality may be justified when applying tests for Markov switching models directly on economic observables, the distribution of a loss differential is generally unknown and may exhibit fatter tails than a Normal distribution (e.g. when using a quadratic loss). Consequently, the above-described problem is more severe in the case of forecast comparisons, and testing for Markov switching in this framework may be very sensitive to the choice of the parametric distribution. In contrast, and as outlined in Section 2.3, threshold models do not rely as heavily on the parametric assumptions on the error terms u_t, and testing is, therefore, more robust in practice.

3 Monte Carlo simulation analysis

In this section, we explore the size and power of our proposed tests in a series of Monte Carlo simulation exercises. We consider both nested and non-nested forecasting models as well as a variety of data generating processes (DGPs).

[9] Under the assumption of normality, the power of the test of Carrasco et al. (2014) relies on serial correlation in the error terms, instead of other deviations from the distribution specified under the null.

First, in Section 3.1 we investigate the size of our proposed DM^{NL} tests for point forecasts. We consider threshold, logistic smooth threshold as well as exponential smooth threshold regression specifications. Additional size results for alternative DGPs as well as density forecasts can be found in Appendix B.1. We also consider forecast encompassing as well as moment-based tests of forecast efficiency in Appendix B.2 and in the Online Appendix, respectively. In all cases, the tests are well-sized for moderate sample sizes.

In Section 3.2, we focus on the power properties of the tests. We consider both cases where the specification of the loss differential under the alternative is aligned with the true DGP as well as situations where it is not. For example, we let the true relative forecasting ability evolve over time according to a threshold model; however, in one case the researcher specifies a threshold model in the loss differential, and thus the specification under the alternative is aligned with the DGP, while in a second case the researcher specifies a smooth logistic threshold model, in which case the alternative and the DGP are not aligned. We also consider situations where the true DGP involves constant deviations from equal predictive ability.

While Section 3.1 and Section 3.2 consider DGPs directly modeled on the forecast error losses, in Section 3.3 we consider DGPs where the non-linear behavior in the forecast loss is generated from a primitive specification of the true underlying data, which allows us to consider both nested and non-nested forecasting models.

Finally, in all the power exercises in the main paper, we consider the case in which the state S_t is observed. However, our test also has power in case the researcher only observes a noisy estimate of S_t – for example, when S_t is a latent variable. As we discuss in Section 3.1and Section 3.2, our results are robust to situations in which the true state is unobserved and only a noisy estimate is available.

The horizon we consider is one (h=1), and the number of Monte Carlo replications is 3,000. For all Monte Carlo results we specified $\gamma \in [0.15, 0.85]$, denoted in terms of the the empirical cdf of S_t, and $\tau \in [0.1, 5]$. Results are not sensitive to reasonable changes in the domain of γ and τ.

3.1 Size results

The underlying data for the point forecast comparison are generated by

$$y_{t+1} = \nu + \delta_1 z_{t,1} + \delta_2 z_{t,2} + e_{t+1}, \tag{24}$$

where $\nu = \delta_1 = \delta_2 = 1$, $e_t \sim_{\text{iid}} \text{N}(0,1)$, $z_{t,1} \sim_{\text{iid}} \text{N}(0,1)$ and $z_{t,2} \sim_{\text{iid}} \text{N}(0,1)$. The parameter vector $\widehat{\beta}_t^{(j)} = [\widehat{\nu}_{t,j}, \widehat{\delta}_{t,j}]$ denotes the OLS estimator $\widehat{\beta}_t^{(j)} = \left(\sum_{i=t-R+1}^{t} z_{i-1}^{(j)\prime} z_{i-1}^{(j)}\right)^{-1} \sum_{i=t-R+1}^{t} z_{i-1}^{(j)\prime} y_i$, where $z_t^{(j)} = [1, z_{t,j}]$. The two point forecasts, both of which are misspecified, are denoted by: $\widehat{f}_{t+1|t}^{(1)} = z_t^{(1)} \widehat{\beta}_t^{(1)}$, and $\widehat{f}_{t+1|t}^{(2)} = z_t^{(2)} \widehat{\beta}_t^{(2)}$. As the misspecification of the two models is symmetric, it is straightforward to show that they have the same predictive ability in expectation. That is, the loss differential, given by

$$\Delta L_{t+1|t} = \left(y_{t+1} - \widehat{f}_{t+1|t}^{(1)}\right)^2 - \left(y_{t+1} - \widehat{f}_{t+1|t}^{(2)}\right)^2, \tag{25}$$

is zero in expectation: $\text{E}(\Delta L_{t+1|t}) = 0$ for all $t = R+1, ..., T$.

We generate time series of $\Delta L_{t+1|t}$ based on eq. (25) for several values of R and P: $R =$

$[25, 50, 100]$ and $P = [50, 100, 200, 1000]$. Then, we estimate the following model on the loss differential to investigate size:

$$\Delta L_{t+1|t} = \mu + \theta \cdot G(S_t; \varphi) + u_t, \tag{26}$$

where $G(\cdot)$ indicates the functional form of the TR, LSTR or ESTR models, described in eq. (2), (3), and (4); $S_t \sim_{\text{iid}} \text{N}(0,1)$ and we treat φ as unknown.

Table 1: Size results for point forecast comparisons

Panel A. ave-W

	TR				ESTR				LSTR			
R/P	50	100	200	1000	50	100	200	1000	50	100	200	1000
25	0.099	0.074	0.072	0.063	0.097	0.077	0.066	0.065	0.098	0.070	0.069	0.061
50	0.102	0.073	0.071	0.062	0.092	0.072	0.070	0.056	0.101	0.081	0.073	0.055
100	0.103	0.069	0.060	0.059	0.097	0.069	0.065	0.056	0.096	0.078	0.069	0.054

Panel B. exp-W

	TR				ESTR				LSTR			
R/P	50	100	200	1000	50	100	200	1000	50	100	200	1000
25	0.112	0.072	0.067	0.062	0.099	0.077	0.064	0.055	0.103	0.072	0.069	0.059
50	0.118	0.083	0.066	0.059	0.105	0.069	0.077	0.055	0.103	0.085	0.071	0.056
100	0.118	0.074	0.063	0.057	0.105	0.073	0.060	0.056	0.101	0.076	0.070	0.052

Panel C. sup-W

	TR				ESTR				LSTR			
R/P	50	100	200	1000	50	100	200	1000	50	100	200	1000
25	0.126	0.081	0.065	0.062	0.124	0.084	0.061	0.058	0.111	0.078	0.068	0.057
50	0.140	0.090	0.071	0.060	0.126	0.081	0.072	0.059	0.103	0.091	0.073	0.059
100	0.142	0.079	0.066	0.058	0.134	0.080	0.063	0.050	0.114	0.074	0.066	0.053

Note: The table displays empirical rejection frequencies of the null hypothesis $\text{H}_0 : \mu = \theta = 0$ for the DM^{NL} test for point forecasts evaluated with the MSFE loss function. The nominal size is 5%. Panels A to C show the results for the ave-W, exp-W, and the sup-W for the threshold model (TR), and the smooth threshold model using an exponential (ESTR) and a logistic (LSTR) function. The results are based on 3,000 MC replications.

Table 1 shows results for point forecast comparison for the null hypothesis defined in eq. (5) for the three different test statistics: sup-W, exp-W and ave-W. The size results are very similar for the TR, ESTR, and LSTR specifications. Overall, the ave-W has the best size properties and delivers size results that are good for $P > 50$ and $R > 25$ for both the threshold model and the smooth threshold model(s). The results of the exp-W test are similar to the ave-W; however, size distortions are slightly bigger in small samples than in the ave-W case. While the sup-W test works well in large samples ($P > 100$), it somewhat over-rejects in smaller samples. The fact that the ave-W has the smallest size-distortions and that the exp-W has smaller size distortions than the sup-W is a property also present in the original test of Hansen (1996b), who found similar results in a small Monte Carlo study for his null hypothesis.[10]

[10] In an unreported Monte Carlo study, we confirm the finding that, in the original test of Hansen (1996b), the smallest empirical rejection frequencies tend to be found for the ave-type test, and the largest for the sup-type test.

As previously mentioned, size results for nested models and density forecasts, as well as moment-based tests (such as forecast encompassing and efficiency), are discussed in Appendix B.1, Appendix B.2 and the Online Appendix, respectively.

3.2 Power results

In this section, we investigate the power properties of the threshold and logistic smooth threshold regression models. In particular, we specify three different alternatives for the loss differential defined in equation (25). The first alternative investigates the power of the proposed test statistics for detecting state dependence ($\theta \neq 0$). The second alternative investigates the empirical rejection frequency when both $\mu \neq 0$ and $\theta \neq 0$. The third alternative investigates power against a constant deviation from the null of equal predictive ability ($\mu \neq 0$).

In order to conduct the power analysis we proceed as follows. Let $\Delta L^{(0)}_{t+1|t}$ be the loss differential obtained from one Monte Carlo draw of eq. (25), normalized by its sample standard deviation (to ensure that the magnitude of the alternative is constant relative to the variation in $\Delta L_{t+1|t}$). For all simulations we use $S_t \sim_{i.i.d.} \mathrm{N}(0,1)$ and $\gamma = 0$.[11] In particular, we define the loss differential under the first alternative, Alternative (1), as

$$\Delta L^{(1)}_{t+1|t}(c) = \Delta L^{(0)}_{t+1|t} + \mu_c + \theta_c \cdot \mathbb{1}(S_t \geq \gamma), \tag{27}$$

where $c = 1, 2, ..., 14$ with $\mu_1 = 0, \mu_2 = 0.085, \mu_3 = 0.170, ..., \mu_{14} = 1.10$, and $\theta_c = -2\mu_c$. Note that $\gamma = 0$ implies that $\mathrm{E}(S_t \geq \gamma) = \frac{1}{2}$. Therefore, it follows that $\mathrm{E}_t \Delta L^{(1)}_{t+1|t} = \mu_c + \mathrm{E}(S_t \geq \gamma)\theta_c = \mu_c - \frac{1}{2}2\mu_c = 0$, i.e. the overall sample has a zero mean and the magnitude of the state-switching coefficient is 0.17 times the standard deviation of $\Delta L^{(0)}_{t+1|t}$, and so forth. In the case where $c = 1$, $\mu_1 = \theta_1 = 0$ implies that the joint null, defined in equation (5), holds. The design of Alternative (1) aims at isolating the power against the threshold alternative only, i.e. keeping the expected value over the full sample at zero. This enables us to compare the power of our tests to the power of the existing Fluctuation and GW test under non-linear alternatives. Therefore, we set the parameter values as described above to leave the expected value of the $\Delta L^{(1)}_{t+1|t}(c)$ over the full sample at zero.

For Alternative (2), the values of μ_c are unchanged but $\theta_c = -\mu_c$, which implies that $\mathrm{E}_t \Delta L_{t+1|t} \neq 0$. In other words, Alternative (2) is a case where both state dependence and a constant deviation from the null hypothesis are present:

$$\Delta L^{(2)}_{t+1|t}(c) = \Delta L^{(0)}_{t+1|t} + \mu_c + \theta_c \cdot \mathbb{1}(S_t \geq \gamma). \tag{28}$$

Alternative (3) considers constant deviation from the null hypothesis, i.e. $\theta_c = 0 \; \forall c$:

$$\Delta L^{(3)}_{t+1|t}(c) = \Delta L^{(0)}_{t+1|t} + \mu_c, \tag{29}$$

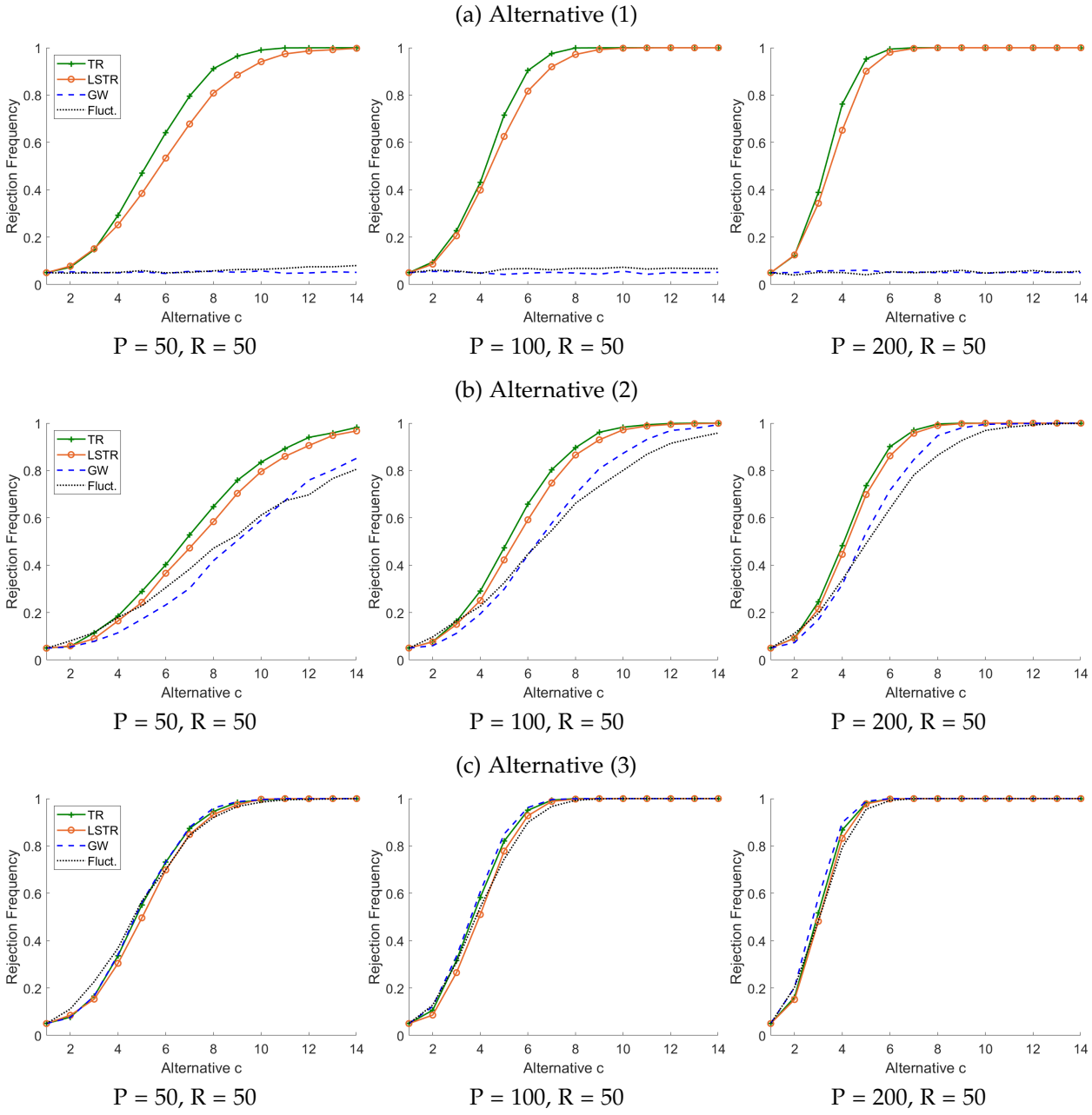
with $\mu_1 = 0, \mu_2 = 0.085, \mu_3 = 0.170, ..., \mu_{14} = 1$.

We then estimate two specifications of the model given in eq. (8): the TR model and the LSTR model. For both models we treat φ as unknown, and we test the null hypothesis defined in eq.

[11] Note that S_t is re-drawn in each Monte Carlo iteration for each alternative; we suppress the respective subscripts for notational convenience.

(5) using the DM^{NL} test defined in eq. (21). Note that since the alternative with state-dependence takes the form of a threshold model, the TR model is aligned with the DGP under the alternative. However, the LSTR is not aligned with the DGP under the alternative, which allows us to assess power in the case of misspecification. In the Online Appendix, we show results of the reverse type of misspecification, i.e. where the DGP under the alternative is a smooth logistic threshold model such that the TR is misspecified.

Figure 2: Size-adjusted power results for point forecast comparison



Note: On the y-axis the figures display size-adjusted empirical rejection frequencies of the null hypothesis $H_0 : \mu = \theta = 0$ for the DM^{NL} test for point forecasts evaluated with the MSFE loss function. The x-axis displays the magnitude of the alternative in units of c. The nominal size is 5%. The solid lines with markers "+" and "o" display the ave-W test results for the TR and LSTR model respectively. The dotted line displays the results of the Fluctuation test by Giacomini and Rossi (2010) and the dashed line displays the results of the GW test. The results are based on 3,000 MC replications.

Figure 2 shows size-adjusted power for the three alternatives defined in equations (27) to (29).[12] We compare the performance of our DM^{NL} with that of Giacomini and White (2006) (GW, who formalized Diebold and Mariano, 1995) and the Giacomini and Rossi (2010) Fluctuation test. The solid lines with markers "+" and "o" display the ave-W results for the threshold model and logistic smooth threshold model, and the dashed and dotted line show the results for the GW and Fluctuation test, respectively.[13] The three figures in Panel (a) show results for Alternative (1), i.e. state dependence without a constant deviation. As we can see, size-adjusted power increases quickly with the magnitude of the alternative as well as with the size of P. Since the DGP under the alternative is a threshold model, the LSTR exhibits rejection frequencies that are smaller than those of the TR but nonetheless high. In turn, the GW and Fluctuation tests have no power to detect the lack of equal predictive ability arising from the state dependence in the relative forecasting performance, and their power remains flat around the nominal size.[14]

The three figures in Panel (b) show results for Alternative (2), i.e. the case of a constant deviation and state dependence. The ave-W test of the TR and LSTR show again good size-adjusted power properties, and due to the presence of a constant deviation, the GW and Fluctuation rejection frequencies also increase as a function of the alternative's magnitude, although they reject less.

The three figures in Panel (c) show results for Alternative (3), i.e. a constant deviation without state-dependence. As expected, the GW test tends to be the most powerful test in this scenario; however, the power of the ave-W for both the TR and LSTR is very similar to that of the GW test.

3.3 Power for DGPs resulting in non-linear forecast error losses

In the previous section, the alternative was modeled directly as a non-linear function of $\Delta L_{t+1|t}$. In this section, we investigate power in two additional cases where the threshold behavior in the loss differential or the forecast encompassing loss is the result of non-linearities in the underlying data.

Common component DGP:
In our first example, two observable variables, y_t and x_t, are driven by a common component, c_t, and unpredictable idiosyncratic components, e_t and u_t respectively:

$$y_{t+1} = \alpha + c_{t+1} + e_{t+1}, \quad x_{t+1} = c_{t+1} + \eta_{t+1}, \quad c_{t+1} = \rho_c c_t + v_{t+1},$$

where α is a constant, $e_{t+1} \sim_{iid} N(0,1)$, $\eta_{t+1} \sim_{iid} N(0,1)$, and e_{t+1}, η_{t+1} and v_{t+1} are mutually independent. Importantly, $v_{t+1} \sim N(0, \sigma^2_{v,t+1})$, i.e. the variance is time-varying such that

$$\sigma_{v,t+1} = \begin{cases} \sigma_{v,L} & \text{if} \quad S_t \geq \gamma \\ \sigma_{v,H} & \text{if} \quad S_t < \gamma, \end{cases}$$

where $\sigma_{v,H} > \sigma_{v,L}$, i.e. the variance of the common component is a threshold-function of the state S_t. In other words, one can interpret x_t as a proxy for the common cycle, and S_t as a proxy for

[12] Results for further values of R and P are very similar and shown in the Online Appendix.
[13] Results for the exp-W and sup-W are virtually identical and available upon request.
[14] Note that the Fluctuation test might have better power in cases where S_t is a persistent variable.

the unobserved strength of the common cycle.

In fact, in periods when $S_t < \gamma$, the shocks to the common component have a larger variance then in periods when $S_t \geq \gamma$, which increases the importance of the common cycle relative to the idiosyncratic error and, therefore, increases the comovement between x_t and y_{t+1}.[15]

The benchmark forecasting model is the simple historical average while the competitor model uses x_t as a predictor:

$$\widehat{f}^{(1)}_{t+1|t} = \frac{1}{R}\sum_{i=t-R+1}^{t} y_i \quad \text{and} \quad \widehat{f}^{(2)}_{t+1|t} = \hat{\alpha}_t + \hat{\beta}_t x_t, \tag{30}$$

where $\hat{\alpha}_t, \hat{\beta}_t$ are estimated by regressing y_{t+1} on a constant and x_t, using a rolling window estimation scheme of size R. These forecating models are similar to the ones that we consider in the empirical analysis in Section 4.

Notice how the threshold in the volatility of the common component generates time-variation in the relative forecasting performance: when $S_t \geq \gamma$, the comovement between y_{t+1} and x_t is weaker and the predictor x_t adds mainly noise to the prediction. However, when $S_t < \gamma$, the forecast using x_t (the observable proxy of the common cycle) outperforms the historical mean prediction because the cyclical component dominates in these periods. Due to the autocorrelation in c_t, this effect will last for some periods even after S_t falls below the threshold again.

For the Monte Carlo study, we set $\alpha = 0.5$, $\rho_c = 0.8$, and $\sigma_{v,L} = 0.1$. The small value for $\sigma_{v,L}$ implies that during calm periods, i.e. when $S_t \geq \gamma$, the common component is of negligible importance for the forecasting performance; the idiosyncratic and unpredictable error e_t dominates. In turn, all else equal, the comovement of y_{t+1} and x_t increases in $\sigma_{v,H}$ such that the competitor model's relative performance in periods of $S_t < \gamma$ also increases in $\sigma_{v,H}$. Therefore, to investigate the power of our methodology to detect periods of predictability, we let $\sigma_{v,H}$ vary over a grid of equally spaced points in the interval $[0.1, 1.6]$.[16] The threshold variable we use is the same as in our empirical application of Section 4, that is the monthly U.S. real GDP growth estimate of Koop et al. (2020), which ranges from 1960:M6 to 2020:M12 (see Section 4 for more details on the variable). We set the true threshold value in the simulations to $\gamma = 1$, and treat it as unknown in the estimation of the threshold model on the loss differential and encompassing loss. The in-sample estimation size is $R = 240$ (as in our empirical application in Section 4), which leads to an out-of-sample size of $P = 487$.

Figure 3 displays the results. Panel (a) shows the power of the DM^{NL} ave-W test for a threshold regression model (TR, solid line with "+" markers) and a logistic smooth threshold regression model (LSTR, solid line with "o" markers), as well as for the Giacomini and White (2006) (dashed line) and Fluctuation tests (dotted line). The x-axis displays the grid of values of $\sigma_{v,H}$ and the y-axis shows the rejection frequency at the 5% percent nominal level. We observe that our proposed methodology has better power than the GW and Fluctuation tests as $\sigma_{v,H}$ increases. This is because, for large values of $\sigma_{v,H}$, the relative performance of the competitor model improves in periods when $S_t < \gamma$ (the persistence in c_t controls how the superior performance of the competitor model smoothes out over time). Therefore, in these periods, the mean squared forecast

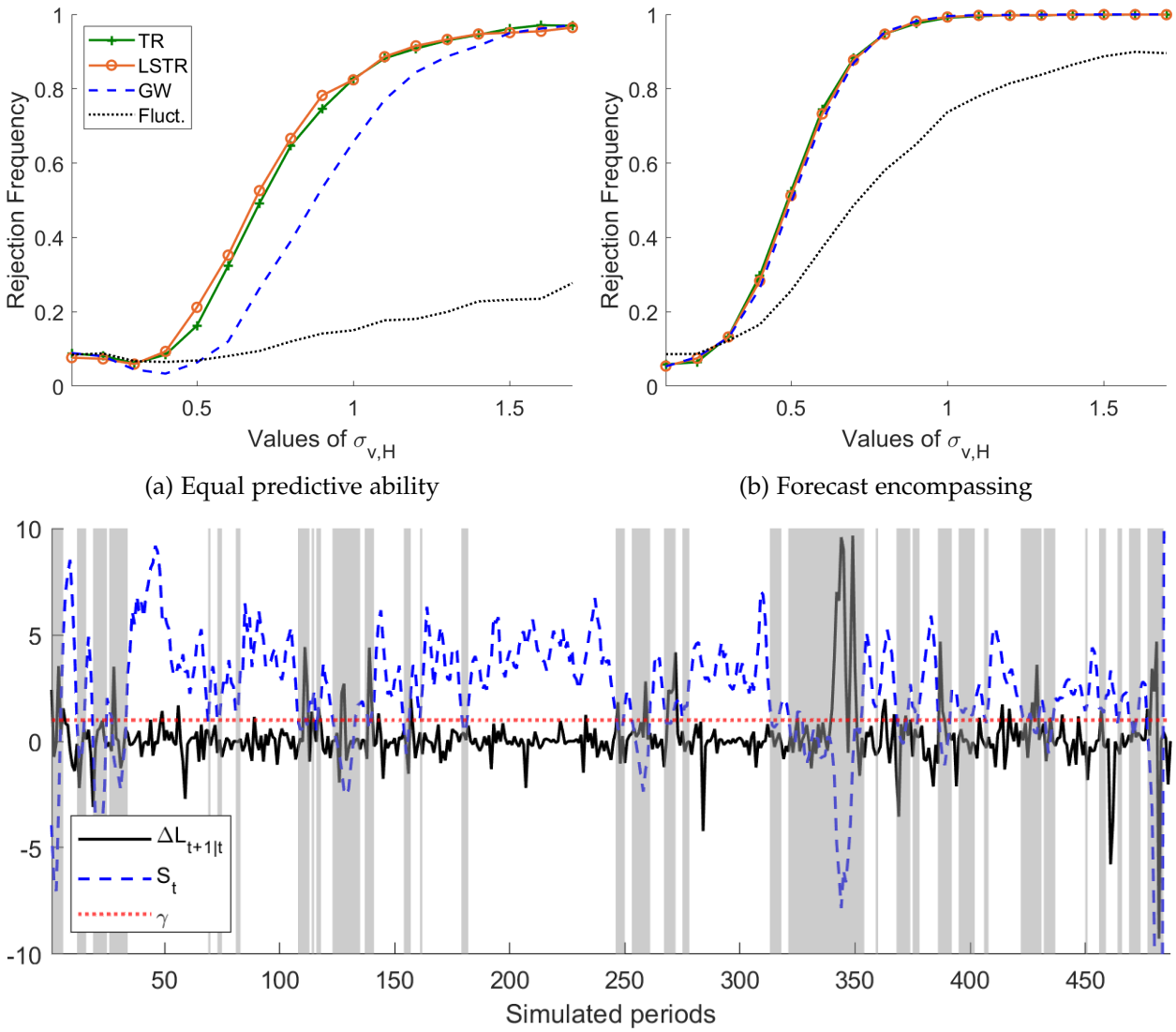
[15] Lagged values of x_{t+1} are correlated with y_{t+1} due to the persistence in the cycle.

[16] Note that a null hypothesis of equal predictive ability or forecast encompassing does not hold for any of the values for $\sigma_{v,H}$ because even for $\sigma_{v,H} = 0.1$ the common component is present in the DGP although with very small signal to noise ratio.

error (MSFE) of the competitor model is lower than that of the historical average.

Panel (b) shows the power of the DM^{NL} ave-W test for forecast encompassing, i.e. whether the first model encompasses the second model (again, the solid line with "+" markers denote the TR specification and the solid line with "o" markers denote the LSTR specification), as well as of the Giacomini and White (2006) (dashed line) and Fluctuation tests (dotted line). Our proposed methodology and Giacomini and White (2006) have very similar power.

Figure 3: Common component DGP: power



(a) Equal predictive ability

(b) Forecast encompassing

(c) Draw of simulated loss differential

Note: Panel (a) shows the power of the equal predictive ability test of the DM^{NL} ave-W as well as of the Giacomini and White (2006) and Fluctuation tests. Panel (b) shows power for the forecast encompassing test of the DM^{NL} ave-W as well as for the Giacomini and White (2006) and Fluctuation tests. The x-axis displays the grid of values of $\sigma_{v,H}$ and the y-axis shows rejection frequencies at a 5% percent nominal level. Panel (c) shows a draw of a simulated loss differential (solid line, positive numbers indicate a better forecast of the model using x_t) alongside S_t (dashed line), and the value of $\gamma = 1$ (dotted line). Grey shaded areas indicate periods where our estimated threshold model assigns a superior predictive ability to the model that uses x_t as a predictor.

Panel (c) shows one draw of the simulated loss differentials (solid line) alongside S_t (dashed line), and the value of $\gamma = 1$ (dotted line). When the dashed line is below the dotted horizontal line, we have that $\sigma_{v,t+1} = \sigma_{v,H}$, i.e. the signal of the common component is relatively stronger.

In turn, grey shaded areas indicate periods for which our threshold model, estimated on the loss differential, assigns a superior predictive ability to the competitor model. In other words, comparing periods where the dashed line is below the horizontal dotted line with the grey shaded areas provides a visual inspection of whether our modeling strategy of the loss differential can identify actual periods of superior predictive ability in the simulated data. As the figure shows, the grey shaded areas coincide with periods where $S_t < \gamma$ in the underlying DGP, indicating that our methodology can recover such periods.

Threshold regression DGP:
In our second example, there is a threshold relationship present in the underlying data that is unknown to the forecaster and, therefore, not modeled in either of the competing prediction models. This translates into a threshold relationship in the loss differentials (encompassing loss). Specifically, the underlying data, y_t, is generated by

$$y_{t+1} = \alpha + \rho_y y_t + \theta_1 z_{t,1} + \theta_2 z_{t,2} + \beta \cdot \mathbb{1}(S_t \geq \gamma) x_t + e_{t+1}, \tag{31}$$

with $e_t, z_{t,1}, z_{t,2} \sim N(0,1)$, $x_t = \rho_x x_{t-1} + \eta_t$, with $\eta_t \sim N(0,2)$, and S_t and γ are the indicator variable and threshold value that we specify below. The variable x_t is only present in the DGP in periods for which S_t is bigger than the threshold γ.

We first consider the case where the two competing forecasting models used to compute the loss differential for conditional mean predictions, labeled benchmark and competitor model respectively, are non-nested[17]:

$$\widehat{f}^{(1)}_{t+1|t} = \hat{\alpha}_{1,t} + \hat{\rho}_{1,t} y_t + \hat{\theta}_1 z_{t,1} \quad \text{and} \quad \widehat{f}^{(2)}_{t+1|t} = \hat{\alpha}_{2,t} + \hat{\rho}_{2,t} y_t + \hat{\theta}_2 z_{t,2} + \hat{\beta}_t x_t, \tag{32}$$

where $\hat{\alpha}_{1,t}, \hat{\rho}_{1,t}$, and $\hat{\theta}_{1,t}$ are estimated by regressing y_t on a constant and $y_{t-1}, z_{t-1,1}$, and $\hat{\alpha}_{2,t}, \hat{\rho}_{2,t}, \hat{\theta}_{2,t}$, and $\hat{\beta}_{2,t}$ are estimated by regressing y_t on a constant and $y_{t-1}, z_{t-1,2}$, and x_{t-1}; note that the non-linear relationship $\beta \cdot \mathbb{1}(S_t \geq \gamma) x_t$ is unknown to the forecaster. All parameters are estimated using a rolling window of size R, given below.

For the forecast encompassing test, the two competing forecasting models are nested[18], such that:

$$\widehat{f}^{(1)}_{t+1|t} = \hat{\alpha}_{1,t} + \hat{\rho}_{1,t} y_t \quad \text{and} \quad \widehat{f}^{(2)}_{t+1|t} = \hat{\alpha}_{2,t} + \hat{\rho}_{2,t} y_t + \hat{\beta}_t x_t. \tag{33}$$

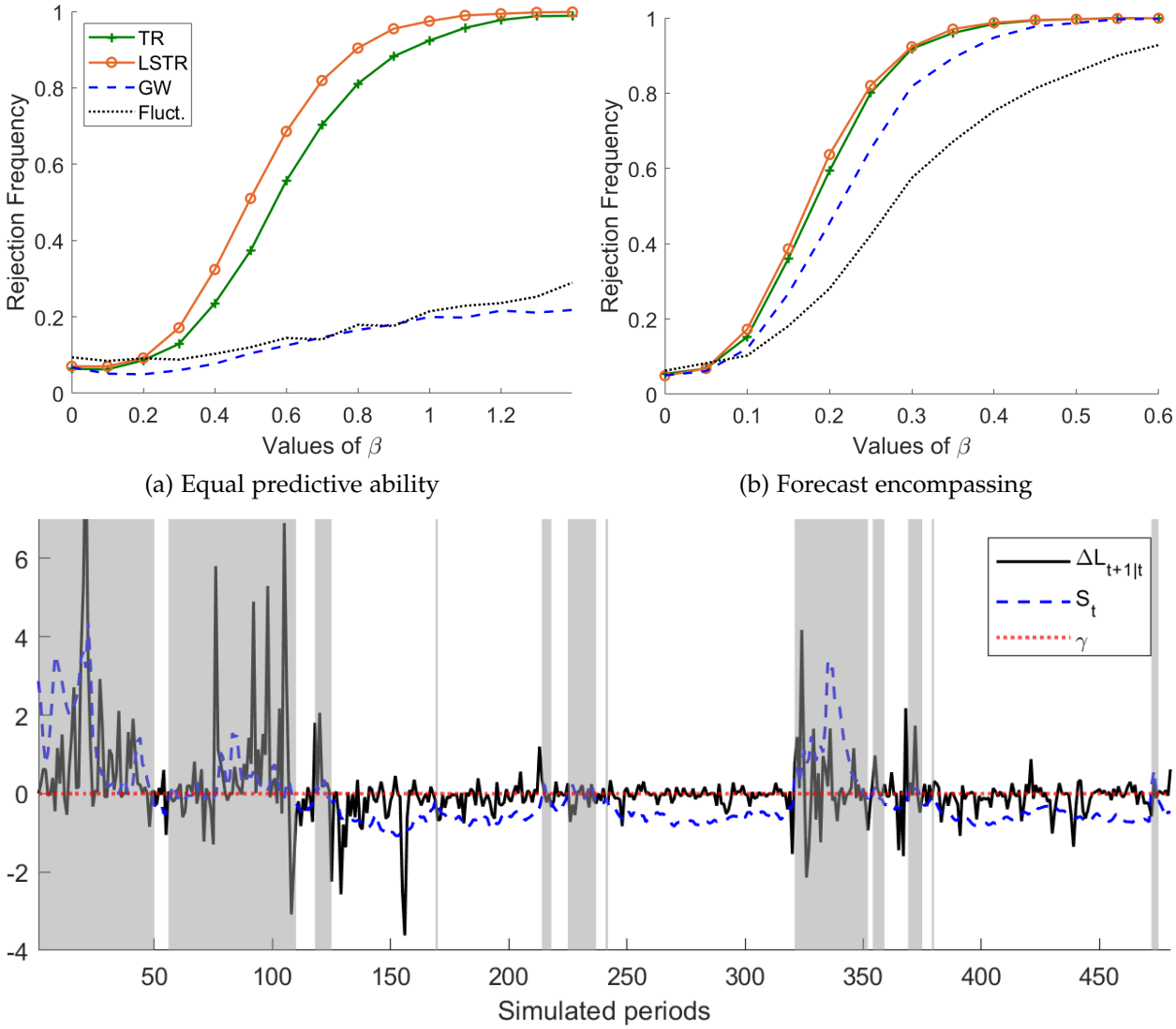
Notice that the competitor model will outperform the benchmark model when $S_t \geq \gamma$, since it incorporates information from x_t into the forecast. However, the competitor model uses x_t as a predictor independently of the value of S_t and, therefore, performs worse than the benchmark when $S_t < \gamma$. The larger β is, all else equal, the more the relative performance of the two competing models differs in the two states $S_t \geq \gamma$ and $S_t < \gamma$. Therefore, in our Monte Carlo study, we set $\alpha = \rho_y = \rho_x = \theta_1 = \theta_2 = 0.8$, and let β vary over a grid of equally spaced points in the interval $[0, 1.4]$. For the threshold indicator variable S_t we use the adjusted National Financial Conditions Index (ANFCI), computed by the Chicago Fed (see the empirical application in Section 5 for details), from 1971:M1 to 2019:M12. In the simulation of the DGP in eq. (31), we

[17] Since the common component DGP's forecasting models were nested and because it is known that GW performs better in finite sample when using non-nested models, we added the predictors $z_{t,1}$ and $z_{t,2}$.

[18] We dropped $z_{t,1}$ and $z_{t,2}$ from the predictions to avoid rejections of the null of "no forecast encompassing" due to $z_{t,1}$ or $z_{t,2}$.

set the true threshold value γ to zero, and treat it as unknown when estimating the threshold model on the loss differential and the encompassing loss. We set $R = 120$ (as in our empirical application in Section 5) for the in-sample estimation window size of the models in eq. (32) and (33), which results in $P = 480$ given ANFCI sample.

Figure 4: Threshold regression DGP: power



(a) Equal predictive ability

(b) Forecast encompassing

(c) Draw of simulated loss differential

Note: Panel (a) shows the power of the equal predictive ability test of the DM^{NL} ave-W as well as of the Giacomini and White (2006) and Fluctuation tests. Panel (b) shows power for the forecast encompassing test of the DM^{NL} ave-W as well as for the Giacomini and White (2006) and Fluctuation tests. The x-axis displays the grid of values of β and the y-axis shows rejection frequencies at a 5% percent nominal level. Panel (c) shows a draw of a simulated loss differential (solid line, positive numbers indicate a better forecast of the model using x_t) alongside S_t (dashed line), and the value of $\gamma = 0$ (dotted line). Grey shaded areas indicate periods where our estimated threshold model assigns a superior predictive ability to the model that uses x_t as a predictor.

Figure 4 displays the results. Panel (a) shows power of the DM^{NL} ave-W test (solid line with "+" markers for TR, solid line with "o" markers for LSTR) for equal predictive ability, as well as for the Giacomini and White (2006) (dashed line) and Fluctuation test (dotted line). The x-axis displays the grid of values of β and the y-axis shows the rejection frequency at a 5% percent nominal level. We observe that our proposed DM^{NL} ave-W test statistic substantially outperforms

the Giacomini and White (2006) and Fluctuation test as β increases.

Panel (b) shows power results for the forecast encompassing test, i.e. whether the first model is encompassing the second model in eq. (33). Again, our threshold model outperforms the GW and Fluctuation tests as β increases.

Panel (c) shows an example of simulated loss differentials (solid line) alongside S_t (dashed line) and the value of $\gamma = 0$ (dotted line). The grey shaded areas again denote periods when the competitor model predicts better than the benchmark according to the estimated threshold model. If the grey shaded areas coincide with periods when the true state is such that $S_t \geq \gamma$ (and some periods thereafter due to the persistence in y_t), our estimation strategy on the loss differential identified the periods where the competitor model forecasts better than the benchmark. Figure 4 provides visual evidences that this is clearly the case.

The case of state variables observed with measurement error:
In the Monte Carlo results shown above, the state S_t is observed. However, in practice, the true state variable may be latent and, therefore, only available with measurement error, a case which we investigate in the Online Appendix.

In particular, we consider the scenario where the researcher does not observe S_t but only a noisy measure $\tilde{S}_t = S_t + v_t$, with $v_t \sim N(0, \sigma_v^2)$. Therefore, when estimating the threshold model on the loss differential (or forecast encompassing moment), the researcher has to condition on $\tilde{S}_t$ instead of S_t itself:

$$\mathrm{E}_t \Delta L_{t+1|t} = \mu + \theta \cdot \mathbb{1}\left(\tilde{S}_t \geq \gamma\right).$$

We find that, when setting the standard deviation of the noise term to a quarter of the standard deviation of the "signal" S_t, power results are still very good using our $\mathrm{DM}^{\mathrm{NL}}$ test.

4 Empirical application: uncovering pockets of predictability in equity premia

Financial return predictability is typically time-varying and elusive. As noted by Pesaran and Timmermann (1995), Rapach and Wohar (2006), and Rapach and Zhou (2013), the predictability of stock market returns appears only when focusing on special sub-samples; Goyal and Welch (2003, 2008) similarly find that predictors that successfully forecast equity premia, the U.S. returns or dividend price ratios typically change over time. Instabilities are widespread: Paye and Timmermann (2006), for example, cannot reject the presence of structural breaks in stock return predictive regressions in several countries and Rossi (2006, 2013b) find similar results for exchange rate returns. As summarized in Timmermann (2008), "... there appear to be pockets in time where there is modest evidence of local predictability; (...) the best forecasting method can be expected to vary over time, and there are likely to be periods of model breakdown where no approach seems to work". It is then inevitable that one must confront time variation when evaluating financial models' predictive ability in an attempt to track their "local" forecasting performance.

As discussed in Timmermann (2008) and Paye and Timmermann (2006), the predictability of equity premia could be caused by market inefficiencies. If that is the case, then rational investors will take the opportunity to trade and make profits. However, if a large number of investors engage in taking advantage of the predictability, their behavior will eventually eliminate the

predictability altogether. This implies the existence of short windows of time in which equity premia are predictable, but a low (or no) predictability in the rest of the sample.

In what follows, we attempt to uncover pockets of predictability in U.S. equity premia in out-of-sample. We consider several of the classic economic predictors considered in Goyal and Welch (2008): the book to market ratio (calculated as the ratio of the book value and the market value of the Dow Jones Industrial Average and labeled "BookToMarket"); the default yield spread (calculated as the difference between BAA and AAA-rated corporate bond yields and labeled "DFY"); a monthly inflation measure based on Consumer Price Index (labeled "Inflation"); a stock variance measure computed as the sum of squared daily returns (labeled "StockVar"); the long term government bond yield (labeled "LongYield"); the short term government bond yield (labeled "Tbill"); and the term spread (calculated as the difference between the long term yield on government bonds and the Treasury bill and labeled "Spread").[19]

Then, the economic models are as follows:

$$\mathrm{E}_{t-1} r_t = \nu + \delta z_{t-1}, \tag{34}$$

where r_t is the equity premium, z_{t-1} is the lagged economic predictor and ν is the intercept. As the benchmark model, we focus on the historical mean, also calculated using a rolling window of past returns over the previous twenty years. All models are estimated in a window of the past twenty years of data, i.e. $R = 240$, producing a series of rolling one-step-ahead out-of-sample forecasts.[20]

We estimate the "local" forecasting performance using a non-linear model in the loss differences, where the loss difference is the difference in the squared out-of-sample forecast error of the benchmark (historic rolling mean) minus that of the economic model (eq. 34):

$$\mathrm{E}_t \Delta L_{t+1|t} = \mu + \theta \cdot \mathbb{1}\left(S_t \geq \gamma\right). \tag{35}$$

Following the existing literature on the countercyclicality of equity premia, we use a measure of monthly real GDP growth, computed by Koop et al. (2020), as our indicator variable. Koop et al. (2020) impute a "true", yet unobserved, monthly real GDP growth series based on a mixed-frequency Bayesian Vector Autoregression (BVAR), using accounting identities as well as GDP expenditure-side and GDP income-side estimates as observables. Their monthly GDP growth series ranges from 1960:M6 to 2020:M12.[21]

We chose output growth as the state variable since stock returns are linked, by the net present value theory, to current and future output growth. Using a similar argument, Neely et al. (2014) and Henkel et al. (2011), for example, show that return predictability is linked to economic recessions. In fact, equity premia are countercyclical (Campbell and Cochrane, 1999; Bekaert et al., 2009) and return predictability is correlated with the business cycle (Neely et al., 2014; Dangl and Halling, 2012; Henkel et al., 2011; Rapach et al., 2010). In particular, when predicting

[19] The data are from A. Goyal's website: http://www.hec.unil.ch/agoyal/

[20] Goyal and Welch (2008) use 20 years of monthly data and Harvey et al. (2021) use around 20 years of monthly data for their results, both based on a rolling-window estimation scheme. We provide robustness results for different sample sizes below.

[21] We are using the January 2021 vintage of Koop et al. (2020), who update the monthly real GDP growth series when new data become available. In unreported results, we find that using other vintages available from Koop et al. (2020) leads to the same results as presented here.

monthly equity premia, Neely et al. (2014) find that economic predictors, especially the term spread and default yield spread, as well as technical indicators, beat the benchmark mean forecast during NBER-dated recession periods. Relative to Neely et al. (2014) and Henkel et al. (2011), our analysis has the advantage that we can formalize the link between predictability and the business cycle using a non-linear model for the forecast error loss.

The idea, formalized in eq. (35), is to capture Timmermann's (2008) "pockets of predictability", where the pockets of predictability depend on the state of the business cycle. That is, the relative performance of the models changes over time depending on whether real GDP growth is higher (or lower) than an unknown threshold value. Note that, when the loss differential is positive, the economic model is better than the benchmark (the historical mean).

In addition, we are interested in testing for forecast encompassing; in fact, it could be the case that some of the predictors result in forecasts that are similar to the mean forecast in terms of accuracy but nonetheless have predictive power that is neglected by the benchmark model. To assess whether the benchmark forecast encompasses the forecast using an economic predictor in a state-dependent manner, we estimate:

$$\mathrm{E}_t(e^2_{t+1|t,1} - e_{t+1|t,1}e_{t+1|t,2}) = \mu + \theta \cdot \mathbb{1}\left(S_t \geq \gamma\right), \tag{36}$$

where $e_{t+1|t,1}$ denotes the forecast error of the mean model, and $e_{t+1|t,2}$ denotes the forecast error of the model with the economic predictor.

Table 2 reports the results of equal predictive ability (Panel A) and the forecast encompassing tests (Panel B). Since our forecasting exercise is most closely related to Neely et al. (2014), our baseline results use the same out-of-sample period as Neely et al. (2014), from 1966:M1 to 2011:M12; results for alternative samples are similar and shown in the Online Appendix. For each predictor, listed in the first column, we report the p-values for the sup-W, ave-W, and exp-W test. In addition, we report the p-value of the Diebold and Mariano (1995)/Giacomini and White (2006) (DM/GW) and the Fluctuation tests, and the out-of-sample size (for the Fluctuation we only indicate whether the p-value is smaller or larger than 0.10 or 0.05). For the case where the $\mathrm{DM}^{\mathrm{NL}}$ tests reject the null hypothesis of equal performance/forecast encompassing, we report the estimated parameters of the model defined in eq. (35) in Table 3. In addition, Table 3 reports the results of t-tests on the parameters, a Wald test on the sum of the parameters, the estimated threshold parameter, and the frequency of the states.

Panel A of Table 2 shows that we find evidence of pockets of predictability when forecasting using the term spread. The estimate of μ, shown in Table 3, is positive indicating that the loss difference is positive when real GDP growth is lower than the threshold value, in which case the economic model has a better predictive ability than the benchmark model. However, when real GDP growth is high, the loss difference becomes negative. That is, the spread adds noise to the prediction of the returns during periods of at least moderate GDP, while the opposite is true when GDP growth is low. Panel B of Table 2 shows that the results regarding the term spread hold when testing for forecast encompassing as well.

Notice that, unlike our proposed tests, the GW test does not find significant differences between the model with the spread and the benchmark. This is because our test is more powerful to detect pockets of predictability when there are instabilities associated with non-linearities. Notice that the Giacomini and Rossi (2010) Fluctuation test statistic is not bigger than the critical

value in the case of the spread either; hence, even though the Fluctuation test is robust to instabilities in the relative forecasting performance, nevertheless it is less powerful than the test proposed in this paper and, in our data, never finds evidence of predictive ability, possibly due to its sporadic appearance over time.

Our results are related to previous work, for instance, Neely et al. (2014), who found evidence that the term spread is capable of predicting equity premia during economic recessions. However, they differ in two important ways. Firstly, our paper is the first to use a monthly real GDP growth series and formally test for non-linearity in the loss differentials; in fact, previous papers relied on pre-defined NBER recession dummies or quarterly GDP data (Rapach et al., 2010). Secondly, using monthly real GDP growth in combination with our methodology, we find that the pockets of predictability are correlated with, but not limited to, recessionary periods. In particular, there are 148 months that our model identified as pockets of predictability but which are not classified as NBER recession dates. Our results, therefore, suggest that the pockets of predictability are not only correlated with the business cycle but also present outside NBER recession dates.

Table 4 sheds further light on what drives the relative forecasting performance. It shows the two competing forecasts' MSFEs for different subsamples: the full out-of-sample as well as the two subsamples identified by $S_t \geq \hat{\gamma}$ and $S_t < \hat{\gamma}$. In periods of high real GDP growth, i.e. when $S_t \geq \hat{\gamma}$, the performance improves for both forecasts relative to the full sample, but less so for the forecast obtained with the spread as a predictor. In turn, during periods of low real GDP growth, i.e. when $S_t < \hat{\gamma}$, both the simple mean as well as the forecast obtained with the economic predictor perform worse, but the latter less so.

We further conducted a number of robustness checks. In the first robustness check, we increased (decreased) R, the in-sample estimation window size, to 300 (180) observations. Results are reported in Appendix C - in particular, see Table C.1, for the test of equal predictive ability, and Table C.2, for the forecast encompassing test. The results are similar even with different in-sample estimation window sizes.

In the second robustness check, reported in Table C.3, we add monthly real GDP growth as a linear control variable in eq. (35), i.e. $\mathrm{E}_t \Delta L_{t+1|t} = \mu + \theta \cdot \mathbb{1}(S_t \geq \gamma) + \phi S_t$, and test $\mu = \theta = 0$. As in the baseline result in Table 2, our test rejects the null hypothesis for the spread, pointing out that a simple linear conditional test would not be sufficient to uncover the non-linear relationship between GDP growth and the pockets of predictability.

In the third robustness exercise, reported in the Online Appendix, we show that our results are robust to specifying the out-of-sample period as 1966:M1 to 2006:M12 (pre-financial crisis), 1966:M1 to 2017:M12 (the end of the Goyal and Welch, 2008 dataset), or 1960:M6 (the start of the monthly GDP series) to 2011:M12.

For the spread, for which we found that the economic model performs sometimes better than the benchmark, Figure 5 reports the loss differences ($\Delta L_{t+1|t}$, solid line) over time, together with the monthly real GDP growth series (S_t, dashed line) that triggers the state-switching. Shaded areas depict periods where the economic model has, on average, a lower squared forecast error than the benchmark model (the loss difference is positive on average). Periods that are not shaded indicate times in which the benchmark model performs better than the economic model. The figure shows that there are several pockets of predictability, where the economic model predicts slightly better than the benchmark. Furthermore, these pockets persist for a few periods and are

interrupted by periods when the economic model performs worse than the benchmark, causing the average performance of the model to be poor over the entire sample.

Farmer et al. (2019) also find evidence in favor of pockets of in-sample predictability in U.S. equity premia using the term spread. Their methodology is very different from ours, as they employ a time-varying parameter model estimated non-parametrically while we model directly the forecast loss differential. In addition, they focus on equity premia at a daily frequency, while we have data at a monthly frequency. Despite these differences, several of their identified pockets are similar to ours.

In particular, our pockets identified for the beginning of the 1970s, the mid-1970s, and the beginning of the 1980s coincide with the in-sample findings of Farmer et al. (2019) for the term spread. Moreover, Farmer et al. (2019) show out-of-sample results for daily data using a bivariate model with the term spread and the T-bill as predictors. Again, the pockets they find at the beginning of the 1970s, the mid-1970s, and the beginning of the 1980s coincide with ours for the term spread.

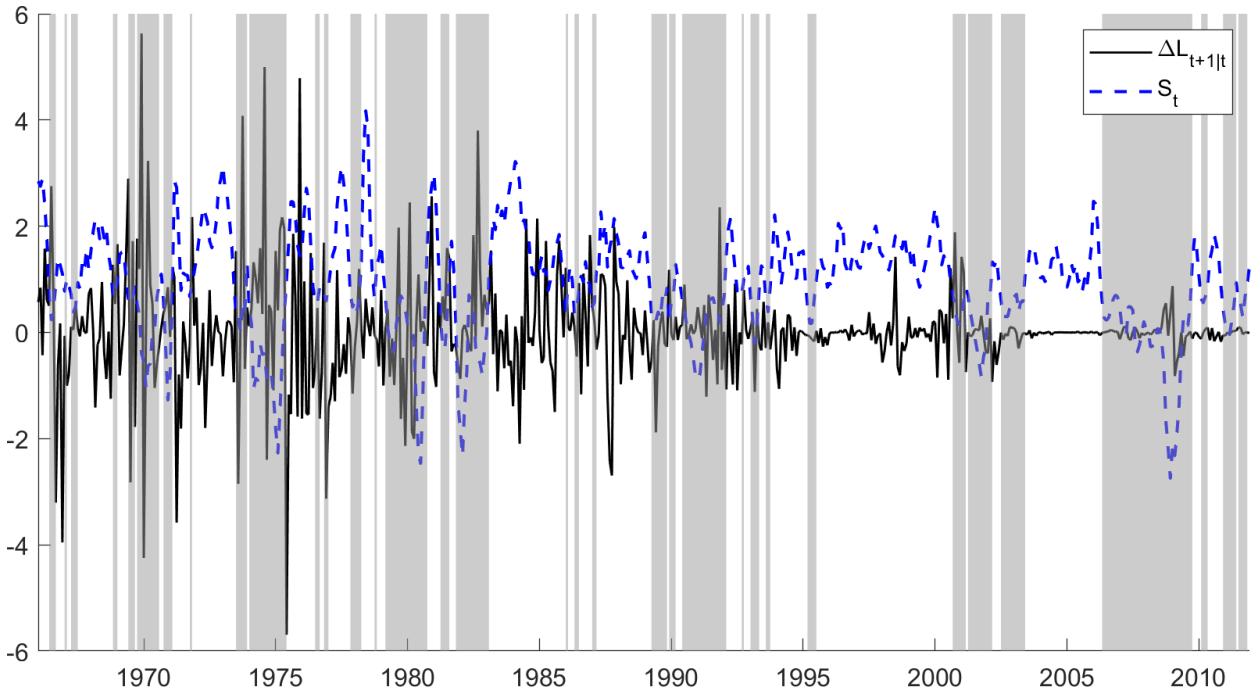
Table 2: State-dependence in equity premium forecasts

Variable Name	DM^{NL}			Alternative statistics		Sample size
			Panel A. Loss differential			
	sup-W	ave-W	exp-W	GW	Fluct.	P
DFY	0.144	0.278	0.242	0.523	$<$ **0.10**	552
Inflation	0.586	0.538	0.539	0.602	$>$ 0.10	552
StockVar	0.515	0.463	0.469	0.238	$>$ 0.10	552
LongYield	0.855	0.729	0.743	0.700	$>$ 0.10	552
Spread	**0.020**	**0.031**	**0.020**	0.570	$>$ 0.10	552
Tbill	0.774	0.724	0.734	0.896	$>$ 0.10	552
BookToMkt	0.408	0.156	0.194	**0.070**	$<$ **0.10**	552
			Panel B. Forecast encompassing			
	sup-W	ave-W	exp-W	GW	Fluct.	P
DFY	0.453	0.458	0.465	0.416	$>$ 0.10	552
Inflation	**0.034**	**0.021**	**0.024**	**0.054**	$<$ **0.05**	552
StockVar	0.576	0.566	0.575	0.504	$>$ 0.10	552
LongYield	0.768	0.449	0.507	0.260	$>$ 0.10	552
Spread	**0.002**	**0.001**	**0.001**	**0.002**	$<$ **0.05**	552
Tbill	0.396	0.129	0.165	**0.073**	$>$ 0.10	552
BookToMkt	0.460	0.270	0.307	0.163	$>$ 0.10	552

Note: The table shows p-values of tests of equal predictive ability (Panel A) and forecast encompassing (Panel B) using the DM^{NL}, the GW, and the Fluctuation test (for the Fluctuation test we only indicate whether the p-value is smaller or larger than 0.10 or 0.05). Bold indicates significance at the 10% level. The in-sample estimation window size, R, is 240.

Finally, note that, in this paper, we focus on detecting "pockets of predictability" in historical data and linking it to the time-variation in an economic threshold variable. For readers interested in methodologies specifically tailored for real-time detection, Inoue and Rossi (2015) and Harvey et al. (2021) propose monitoring procedures to detect structural changes. They suggest sequentially

Figure 5: Pockets of predictability using the term spread



Note: The figure shows the estimated $\Delta L_{t+1|t}$ (solid line, positive numbers indicate that the economic model forecasts better than the mean model) together with the monthly real GDP growth measure S_t (dashed line), which triggers the state-switching, for the term spread as the predictor. Shaded areas indicate periods where the economic model performs better than the benchmark, i.e. they show the pockets of predictability.

Table 3: State-dependent estimates: eq. (35) and (36)

Predictor	Parameter estimates				Parameter tests			State characteristics	
	$\hat{\mu}$	$\hat{\theta}$	$\hat{\mu}+\hat{\theta}$	$\hat{\gamma}$	$\hat{\mu}=0$	$\hat{\theta}=0$	$\hat{\mu}+\hat{\theta}=0$	$\bar{S}$	$\mathrm{P}(S_t \geq \hat{\gamma})$
				Panel A. Loss differential					
Spread	0.190	−0.284	−0.093	0.779	**3.793**	**−3.248**	1.554	0.922	0.592
				Panel B. Forecast encompassing					
Spread	0.288	−0.274	0.014	0.779	**6.113**	**−2.819**	0.027	0.922	0.592
Inflation	−0.006	0.092	0.086	0.086	−0.221	0.501	0.223	0.922	0.842

Note: The table shows the parameter estimates associated with the equation in (35) and (36). The columns $\hat{\mu}, \hat{\theta}, \hat{\mu}+\hat{\theta}$, and $\hat{\gamma}$ show the parameter estimates associated with the sup-W statistic. The columns $\hat{\mu}=0, \hat{\theta}=0$, and $\hat{\mu}+\hat{\theta}=0$ show the values of the statistic when using a t-test or a Wald test respectively, for testing the hypothesis that the parameters, or their sum, are equal to zero. The critical values of a Wald test, with one restriction, at the 5% and 10% level are 2.706 and 3.842. Bold numbers indicate a rejection at the 10% level. The column $\bar{S}$ shows the average value of the conditioning variable S_t, and $\mathrm{P}(S_t \geq \hat{\gamma})$ shows the relative frequency of being in the state where both μ and θ are present.

repeating t-tests over short time periods and control the overall rejection rates. For example, in their application to predictive regressions, Harvey et al. (2021) find that the one-month ahead equity premium had been predictable at several points in time and that such episodes could have been detected in real-time by their methodology. Unlike us, Harvey et al. (2021) do not find pockets for the term spread as a predictor. As pointed out by Harvey et al. (2021) themselves when comparing their results to Neely et al. (2014), these difference can stem from the longer sample of the data we are using and the fundamental difference between an ex post analysis of predictability, which is our approach, and the real-time monitoring approach of Harvey et al. (2021).

Table 4: Mean squared forecast errors in subsamples identified by the state variable

Predictor	Mean forecast (MF)			Economic predictor (EP)			Rel. MSFE: EP/MF	
	Full	$S_t < \hat{\gamma}$	$S_t \geq \hat{\gamma}$	Full	$S_t < \hat{\gamma}$	$S_t \geq \hat{\gamma}$	$S_t < \hat{\gamma}$	$S_t \geq \hat{\gamma}$
Spread	1.000	1.347	0.761	0.992	1.283	0.792	0.952	1.041

Note: The column labeled "Full" shows the MSFEs over the full out-of-sample period. The column labeled $S_t < \hat{\gamma}$ ($S_t \geq \hat{\gamma}$) shows the MSFE in the subsample where S_t is lower (greater) than the threshold value. The columns labeled "Rel. MSFE: EP/MF" show the ratio of the MSFE of the economic model and the mean model in the respective subsamples; a number smaller than one indicates a superior performance of the economic model. To ease the comparison between the subsamples, we normalized all values by the MSFE of the mean forecast in the full sample.

5 Empirical application: forecasting industrial production

Non-linearities are also widespread in macroeconomic forecasting. As the literature has shown, the relative forecasting performance of the models is unstable and varies over time (see Rossi, 2013a for a reference). Moreover, while parsimonious statistical models tend to forecast well in regular times, models that use economic predictors are more accurate in disruptive times, such as during recessions and financial crises (see Chauvet and Potter, 2013).

In this section, we investigate the state dependence in models' relative forecasting performance when predicting U.S. industrial production (IP). Adrian et al. (2019) demonstrate the importance of financial conditions when forecasting the distribution of output growth, particularly tail risk, advocating for a non-linear relationship between financial stability and macroeconomic performance. They suggest that financial conditions can be important for growth for two reasons. In structural models, frictions in either the supply of or the demand for credit can result in non-linear equilibrium relationships between financial conditions and growth (see He and Krishnamurthy, 2011 and Brunnermeier et al., 2013, among others). On the other hand, financial variables, by virtue of being fast-moving variables, can provide more timely signals about negative shocks to the economy. Granziera and Sekhposyan (2019) further show that financial conditions can be useful for model selection when forecasting industrial production.

Motivated by the evidence above, we model the state-dependence in the relative forecasting performance of models of IP growth in terms of financial conditions and use the Chicago Fed's adjusted National Financial Conditions Index (ANFCI) as an indicator defining the state. The

ANFCI is a measure of financial conditions constructed as a weighted average of 105 measures of financial activity, adjusted to remove the variation associated with economic activity and inflation. By construction, values above zero indicate financial conditions that are tighter than average. The series are from FRED database of the Federal Reserve Bank of St. Louis. The relative forecasting performance, however, could vary over time not only when the ANFCI is greater than zero, but also when it is above/below some other threshold value. We let the threshold be unknown and estimate it using our procedure. Our proposed tests detect the thresholds in the models' relative forecasting performance across financial cycles.

The forecasting environment is similar to Granziera and Sekhposyan (2019). The benchmark is an autoregressive model of order p (AR(p)):

$$y_{t+1} = \alpha + \rho(L_p)y_t + e_t,$$

where y_t is industrial production growth, $\rho(L_p) = \rho + \rho_1 L + ... + \rho_p L^p$ is a lag-polynomial with a lag length (p) chosen by the BIC, which is re-estimated at each point in time. The competitor is an autoregressive-distributed lag model (ADL(p,q)) using one economic predictor at-a-time:

$$y_{t+1} = \alpha + \rho(L_p)y_t + \phi(L_q)x_t + e_t,$$

where $\phi(L_q) = \phi + \phi_1 L + ... + \phi_p L^p$. The lag length p of the autoregressive component is given from the AR(p), and we (re-)estimate q via the BIC, again, at each point in time.

We consider four macroeconomic variables and four financial variables as predictors (label and mnemonics are reported in parentheses): new privately-owned housing units started (labeled "Housing Starts," HOUST), the non-farm vacancies divided by the number of unemployed (labeled "VacancyToUr", HelpHWIURATIO), the employment level (labeled "Employment", CE16OV), new orders of durable goods (labeled "New Orders", AMDMNOx), outstanding consumer credit, measured by outstanding total nonrevolving credit owned and securitized (labeled "Consumer Credit", NONREVSL), the spread between the one-year Treasury rate minus the Federal Funds rate (FFR) (labeled "One Year Spread", T1YFFM), the spread between the ten-year Treasury rate and the FFR (labeled "Ten Year Spread", T10YFFM), and the Moody's Baa Corporate Bond rate minus the FFR (labeled "Credit Spread", BAAFFM). These indicators are broadly considered to be leading indicators for real economic activity. The data is from the FRED-MD database of McCracken and Ng (2016) and ranges from January 1959 to December 2019. We transform the data as recommended in McCracken and Ng (2016) to ensure stationarity.

The forecasting horizon we consider is one month. The parameters $\alpha, \rho(L_p)$ and $\phi(L_q)$ are estimated using a rolling window of 10 years, i.e. 120 observations. Given the one-step-ahead forecasts of the AR(p) and ADL(p,q), we compute the loss differential, $\Delta L_{t+1|t}$ as the difference between the squared forecast error of the AR(p) and the squared forecast error of the ADL(p,q). Then, we model the loss differential as

$$\mathrm{E}_t \Delta L_{t+1|t} = \mu + \theta \cdot \mathbb{1}(S_t \geq \gamma) \tag{37}$$

and we test the null hypothesis $\mathrm{H}_0 : \mu = \theta = 0$. Since the ANFCI series starts in 1971:M1, our out-of-sample size is 587 observations.

Table 5 shows results for our test statistics. When we use the VacancyToUR, New Orders, and Consumer Credit as predictors, our sup-W, ave-W, and exp-W reject the null hypothesis of equal predictive ability. In fact, the GW test rejects equal predictive ability in these cases as well, yet it is not suitable to shed light on the presence of state-dependence. The fluctuation test (Fluct.), on the other hand, does not reject equal predictive ability.

Table 5: State-dependence in industrial production forecasts

Variable Name	DM^{NL}			Alternative statistics		Sample size
	sup-W	ave-W	exp-W	GW	Fluct.	P
Housing Starts	0.135	0.148	0.135	0.641	> 0.10	587
VacancyToUr	**0.021**	**0.022**	**0.023**	**0.045**	> 0.10	587
Employment	0.218	0.309	0.294	0.362	> 0.10	587
New Orders	**0.004**	**0.003**	**0.003**	**0.039**	> 0.10	587
Consumer Credit	**0.027**	**0.049**	**0.043**	**0.020**	> 0.10	587
Ones Year Spread	0.598	0.668	0.681	0.430	> 0.10	587
Ten Year Spread	0.755	0.641	0.666	0.422	> 0.10	587
Credit Spread	0.523	0.733	0.719	0.740	> 0.10	587

Note: The table shows p-values of tests of equal predictive ability using the DM^{NL} sup-W, ave-W, exp-W tests as well as GW and Fluctuation test (for the Fluctuation test we only indicate whether the p-value is smaller or larger than 0.10 or 0.05). Bold indicates significance at the 10% level. The in-sample estimation window size, R, is 120.

The coefficient estimates of the threshold model for cases where the test rejects are reported in Table 6. The estimates strongly suggest the presence of state dependence for the VacancyToUR index and New Orders. Whenever financial conditions are sufficiently tight, i.e. $S_t \geq \hat{\gamma}$, the models using either the VacancyToUR or the New Orders as a predictor performs better, as indicated by $\theta > 0$ and $\theta > |\mu|$. Moreover, the estimated threshold value ($\hat{\gamma}$) is 0.767 for the New Orders. The fact that $\hat{\gamma}$ is considerably above zero, the ad-hoc threshold at which the ANFCI indicates tighter financial conditions, shows the usefulness of our method since our method does not rely on the ad-hoc threshold value, and, in fact, the estimated threshold value is much larger. In fact, when using a GW test on the subsample of the loss differential for which $S_t \geq 0$, we cannot reject the null hypothesis of equal predictive ability (the value of the test statistic of 0.264); when using the GW test on the subsample identified by $S_t \geq 0.767$, the value of the test statistic is 2.168. In other words, using the value of zero as a threshold would not have allowed us to identify the periods of superior predictability when using the new orders as a predictor.

For Consumer Credit, we find that the AR(p) always outperforms the ADL(p,q) model, as both $\hat{\mu}$ as well as $\hat{\mu} + \hat{\theta}$ are negative, albeit only statistically so in regular times. When financial conditions are tight, there is little statistical difference between the ADL(p,q) model relative to the parsimonious autoregressive benchmark. In addition, across all these models, the identified thresholds are different, yet capture states that dominate at most 35% of the sample period. Furthermore, S_t and the identified threshold enable the researcher to select models and pick next period's most accurate model.

Given that our test rejects the null hypothesis, Table 7 investigates which model performs the best in the presence of state dependence (that is when the predictors are VacancyToUr and New Orders). The table shows the MSFE of the two competing models for the full out-of-sample period

Table 6: Parameter estimates given the presence of state dependence

Frequency	Parameter estimates				Parameter tests			State freq.
	$\hat{\mu}$	$\hat{\theta}$	$\hat{\mu}+\hat{\theta}$	$\hat{\gamma}$	$\hat{\mu}=0$	$\hat{\theta}=0$	$\hat{\mu}+\hat{\theta}=0$	$P(S_t \geq \hat{\gamma})$
VacancyToUR	−0.009	0.096	0.087	−0.050	−0.270	**2.867**	**70.628**	0.359
New Orders	−0.028	0.082	0.054	0.767	−0.862	**2.469**	**44.570**	0.160
Consumer Credit	−0.022	0.005	−0.017	0.136	**−4.035**	0.422	2.145	0.290

Note: The table shows the parameter estimates associated with eq.. (37). The columns labeled $\hat{\mu}, \hat{\theta}, \hat{\mu}+\hat{\theta}$, and $\hat{\gamma}$ show the parameter estimates associated with sup-W. The columns $\hat{\mu}=0$, $\hat{\theta}=0$, and $\hat{\mu}+\hat{\theta}=0$ show the values of the statistic when using a t-test (F-test) for testing the hypothesis that the parameters (or their sum) are equal to zero. The critical values of the F-test with one restriction at the 5% and 10% level are 2.706 and 3.842. Bold numbers indicate a rejection at the 10% level. $P(S_t \geq \hat{\gamma})$ shows the relative frequency of being in the state where both μ and θ are present.

as well as the out-of-sample periods where $S_t < \hat{\gamma}$ and $S_t \geq \hat{\gamma}$. Moreover, in the columns labeled "Rel. MSFE ADL(p,q)/AR(p)," the table shows the relative MSFE of the two models over the two sub-samples. We report relative MSFEs, where the MSFE has been normalized by the full sample MSFE of the AR(p) model. The first row shows the results for VacancyToUr: the MSFE of both the AR(p) and ADL(p,q) increase during times of financial stress (where $S_t \geq \hat{\gamma}$); however, the increase in the MSFE of the ADL(p,q) is smaller than the increase for the AR(P) model. In other words, while both models have a higher MSFE during times of financial stress, the information in the Help-Wanted Index ratio helps reduce the increase in the MSFE. The pattern is the same for new orders.

Figure 6 plots the estimated $\Delta L_{t+1|t}$, highlighting the periods where the economic model performed better (shaded areas). Note that, while the shaded areas coincide with several recessions, they are not necessarily limited to recessionary dates. For instance, the VacancyToUr improved the predictive performance in the mid-1980s and pre-2000s, i.e. outside and before recessions. Thus, the non-linear dependence is not necessarily tied to the state of the business cycle.

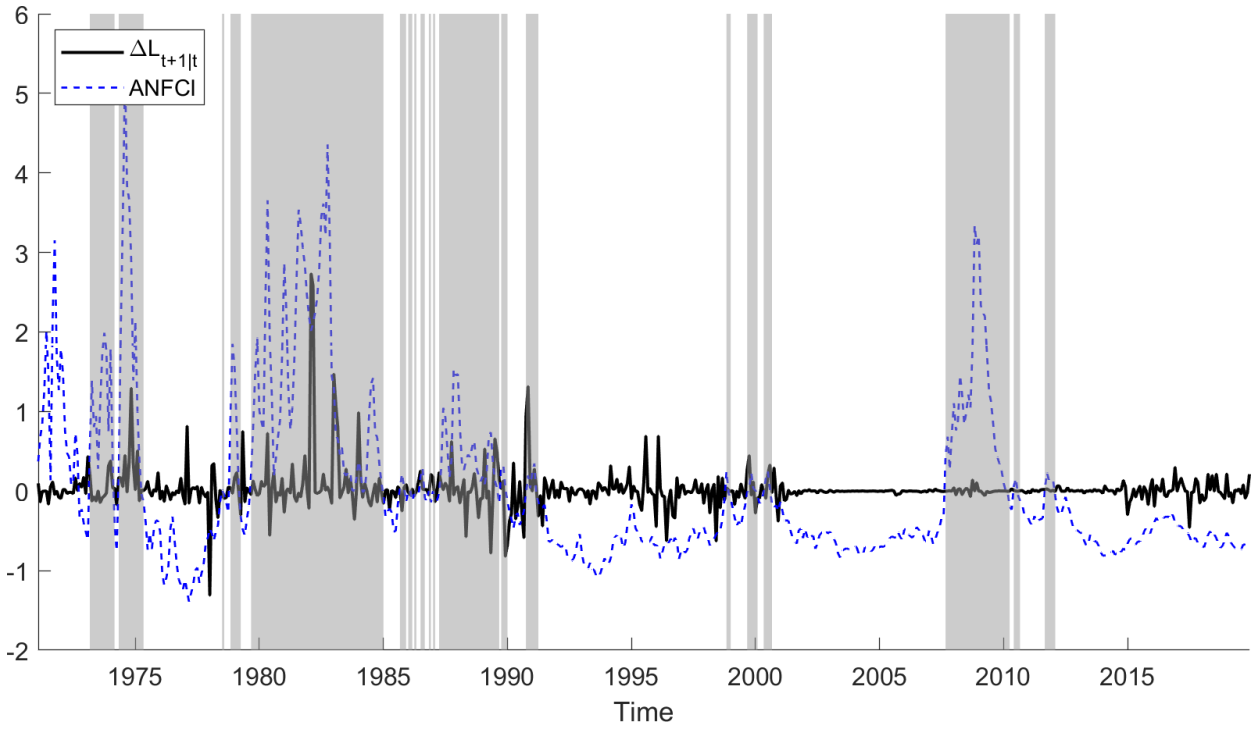
Table 7: MSFEs in subsamples identified via the threshold model

Frequency	AR(p)			ADL(p,q)			Rel. MSFE: ADL(p,q)/AR(p)	
	FS	$S_t < \hat{\gamma}$	$S_t \geq \hat{\gamma}$	FS	$S_t < \hat{\gamma}$	$S_t \geq \hat{\gamma}$	$S_t < \hat{\gamma}$	$S_t \geq \hat{\gamma}$
VacancyToUr	1.000	0.642	1.634	1.003	0.700	1.539	1.091	0.942
New Orders	1.000	0.648	2.824	1.040	0.713	2.736	1.101	0.969
Consumer Credit	1.000	0.631	1.898	1.020	0.662	1.891	1.049	0.996

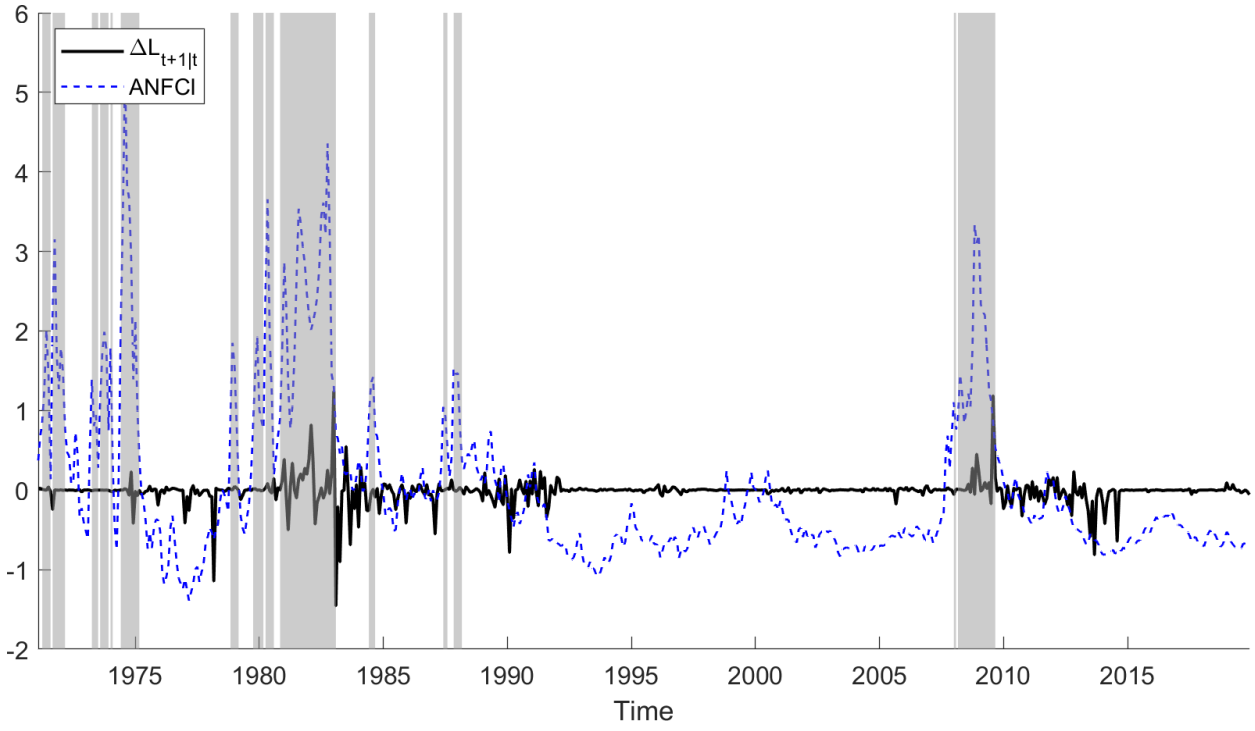
Note: The column labeled "Full" shows the full out-of-sample MSFEs. The column labeled $S_t < \hat{\gamma}$ ($S_t \geq \hat{\gamma}$) shows the MSFE in the subsample where S_t is lower (greater) than the threshold value. The columns labeled "Rel. MSFE: ADL(p,q)/AR(p)" show the ratio of the MSFE of the ADL(p,q) and AR(p) model in the respective subsamples; a number smaller than one indicates a superior performance of the ADL(p,q). To ease the comparison between the subsamples, we normalized all values by the MSFE of the full sample of the AR(p) forecast.

The appendix shows that our results are largely robust to the choice of the in-sample estimation window (see Table D.1 in Appendix D) or including the ANFCI as a linear control variable (see Table D.2 in Appendix D).

Figure 6: State dependence in the relative forecasting performance



(a) Predictor: VacancyToUr



(b) Predictor: New Orders

Note: The figure shows $\Delta L_{t+1|t}$ (solid line, positive numbers indicate that the ADL(p,q) forecasts better than the AR(p)) together with the state variable, ANFCI (dashed line). Grey shaded areas highlight periods where the ADL(p,q) model outperforms the AR(p) benchmark. Panel (a) shows the results when using VacancyToUr as the predictor. Panel (b) shows the results when using New Orders as the predictor.

6 Conclusion

We have developed methodologies that allow researchers to test for the presence of non-linearities in the relative and absolute forecast error losses. Non-linear changes in the forecasting performance that feature non-linearities could be naturally generated in the presence of models with omitted non-linearities. For example, if the true data generating process is a threshold model, the forecast error loss of a model with constant parameters may feature non-linear, threshold-type time-variation. Hence, our paper is extremely useful in situations where the researcher expects the forecast performance to be state dependent. Currently, the only approach available to researchers is Giacomini and Rossi (2010), and, as we show, it is not the best approach to handle non-linear forecast error losses when the non-linearity is a function of an observable variable.

Our testing framework assumes that the parameters of the forecasting models are estimated using a rolling window scheme, i.e. we evaluate forecasting methods rather than forecasting models, and we allow for nested and non-nested models. Results from a Monte Carlo study indicate good size and power properties of the test statistics for moderate sample sizes.

In the first empirical application, we document the existence of state dependence in the relative forecasting performance when predicting stock returns. In particular, a simpler model performs better during times of high real GDP growth, whereas term spread model has a better forecasting performance during times of low real GDP growth. Hence, our results link predictability in returns to current and future output growth. Existing tests, such as the Giacomini and White (2006) and Giacomini and Rossi (2010) tests, cannot detect these "pockets of predictability" as they lack power against state dependence.

Since non-linearities are also widespread in macroeconomic forecasting, in the second empirical application we investigate the relative forecasting performance of an autoregressive versus an autoregressive-distributed lag model for forecasting U.S. industrial production. The choice of our state variable, the adjusted National Financial Conditions Index, is motivated by Adrian et al. (2019) who demonstrated the importance of financial conditions for forecasting the distribution of output growth, particularly tail risk, advocating for a non-linear relationship between financial stability and macroeconomic performance. Our proposed test detects state dependence in the loss differentials of several predictors.

References

Adrian, T., Boyarchenko, N., and Giannone, D. (2019). Vulnerable growth. *American Economic Review*, 109(4):1263–89.

Amisano, G. and Giacomini, R. (2007). Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business & Economic Statistics*, 25:177–190.

Andrews, D. and Ploberger, W. (1994). Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica*, 62:1383–1414.

Andrews, D. W. K. (1991). An empirical process central limit theorem for dependent nonidentically distributed random variables. *Journal of Multivariate Analysis*, 38:188–203.

Andrews, D. W. K. (1993). An introduction to econometric applications of empirical process theory for dependent random variables. *Econometric Reviews*, 12:183–216.

Bekaert, G., Engstrom, E., and Xing, Y. (2009). Risk, uncertainty, and asset prices. *Journal of Financial Economics*, 91:59–82.

Brunnermeier, M. K., Eisenbach, T. M., and Sannikov, Y. (2013). *Macroeconomics with financial frictions: A survey*, volume 2 of *Econometric Society Monographs*, page 3–94. Cambridge University Press.

Campbell, J. Y. and Cochrane, J. (1999). Force of habit: aconsumption-based explanation of aggregate stock market behavior. *Journal of Political Economy*, 107:205–251.

Carrasco, M., Hu, L., and Ploberger, W. (2014). Optimal test for markov switching parameters. *Econometrica*, 82:765–784.

Chauvet, M. and Potter, S. (2013). Forecasting Output. In Elliot, G. and Tmmermann, A., editors, *Handbook of Economic Forecasting*, volume 2, chapter 3, pages 141–194. Elsevier Publications.

Cho, J. S. and White, H. (2007). Testing for regime switching. *Econometrica*, 75:1671–1720.

Clark, T. and McCracken, M. (2001). Tests of Equal Forecast Accuracy and Encompassing for Nested Models. *Journal of Econometrics*, 105:85–110.

Clark, T. and West, K. (2006). Using Out-of-sample Mean Squared Prediction Errors to Test the Martingale Difference Hypothesis. *Journal of Econometrics*, 135:155–186.

Clark, T. and West, K. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138:291–311.

Dangl, T. and Halling, M. (2012). Predictive regressions with time-varying coefficients. *Journal of Financial Economics*, 106:157–181.

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 64:247–254.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 74:33–43.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13:253–263.

Dotsey, M., Fujita, S., and Stark, T. (2018). Do phillips curves conditionally help to forecast inflation. *International Journal of Central Banking*, 14:43–92.

Doukhan, P., Massart, P., and Rio, E. (1995). Invariance principles for absolutely regular empirical processes. *Annales de l'Institut H. Poincaré*, 31:393–427.

Farmer, L., Schmidt, L., and Timmermann, A. (2019). Pockets of predictability. *mimeo*.

Garcia, R. (1998). Asymptotic null distribution of the likelihood ratio test in markov switching models. *International Economic Review*, 39:763–788.

Giacomini, R. and Rossi, B. (2010). Forecast comparisons in unstable environments. *Journal of Applied Econometrics*, 25:595–620.

Giacomini, R. and White, H. (2006). Test of conditional predictive ability. *Econometrica*, 74:1545–1578.

Goyal, A. and Welch, I. (2003). Predicting the equity premium with dividend ratios. *Management Science*, 49:639–654.

Goyal, A. and Welch, I. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21:1455–1508.

Granziera, E. and Sekhposyan, T. (2019). Predicting relative forecasting performance: an empirical investigation. *International Journal of Forecasting*, 35:1636–1657.

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

Hansen, B. E. (1992). The likelihood ratio test under non-standard conditions: Testing the Markov switching model of GNP. *Journal of Applied Econometric*, 7:61–82.

Hansen, B. E. (1996a). Erratum: The likelihood ratio test under non-standard conditions: Testing the Markov switching model of GNP. *Journal of Applied Econometrics*, 11:195–198.

Hansen, B. E. (1996b). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica*, 64:413–430.

Hansen, B. E. (1996c). Stochastic equicontinuity for unbounded dependent heterogeneous arrays. *Econometric Theory*, 12:347–359.

Harvey, D., Leybourne, S. J., Sollis, R., and Taylor, A. M. R. (2021). Real-time detection of regimes of predictability in the US equity premium. *Journal of Applied Econometrics*, 36:45–70.

He, Z. and Krishnamurthy, A. (2011). A model of capital and crises. *The Review of Economic Studies*, 79(2):735–777.

Henkel, S. J., Martin, J. S., and Nardari, F. (2011). Time-varying short-horizon predictability. *Journal of Financial Economics*, 99:560–580.

Inoue, A. and Rossi, B. (2015). Recursive predictability tests for real time data. *Journal of Business and Economic Statistics*, 23:336–345.

Koop, G., McIntyre, S., Mitchell, J., and Poon, A. (2020). Reconciled estimates of monthly GDP in the US. EMF Research Papers 37, Economic Modelling and Forecasting Group.

Liew, J. and Vassalou, M. (2000). Can book-to-market, size, and momentum be risk factors that explain economic growth. *Journal of Financial Economics*, 57:221–245.

McCracken, M. W. (2019). Tests of conditional predictive ability: Some simulation evidence. Working Paper Nr. 011C, Federal Reserve Bank of St. Louis.

McCracken, M. W. and Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business and Economic Statistics*, 34:574–589.

Neely, J. N., Rapach, D. E., Tu, J., and Zhou, G. (2014). Forecasting the equity risk premium: The role of technical indicators. *Management Science*, 60:1772–1791.

Paye, B. and Timmermann, A. (2006). Instability of return prediction models. *Journal of Empirical Finance*, 13:274–315.

Pesaran, M. H. and Timmermann, A. (1995). Predictability of stock returns: Robustness and economic significance. *Journal of Finance*, 50:1201–1228.

Qu, Z. and Zhuo, F. (2020). Likelihood Ratio-Based Tests for Markov Regime Switching. *The Review of Economic Studies*, 88(2):937–968.

Rapach, D., Strauss, J., and Zhou, G. (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *Review of Financial Studies*, 23:821–862.

Rapach, D. and Wohar, M. (2006). Structural breaks and predictive regression models of aggregate U.S. stock returns. *Journal of Financial Econometrics*, 4:238–274.

Rapach, D. and Zhou, G. (2013). Forecasting stock returns. In Elliott, G., Granger, C., and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 2, pages 328–383. Elsevier.

Rossi, B. (2006). Are exchange rates really random walks? Some evidence robust to parameter instability. *Macroeconomic Dynamics*, 10:20–38.

Rossi, B. (2013a). Advances in Forecasting under Model Instability. In Elliot, G. and Tmmermann, A., editors, *Handbook of Economic Forecasting*, volume 2, chapter 21, pages 1203–1324. Elsevier Publications.

Rossi, B. (2013b). Are exchange rates predictable? *Journal of Economic Literature*, 51:1063–1119.

Stock, J. H. and Watson, M. W. (2009). Phillips curve inflation forecasts. In Fuhrer, J., Kodrzycki, Y. K., Sneddon Little, J., and Olivei, G. P., editors, *Understanding Inflation and the Implications for Monetary Policy: A Phillips Curve Retrospective*. MIT Press.

Teräsvirta, T. (2006). Forecasting economic variables with nonlinear models. In Elliott, G., C. G. and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 1, pages 414–457. North-Holland.

Timmermann, A. (2008). Elusive return predictability. *International Journal of Forecasting*, 24:1–8.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, 64:1067–1084.

White, H. (2001). *Asymptotic Theory for Econometricians*. Emerald Group Publishing Limited.

A Theory

A.1 Proof of Proposition 1

Proof of Proposition 1. According to Theorem 3.49 in White (2001), if A_t is α-mixing (or strong mixing) with coefficients of size $-\delta$, $\delta > 0$, so is any measurable function of a finite number of lags of A_t. Under A1(i), $\delta > \nu/(\nu-1)$ and $\nu > 1$, such that $\delta > 0$, and as absolute regularity implies α-mixing, A1(i) implies that any measurable function of a finite number of lags of A_t is absolutely regular. By A1(ii) and A4, $\mathcal{L}_{t+h|t}$ and X_t are measurable functions of a finite number of lags of A_t, and thus, under A1(i), they are absolutely regular with coefficients of size $-\delta$. Consequently, $(\mathcal{L}_{t+h|t}, X_t)$ is strictly stationary and absolutely regular with mixing coefficients $\eta(m) = O\left(m^{-\delta}\right)$ for some $\delta > \nu/(\nu-1)$ and $\nu > 1$, and thus satisfying assumption 1(i) in Hansen (1996b). Further, A2 implies that assumptions 1(ii)-(iii) in Hansen (1996b) hold, and A3 that assumptions 2 and 3 in Hansen (1996b) are satisfied. Thus, under A1 to A4, the result follows from Theorem 1 of Hansen (1996b).

A.2 The case of multiple threshold variables

For now, we have treated the threshold variable S_t as known, and only the threshold γ as unknown. As noted by Hansen (1996b), in practice, the researcher might have several potential threshold variables S_t at hand and needs to decide which variable to include. This case can be naturally accommodated in our framework of the threshold regression model, i.e. when $G(S_t;\gamma) = \mathbb{1}\{S_t \geq \gamma\}$, and we sketch the procedure in the following.

Let D denote a finite set of index numbers, from 1 to $\bar{d}$, for candidate threshold variables, such that $S_t(d)$, $d \in D$, denotes the candidate threshold variable indexed by d. Eq. (1) then becomes

$$\mathcal{L}_{t+h|t} = X_t'\mu + X_t'\theta \cdot \mathbb{1}\{S_t \geq \gamma\} + u_{t+h}. \tag{38}$$

Conditional on a value $(\gamma, d) \in (\Gamma \text{ x } D)$, the estimation of eq. (38) is analogue to that in the model described in eq. (1). Further, and to simplify notation, let all terms of Section 2.3 that are a function of γ be defined analogously as a function of (γ, d). Further, let

$$\mathrm{DM}^{\mathrm{NL}}_{\gamma,\mathrm{D}}: \quad g_{\Gamma,D}(W_p) = \begin{cases} \sup_{d\in D}\sup_{\gamma\in\Gamma} W_P(\gamma,d) \\ \frac{1}{D}\sum_D \int_\Gamma W_P(\gamma,d)\mathrm{d}w(\gamma,d) \\ \ln\left(\frac{1}{D}\sum_D \int_\Gamma \exp(\frac{1}{2}W_P(\gamma,d))\mathrm{d}w(\gamma,d)\right) \end{cases} \tag{39}$$

denote the statistic that takes the supremum over both $\gamma \in \Gamma$ and $d \in D$. It is straightforward to show that the test statistic in eq. (39) has as an asymptotic distribution for point and density forecasts that is analogue to that derived in Proposition 1. We now state the necessary assumptions and then the corollary that accommodates the case of testing for a threshold model when there is more than one candidate threshold variable.

Assumption A.A1 *(i) For all $d \in D$, where D is a finite set of index numbers, $(A_t, X_t, S_t(d))$ is strictly stationary and absolutely regular with mixing coefficients $\eta(m) = O(m^{-\delta})$ for some $\delta > v/(v-1)$ and $v > 1$. (ii) The estimation window size (R) is finite and the estimation scheme is a rolling window*

estimation.

Assumption A.A2 *For $r > v > 1$, $E|Q_t|^{4r} < \infty$, $E|u_t|^{4r} < \infty$, $\inf_{d\in D} \inf_{\gamma\in\Gamma}\ det\big(M(\gamma,\gamma,d,d)\big) > 0$.*

Assumption A.A3 *Let $r > v$ and let S_t have a density function $g(S_t)$ such that $\sup_{s\in\mathbb{R}^{\bar{d}}} g(s) = \bar{g} < \infty$.*

Assumption A.A4 *$f^{(i)}_{t+h|t}(.)$ is a measurable function of lags of A_t, for $i = 1, 2$.*

Corollary 1 *Let $g_{\Gamma,D}(W_p)$ be one of the statistics defined in eq. (39) . Then, under A.A1 to A.A4 and H_0 defined in eq. (5): $E\left(\mathcal{L}_{t+h|t}\right) = 0$ for all $t = R + h, ..., T$, we have*

$$\lim_{P\to\infty} g_{\Gamma,D}\big(W_P(\gamma,d)\big) \to g_{\Gamma,D}\big(\chi^2(\gamma,d)\big), \tag{40}$$

where $\chi^2(\gamma)$ is a chi-square distribution with degrees of freedom $rank(H_r)$, and $g_{\Gamma,D}\big(\chi^2(\gamma,d)\big)$ can be completely characterized by its covariance kernel $K(\gamma_1,\gamma_2,d_1,d_2)$.

Given A.A1 to A.A4, the proof of Corollary 1 follows from Proposition 1, invoking Theorem 3 of Hansen (1996b). The algorithm to simulate the critical values is similar to the algorithm described in Section 2.4, and is given below.
Draw a set of standard Normal random variates $\{v_{t,j}\}_{t=R}^{T-h+B}$:
Simulation Algorithm 2. For each $j = 1, ..., J$, do the following steps:

1. Draw a set of standard Normal random variates $\{v_{t,j}\}_{t=R}^{T-h+B}$:
   (a) Select a threshold variable $S_t(d)$, $d \in D$.
      i. Calculate $\widehat{\lambda}^j_P(\gamma,d) = \frac{1}{\sqrt{P}}\frac{1}{\sqrt{1+B}}\sum_{b=0}^{B}\sum_{t=R}^{T-h}\widehat{s}_{t+h}(\gamma,d)v_{t+b,j}$;
      ii. Using $\widehat{\lambda}^j_P(\gamma,d)$, calculate:
      $W^j_P(\gamma,d) = \widehat{\lambda}^j_P(\gamma,d)'M_P(\gamma,\gamma,d,d)^{-1}H_r\left[H_r'\widehat{V}^*_P(\gamma,d)H_r\right]^{-1}H_r'M_P(\gamma,\gamma,d,d)^{-1}\widehat{\lambda}^j_P(\gamma,d)$;
      iii. Repeat (i)-(ii) for all $\gamma \in \Gamma$;
   (b) Repeat (a) for all $d \in D$;
2. Compute $W^j_P = g_\Gamma\big(W^j_P(\gamma,d)\big)$.

After J iterations, we obtain a set of $\{W^j_P\}_{j=1}^J$ draws from the asymptotic distribution, which we can use to construct critical values and p-values. In particular, the approximate p-value is given by $\widehat{p}(J) = \frac{1}{J}\sum_{j=1}^{J}\mathbb{1}(W_P > W^j_P)$, where W_P denotes the value of the test statistic computed using the actual data.

B Additional Monte Carlo results

B.1 Nested models and density forecasts

This section shows Monte Carlo results for the size of our test statistic for a point forecast comparison of nested models and a density forecast comparison for non-nested models.

Point forecast comparison - nested models:
The underlying data for point forecasts of nested models is generated by

$$y_t = \beta + e_t, \tag{41}$$

with $e_t \sim_{\text{iid}} \text{N}(0,1)$ and β a constant parameter. Let $\widehat{\beta}_t = \frac{1}{R}\sum_{i=t-R+1}^{t} y_i$ denote the OLS estimate of β. The two point forecasts are $\widehat{f}^{(1)}_{t+1|t} = 0$, and $\widehat{f}^{(2)}_{t+1|t} = \widehat{\beta}_t$ respectively. For $\beta = \frac{1}{\sqrt{R}}$, the expected squared forecast error difference is zero in expectation, i.e. the loss differential

$$\Delta L_{t+1|t} = \left(y_{t+1} - \widehat{f}^{(1)}_{t+1|t}\right)^2 - \left(y_{t+1} - \widehat{f}^{(2)}_{t+1|t}\right)^2, \tag{42}$$

is zero in expectation: $\text{E}(\Delta L_{t+1|t}) = 0$ for all $t = R+1, ..., T$.

Density forecast comparison:
The data for the density forecasts comparison is generated by the DGP in eq. (24). The two competing density forecasts are both based on a normal density, given by $\phi(x|b,\sigma^2)$, where x denotes the value at which the density is evaluated, b denotes the conditional mean forecasts, and σ^2 the conditional variance forecast. The two conditional means of the normal densities are the same as the point forecasts in PF1, i.e. $\widehat{b}^{(1)}_{t+1|t} = z^{(1)}_t \widehat{\beta}^{(1)}_t$, and $\widehat{b}^{(2)}_{t+1|t} = z^{(2)}_t \widehat{\beta}^{(2)}_t$, with $\widehat{\beta}^{(j)}_t = \left(\sum_{i=t-R+1}^{t} z^{(j)'}_{i-1} z^{(j)}_{i-1}\right)^{-1} \sum_{i=t-R+1}^{t} z^{(j)'}_{i-1} y_i$ and $z^{(j)}_t = [1, z_{t,j}]$. In turn, the variance forecasts is based on the in-sample estimate of the error variance: $\widehat{\sigma}^{2(j)}_{t+1|t} = \frac{1}{R}\sum_{i=t-R+1}^{t}\left(y_i - z^{(j)}_{i-1}\widehat{\beta}^{(j)}_t\right)^2$. The two density forecasts, both of which are misspecified, are denoted by: $\widehat{f}^{(1)}_{t+1|t} = \phi\left(y_{t+1}|\widehat{b}^{(1)}_{t+1|t}, \widehat{\sigma}^{2(1)}_{t+1|t}\right)$, and $\widehat{f}^{(2)}_{t+1|t} = \phi\left(y_{t+1}|\widehat{b}^{(2)}_{t+1|t}, \widehat{\sigma}^{2(2)}_{t+1|t}\right)$. The loss differential is then given by

$$\Delta L_{t+1|t} = \log\left(\widehat{f}^{(1)}_{t+1|t}(y_{t+1})\right) - \log\left(\widehat{f}^{(2)}_{t+1|t}(y_{t+1})\right), \tag{43}$$

and is zero in expectation: $\text{E}(\Delta L_{t+1|t}) = 0$ for all $t = R+1, ..., T$.

The model we use for evaluating the size of the DM^{NL} is given in eq. (26). Table B.1 shows results for point forecasts of the nested models. We observe some under-rejections, which are in line in terms of their magnitude with the results of McCracken (2019) for nested models. The results of the density forecast comparison, displayed in Table B.2, are overall good; we observe some under-rejection for $R = 25$, as well as some over-rejection for $P \leq 100$, but well-sized results for $R = 100$ and $P \geq 200$.

Table B.1: Size results for a point forecast comparison — nested models

Panel A. ave-W

	TR				ESTR				LSTR			
R/P	50	100	200	1000	50	100	200	1000	50	100	200	1000
25	0.082	0.061	0.042	0.032	0.084	0.051	0.044	0.027	0.089	0.054	0.044	0.035
50	0.094	0.058	0.047	0.035	0.093	0.067	0.041	0.026	0.094	0.060	0.043	0.024
100	0.104	0.062	0.050	0.027	0.100	0.063	0.046	0.031	0.108	0.069	0.048	0.035

Panel B. exp-W

	TR				ESTR				LSTR			
R/P	50	100	200	1000	50	100	200	1000	50	100	200	1000
25	0.100	0.069	0.045	0.036	0.099	0.058	0.046	0.033	0.094	0.059	0.044	0.036
50	0.110	0.064	0.050	0.035	0.109	0.070	0.049	0.035	0.099	0.061	0.044	0.025
100	0.127	0.071	0.049	0.037	0.115	0.069	0.052	0.034	0.115	0.071	0.049	0.037

Panel C. sup-W

	TR				ESTR				LSTR			
R/P	50	100	200	1000	50	100	200	1000	50	100	200	1000
25	0.125	0.084	0.053	0.042	0.135	0.077	0.057	0.041	0.109	0.067	0.048	0.040
50	0.137	0.078	0.057	0.040	0.141	0.074	0.056	0.040	0.107	0.070	0.048	0.034
100	0.152	0.085	0.060	0.044	0.150	0.082	0.058	0.039	0.131	0.077	0.050	0.039

Note: The table displays empirical rejection frequencies of the null hypothesis $H_0 : \mu = \theta = 0$ for the DM^{NL} test for point forecasts of nested models evaluated with the MSFE loss function. The nominal size is 5%. Panels A to C show the results for the three DM^{NL} tests: the sup-W, exp-W and ave-W. The results are based on 3,000 MC replications.

Table B.2: Size results for a density forecast comparison

Panel A. ave-W

	TR				ESTR				LSTR			
R/P	50	100	200	1000	50	100	200	1000	50	100	200	1000
25	0.095	0.059	0.039	0.026	0.083	0.058	0.038	0.023	0.091	0.060	0.048	0.030
50	0.114	0.069	0.050	0.036	0.105	0.072	0.055	0.035	0.110	0.078	0.058	0.036
100	0.113	0.078	0.059	0.053	0.111	0.078	0.062	0.050	0.115	0.085	0.067	0.048

Panel B. exp-W

	TR				ESTR				LSTR			
R/P	50	100	200	1000	50	100	200	1000	50	100	200	1000
25	0.121	0.070	0.051	0.032	0.109	0.070	0.046	0.034	0.097	0.062	0.050	0.032
50	0.149	0.081	0.058	0.041	0.129	0.077	0.062	0.043	0.119	0.080	0.061	0.037
100	0.129	0.083	0.060	0.053	0.129	0.079	0.065	0.051	0.119	0.086	0.068	0.048

Panel C. sup-W

	TR				ESTR				LSTR			
R/P	50	100	200	1000	50	100	200	1000	50	100	200	1000
25	0.150	0.080	0.056	0.040	0.146	0.084	0.057	0.042	0.111	0.075	0.052	0.037
50	0.180	0.103	0.064	0.046	0.161	0.093	0.074	0.050	0.136	0.088	0.067	0.037
100	0.152	0.102	0.064	0.056	0.153	0.088	0.076	0.054	0.131	0.095	0.075	0.044

Note: The table displays empirical rejection frequencies of the null hypothesis $H_0 : \mu = \theta = 0$ for the DM^{NL} test for density forecasts evaluated with the log-score loss function. The nominal size is 5%. Panels A to C show the results for the three DM^{NL} tests: the sup-W, exp-W and ave-W. The results are based on 3,000 MC replications.

B.2 Forecast encompassing test

In this section, we show Monte Carlo results that investigate the finite sample size of a forecast encompassing test. The DGP we assume in our simulations is

$$y_t = \begin{cases} e_t \quad \text{for} \quad t \leq R \\ \beta_{t-1} x_t + \gamma_{t-1} z_t + e_t \quad \text{for} \quad t > R, \end{cases}$$

where R is the in-sample estimation size and e_t, x_t, and z_t are mutually and serially independent standard Normal variates. The coefficients β_t and γ_t are given by $\beta_t = \left(\sum_{i=t-R+1}^{t} x_i y_i\right) / \left(\sum_{i=t-R+1}^{t} x_i^2\right)$ and $\gamma_t = \left(\sum_{i=t-R+1}^{t} z_i y_i\right) / \left(\sum_{i=t-R+1}^{t} z_i^2\right)$. The two competing forecasting models that produce one-step-ahead predictions are $f^{(1)}_{t+1|t} = 0$ and $f^{(2)}_{t+1|t} = \hat{\beta}_t x_{t+1}$, where $\hat{\beta}_t = \left(\sum_{i=t-R+1}^{t} x_i y_i\right) / \left(\sum_{i=t-R+1}^{t} x_i^2\right)$. The two forecast errors are given by $e_{t+1|t,1} = y_{t+1} - f^{(1)}_{t+1|t}$ and $e_{t+1|t,2} = y_{t+1} - f^{(2)}_{t+1|t}$. The moment condition to test for forecast encompassing can be expressed as $ENC_{t+1|t} = e^2_{t+1|t,2} - e_{t+1|t,1} e_{t+1|t,2}$, and we model ENC_t as

$$ENC_{t+1|t} = \mu + \theta \cdot G(S_t; \varphi) + u_t,$$

where $G(\cdot)$ takes the form of the TR, LSTR, or ESTR model, and we test for the null hypothesis of forecast encompassing: $\mu = \theta = 0$. Table B.3 shows the results - size is good for $P \geq 100$ and comparable to the Monte Carlo results for the loss differential. For the sup-W, the test is slightly undersized for $R = 25$ and $P = 1000$, but the rejection frequency is close to the nominal size for $R = 50$ and $R = 100$.

Table B.3: Size results for a forecast encompassing test

Panel A. ave-W

	TR				ESTR				LSTR			
R/P	50	100	200	1000	50	100	200	1000	50	100	200	1000
25	0.084	0.064	0.051	0.051	0.082	0.054	0.047	0.053	0.071	0.062	0.051	0.052
50	0.085	0.066	0.055	0.050	0.083	0.058	0.055	0.051	0.087	0.064	0.059	0.050
100	0.083	0.068	0.059	0.060	0.089	0.062	0.055	0.057	0.084	0.063	0.059	0.055

Panel B. exp-W

	TR				ESTR				LSTR			
R/P	50	100	200	1000	50	100	200	1000	50	100	200	1000
25	0.098	0.054	0.044	0.044	0.087	0.056	0.045	0.041	0.074	0.061	0.051	0.049
50	0.095	0.066	0.049	0.046	0.084	0.050	0.050	0.045	0.091	0.062	0.055	0.048
100	0.091	0.067	0.053	0.056	0.093	0.060	0.051	0.052	0.084	0.060	0.060	0.055

Panel C. sup-W

	TR				ESTR				LSTR			
R/P	50	100	200	1000	50	100	200	1000	50	100	200	1000
25	0.112	0.059	0.046	0.036	0.104	0.055	0.045	0.032	0.078	0.054	0.042	0.045
50	0.109	0.071	0.048	0.042	0.098	0.055	0.048	0.043	0.101	0.069	0.051	0.042
100	0.112	0.074	0.052	0.055	0.114	0.067	0.051	0.051	0.090	0.062	0.055	0.052

Note: The table displays empirical rejection frequencies of the null hypothesis $H_0 : \mu = \theta = 0$ for the DM^{NL} test for forecast encompassing. The nominal size is 5%. Panel A to C show the results for the three DM^{NL} tests: the sup-W, exp-W and ave-W. The results are based on 3,000 MC replications.

C Additional results: pockets of predictability

Table C.1: Loss differential: alternative sizes of R

Variable Name	DM^{NL}			Alternative statistics		Sample size
				Panel A. $R = 180$		
	sup-W	ave-W	exp-W	GW	Fluct.	P
DFY	0.411	0.453	0.468	0.619	> 0.10	612
Inflation	0.850	0.848	0.852	0.803	> 0.10	612
StockVar	0.427	0.278	0.325	0.104	> 0.10	612
LongYield	0.330	0.334	0.334	0.464	$<$ **0.10**	612
Spread	**0.038**	**0.099**	**0.040**	0.810	$<$ **0.05**	612
Tbill	0.286	0.246	0.254	0.445	> 0.10	612
BookToMkt	0.406	0.154	0.197	**0.070**	> 0.10	612
				Panel B. $R = 300$		
	sup-W	ave-W	exp-W	GW	Fluct.	P
DFY	**0.054**	**0.095**	**0.076**	0.224	$<$ **0.05**	492
Inflation	0.333	0.542	0.524	0.864	$<$ **0.10**	492
StockVar	0.569	0.426	0.451	0.162	> 0.10	492
LongYield	0.649	0.560	0.584	0.477	> 0.10	492
Spread	**0.038**	**0.084**	**0.053**	0.770	> 0.10	492
Tbill	0.757	0.692	0.703	0.587	> 0.10	492
BookToMkt	0.158	0.111	0.117	**0.087**	$<$ **0.10**	492

Note: The table shows p-values of tests of equal predictive ability using the DM^{NL}, the GW, and the Fluctuation test (for the Fluctuation test we only indicate whether the p-value is smaller or larger than 0.10 or 0.05). Bold indicates significance at the 10% level. The in-sample estimation window size, R, is 180 and 300 respectively.

Table C.2: Forecast encompassing: alternative sizes of R

Variable Name	DM^{NL}			Alternative statistics		Sample size
				Panel A. $R = 180$		
	sup-W	ave-W	exp-W	GW	Fluct.	P
DFY	0.342	0.267	0.283	0.248	> 0.10	612
Inflation	0.106	**0.066**	**0.085**	**0.094**	$<$ **0.05**	612
StockVar	0.685	0.774	0.767	0.782	> 0.10	612
LongYield	0.767	0.650	0.671	0.382	> 0.10	612
Spread	**0.020**	**0.008**	**0.008**	**0.009**	$<$ **0.05**	612
Tbill	0.405	0.229	0.274	0.106	> 0.10	612
BookToMkt	0.755	0.495	0.533	0.266	> 0.10	612
				Panel B. $R = 300$		
DFY	0.292	0.476	0.455	0.950	> 0.10	492
Inflation	**0.019**	**0.024**	**0.024**	**0.088**	$<$ **0.05**	492
StockVar	0.762	0.815	0.819	0.751	> 0.10	492
LongYield	0.910	0.824	0.838	0.570	> 0.10	492
Spread	**0.027**	**0.023**	**0.026**	**0.027**	$<$ **0.10**	492
Tbill	0.535	0.411	0.445	0.249	> 0.10	492
BookToMkt	0.210	0.199	0.205	0.169	> 0.10	492

Note: The table shows p-values of tests of forecast encompassing using the DM^{NL}, the GW, and the Fluctuation test (for the Fluctuation test we only indicate whether the p-value is smaller or larger than 0.10 or 0.05). Bold indicates significance at the 10% level. The in-sample estimation window size, R, is 180 and 300 respectively.

Table C.3: Testing for state dependence: including the indicator as a linear control variable

Variable Name	DM^{NL}			Alternative statistics		Sample size
			Panel A. Loss differential			
	sup-W	ave-W	exp-W	GW	Fluct.	P
DFY	0.565	0.758	0.721	0.523	$<$ **0.10**	552
Inflation	0.790	0.904	0.904	0.602	$>$ 0.10	552
StockVar	0.281	0.194	0.214	0.238	$>$ 0.10	552
LongYield	0.846	0.783	0.817	0.700	$>$ 0.10	552
Spread	0.144	**0.021**	**0.054**	0.570	$>$ 0.10	552
Tbill	0.808	0.885	0.891	0.896	$>$ 0.10	552
BookToMkt	0.383	0.129	0.204	**0.070**	$<$ **0.10**	552
			Panel B. Forecast encompassing			
	sup-W	ave-W	exp-W	GW	Fluct.	P
DFY	0.324	0.414	0.443	0.416	$>$ 0.10	552
Inflation	0.430	0.441	0.466	**0.054**	$<$ **0.05**	552
StockVar	0.384	0.342	0.349	0.504	$>$ 0.10	552
LongYield	0.709	0.515	0.561	0.260	$>$ 0.10	552
Spread	**0.011**	**0.000**	**0.003**	**0.002**	$<$ **0.05**	552
Tbill	0.489	0.263	0.343	**0.073**	$>$ 0.10	552
BookToMkt	0.574	0.192	0.287	0.163	$>$ 0.10	552

Note: The table shows p-values of tests of equal predictive ability (Panel A) and forecast encompassing (Panel B) using the DM^{NL}, the GW, and the Fluctuation test (for the Fluctuation test we only indicate whether the p-value is smaller or larger than 0.10 or 0.05). When using the DM^{NL} ave-W, we additionally included the indicator variable S_t (monthly real GDP growth), as a linear control. Bold indicates significance at the 10% level. The in-sample estimation window size, R, is 240.

D Additional results: industrial production forecasts

Table D.1: Loss differential: alternative sizes of R

Variable Name	DM^{NL} p-values			Alternative Statistics		Sample Size
	Panel A. $R = 60$					
	sup-W	ave-W	exp-W	GW	Fluct.	P
Housing Starts	**0.079**	0.208	0.102	0.731	> 0.10	587
VacancyToUr	**0.061**	**0.033**	**0.037**	**0.050**	> 0.10	587
Employment	0.389	0.749	0.643	0.820	> 0.10	587
New Orders	**0.018**	**0.043**	**0.033**	**0.068**	$<$ **0.10**	587
Consumer Credit	0.187	0.142	0.154	**0.054**	> 0.10	587
One Year Spread	0.871	0.925	0.921	0.951	> 0.10	587
Ten Year Spread	0.486	0.390	0.436	0.178	> 0.10	587
Credit Spread	0.670	0.591	0.602	0.811	> 0.10	587
	Panel B. $R = 180$					
	sup-W	ave-W	exp-W	GW	Fluct.	P
Housing Starts	**0.049**	**0.057**	**0.055**	0.538	> 0.10	538
VacancyToUr	**0.045**	**0.072**	**0.056**	0.212	$<$ **0.05**	538
Employment	0.235	0.504	0.454	0.725	> 0.10	538
New Orders	0.115	0.182	0.195	0.155	> 0.10	538
Consumer Credit	**0.001**	**0.005**	**0.003**	**0.040**	> 0.10	538
One Year Spread	0.417	0.584	0.585	0.441	> 0.10	538
Ten Year Spread	0.551	0.736	0.737	0.529	> 0.10	538
Credit Spread	0.327	0.267	0.281	0.188	> 0.10	538

Note: The table shows p-values of tests of equal predictive ability using the DM^{NL}, the GW, and the Fluctuation test (for the Fluctuation test we only indicate whether the p-value is smaller or larger than 0.10 or 0.05). Bold indicates significance at the 10% level. The in-sample estimation window size, R, is 60 and 180 respectively.

Table D.2: Testing for state dependence: including the indicator as a linear control variable

Variable Name	DM^{NL} p-values			Alternative statistics		Sample Size
	sup-W	ave-W	exp-W	GW	Fluct.	P
Housing Starts	0.335	0.262	0.270	0.641	> 0.10	587
VacancyToUr	0.163	**0.043**	**0.096**	**0.045**	> 0.10	587
Employment	0.386	0.395	0.444	0.362	> 0.10	587
New Orders	0.136	**0.053**	0.102	**0.039**	> 0.10	587
Consumer Credit	**0.098**	**0.046**	**0.057**	**0.020**	> 0.10	587
One Year Spread	0.329	0.703	0.644	0.430	> 0.10	587
Ten Year Spread	0.374	0.473	0.462	0.422	> 0.10	587
Credit Spread	0.663	0.809	0.804	0.740	> 0.10	587

Note: The table shows p-values of tests of equal predictive ability using the DM^{NL}, the GW, and the Fluctuation test (for the Fluctuation test we indicate whether the p-value is smaller or larger than 0.10 or 0.05). When using the DM^{NL} ave-W, we additionally included the indicator variable S_t (ANFCI), as a linear control. Bold indicates significance at the 10% level. The in-sample estimation window size, R, is 120.

E Forecast evaluation under Markov switching

In this section, we discuss a test for forecast evaluation of relative or absolute forecast error losses in the presence of Markov switching regimes. Note that the notation may differ from the rest of the paper to accommodate the different framework of testing for Markov switching. The test is inspired by Carrasco et al. (2014) (CHP hereafter). The forecast error loss of interest is modeled as:

$$\mathcal{L}_{t+h|t} = \mu + \mu_t + u_{t+h}, \tag{44}$$

where $\mu_t = \mu_S S_t$, S_t is a stationary geometric ergodic two-state univariate first-order Markov chain, μ_S is the magnitude of the change and u_{t+h} is mean zero and satisfies Assumption B1 below. Let $\theta \equiv \{\mu, \sigma^2\}$, and let $\theta_0 \equiv \{\mu_0, \sigma_0^2\}$ denote the parameters under the null, where μ_0 is a parameter of interest and $\sigma_0^2 > 0$ is left unspecified.

Our null hypothesis is described in eq. (5), and is such that $H_0 : E\left(\mathcal{L}_{t+h|t}\right) = 0 \quad \forall t$. Under the model in eq. (44), the null hypothesis can be reparameterized as: $\mu = \mu_S = 0$. Our null hypothesis is different from CHP in two ways: the first is that the latter only test $\mu_S = 0$; the second is that our objective is to test forecast error losses in a relative or absolute forecast performance evaluation, which depends on the estimates of the forecasting models' parameters. Under the alternative $E\left(\mathcal{L}_{t+h|t}\right) \neq 0$, which again can be due to either Markov switching or constant and unequal forecast performance.

To derive the asymptotic distribution of the test, we require the following additional assumptions:

Assumption B1 *The latent variable μ_t is defined as $\mu_t = \mu_S S_t$, where μ_S is a finite scalar constant, S_t is a stationary geometric ergodic finite-state univariate first-order Markov chain with $var(S_t) = 1$ and covariance $cov\left(S_t, S_{t-i}\right) = \rho^i$, $\rho \neq 0$ and $-1 < \rho < 1$. Furthermore, μ_t is strongly exogenous to A_t, $t = 1, ..., T$, such that the joint likelihood of $A_1, ..., A_T, \mu_1, ..., \mu_T$ factorizes as $\Pi_{t=1}^{T} f\left(A_1, ..., A_T; \theta\right) q\left(\mu_t | \mu_{t-1}, ..., \mu_1; \rho\right)$*

and the values of $\mu_t + \mu$ belong to some compact set containing μ.

Assumption B2 *Let the conditional log-density of $\mathcal{L}_{t+h|t}$ be Normal and be denoted by ℓ_t under the null hypothesis. Let N_0 be a neighborhood around θ_0, where θ_0 is an interior point of N_0; the information matrix $\mathcal{I}(\theta_0) = E_0\left(||\ell_t^{(1)}(\theta_0)\,\ell_t^{(1)}(\theta_0)'||^{20}\right)$ is nonsingular.*

For convenience, we maintain Assumption A1. Assumption B1 specifies the behavior of the time variation and it requires that, under the null, the distribution of the data A_t and that of η_t are mutually independent. Assumption B2 makes a convenient distributional assumption which implies that the asymptotic distribution of our test statistic is the same as in CHP — for details see the proof of Proposition 2. The following proposition provides the result of our test for Markov switching in the forecast error losses.

Proposition 2 *Let DM^{NL}: $g(TS_P) = \sup_{\rho\in[\underline{\rho},\overline{\rho}]} TS_P(\rho)$, and $TS_P(\rho) = \frac{1}{2}\left(\max\left(0, \frac{\Gamma_P^*(\rho)}{\sqrt{\widehat{\xi}(\rho)'\widehat{\xi}(\rho)}}\right)\right)^2$, where $\Gamma_P^*(\rho) = P^{-1/2}\sum_t \eta_t^*\left(\rho, \widehat{\theta}_0\right)$, and*

$$\eta_t^*(\rho,\theta) = \frac{1}{2}\left\{\left[\ell_t^{(2)}(\theta) + \ell_t^{(1)}(\theta)\,\ell_t^{(1)}(\theta)'\right] + 2\sum_{\tau<t}\rho^{(t-\tau)}\ell_t^{(1)}(\theta)\,\ell_\tau^{(1)}(\theta)'\right\}.$$

$\widehat{\xi}(\rho)$ is the residual of a regression of $\eta_t\left(\rho, \widehat{\theta}_0\right)$ on $\ell_t^{(1)}\left(\widehat{\theta}_0\right)$ and $\widehat{\theta}_0$ is the constrained ML estimator of θ under the null. Then, under A1, B1, B2 and H_0 defined in eq. (5): $E\left(\mathcal{L}_{t+h|t}\right) = 0$ for all $t = R+h, ..., T$:

$$g(TS_P) = \sup_{\rho\in[\underline{\rho},\overline{\rho}]} TS_P(\rho) \underset{d}{\to} \sup_{\rho\in[\underline{\rho},\overline{\rho}]} \frac{1}{2}\left(\max(0,K)\right)^2, \tag{45}$$

where $K = sign(\rho)\sqrt{1-\rho^2}\sum_{i=0}^{\infty}\rho^i Z_i$, where $sign(\rho) = 1$ if $\rho > 0$, zero if $\rho = 0$ and equal to -1 if $\rho < 0$, and Z_i are iid standard Normal variables. The DM^{NL} test rejects H_0 defined in eq. (5) when $g_\Gamma(TS_P) > \phi_\alpha$, where ϕ_α is the critical value (for a nominal size of α) in Table E.1 below, where either $\underline{\rho} = -0.7$, $\overline{\rho} = 0.7$ or $\underline{\rho} = -0.98$, $\overline{\rho} = 0.98$.

Proof of Proposition 2. From a similar argument as that in the proof of Proposition 1, since the forecast errors loss $\mathcal{L}_{t+h|t}$ is a measurable functions of a finite number of lags of A_t, under A1(i) they are absolutely regular with coefficients of size $-\delta$. Consequently, $\mathcal{L}_{t+h|t}$ is strictly stationary and absolutely regular with mixing coefficients $\eta(m) = O\left(m^{-\delta}\right)$ for some $\delta > \nu/(\nu-1)$ and $\nu > 1$. Under Assumptions A1, B1 and B2, the assumptions in CHP hold. In particular, let $cov()$ denote the covariance and let

$$\begin{aligned} d^*(\rho) \equiv d^*(\rho,\theta_0) &= \mathcal{I}(\theta_0)^{-1} cov\left(\eta_t^*(\rho,\theta_0), \ell_t^{(1)}(\theta_0)\right) \\ &= \mathcal{I}(\theta_0)^{-1} cov\left(\eta_t^*(\rho,\theta_0), \begin{pmatrix} \ell_{\mu,t}^{(1)}(\theta_0) & \ell_{\sigma^2,t}^{(1)}(\theta_0) \end{pmatrix}\right) \\ &= \mathcal{I}(\theta_0)^{-1}\left(\, cov\left(\eta_t^*(\rho,\theta_0), \ell_{\mu,t}^{(1)}(\theta_0)\right),\ cov\left(\eta_t^*(\rho,\theta_0), \ell_{\sigma^2,t}^{(1)}(\theta_0)\right)\,\right). \end{aligned}$$

Under normality, $\eta_t^*(\rho,\theta) = \frac{1}{2\sigma^4}\left[\left(u_{t+h}^2 - \sigma^2\right) + 2\sum_{\tau<t}\rho^{(t-\tau)}u_{t+h}u_{\tau+h}\right]$ and $\ell_{\mu,t}^{(1)}(\theta_0) = u_{t+h}/\sigma_0^2$; therefore, $cov\left(\eta_t^*(\rho,\theta_0), \ell_{\mu,t}^{(1)}(\theta_0)\right) = 0$. Furthermore, because of the Normality assumption,

the matrix $\mathcal{I}(\theta_0)^{-1}$ is block diagonal. Thus, the first element of the vector $d^*(\rho)$ equals zero. This implies that $d^*(\rho)' \ell_t^{(1)}\left(\widehat{\theta}_0\right) = 0$ since (i) we just showed that the first element of $d^*(\rho)$ is zero; and (ii) the second element of $\ell_t^{(1)}\left(\widehat{\theta}_0\right) = 0$ because this component of the score is evaluated at the constrained MLE of σ^2. Thus, $\Gamma_P^*(\rho) = P^{-1/2}\sum_t \eta_t^*\left(\rho, \widehat{\theta}_0\right) = P^{-1/2}\sum_t \left(\eta_t^*\left(\rho, \widehat{\theta}_0\right) - d^*(\rho)' \ell_t^{(1)}\left(\widehat{\theta}_0\right)\right)$, as $d^*(\rho)' \ell_t^{(1)}\left(\widehat{\theta}_0\right) = 0$. Consequently, the arguments of Lemma C.1 of Carrasco et al. (2014) apply to eq. (45), and the results follow from Theorem 3.1 of Carrasco et al. (2014).

Table E.1: Critical Values (ϕ_α)

α	$\rho \in [-0.7, 0.7]$	$\rho \in [-0.98, 0.98]$
1%	3.96	4.52
5%	2.45	2.99
10%	1.82	2.32

Note: The critical values are taken from Carrasco et al. (2014).

Table E.2 reports size results for the test $g(TS_P)$ using the data generated according to the DGP given in Section 3.1 (labeled DGP1) and the point forecast comparison of nested models of Appendix B.1 (labeled DGP2). The table shows that the test tends to over-reject in small samples (P< 200, R < 25), but is well-sized for larger sample sizes. As before, large sample results for the nested model case of DGP2 are slightly undersized.

Table E.2: Size results for forecast comparison test with Markov switching

	DGP1								DGP2							
	Size 5 %				Size 10 %				Size 5 %				Size 10 %			
R/P	50	100	200	1000	50	100	200	1000	50	100	200	1000	50	100	200	1000
25	0.088	0.078	0.064	0.065	0.161	0.142	0.127	0.119	0.059	0.046	0.035	0.028	0.116	0.086	0.076	0.066
50	0.072	0.065	0.062	0.052	0.137	0.129	0.120	0.116	0.069	0.049	0.039	0.028	0.133	0.099	0.088	0.067
100	0.074	0.060	0.053	0.052	0.143	0.118	0.110	0.113	0.077	0.059	0.044	0.032	0.147	0.108	0.091	0.068

Note: The table displays empirical rejection frequencies of the null hypothesis $H_0 : \mu = \mu_S = 0$ for the DM^{CHP} test. Size 5% and 10% denote the nominal size. DGP1 is given in Section 3.1, and DGP2 refers to the point forecast comparison of the nested models given Appendix B.1. The results are based on 1,000 MC replications and using the critical values of Table E.1 with $\rho \in [-0.98, 0.98]$.

For a model that includes autoregressive components the distribution depends on the autoregressive component and can be derived from the asymptotic distribution of $\sup_{\rho \in [\underline{\rho}, \bar{\rho}]} v_T(\theta_0, \rho)$. For an AR(1), similar to Carrasco et al. (2014), the critical values can be simulated from:

$$\frac{\sqrt{1-\rho^2}\,|1-\rho\phi|}{|\rho-\phi|}\Big[\sum_{i=0}^{\infty}\rho^i Z_i - \frac{(1-\phi^2)}{(1-\phi\rho)}\sum_{i=0}^{\infty}\phi^i Z_i\Big]. \tag{46}$$