Check for updates

Journal of Time Series Analysis

WILEY

ORIGINAL ARTICLE OPEN ACCESS

Forecast Evaluation for Functional Data

Jack Fosten[1] | Piotr Kokoszka[2] | Tim Kutta[3] | Shixuan Wang[4]

[1]Bayes Business School, City St George's University of London, London, UK | [2]Colorado State University, Fort Collins, Colorado, USA | [3]Aarhus University, Aarhus, Denmark | [4]University of Reading, Reading, UK

Correspondence: Jack Fosten (jack.fosten@city.ac.uk)

Received: 21 April 2026 | **Revised:** 7 October 2026 | **Accepted:** 10 October 2026

Keywords: forecast evaluation | functional data | predictive

ABSTRACT

This paper proposes methods for comparing the accuracy of two competing sets of functional forecasts. This is increasingly relevant as many economic and financial forecasters are interested in predicting variables which are observed as functional data objects. However, to date there has been very little research on formal statistical tests to evaluate the relative accuracy of competing functional forecasts. We propose a suite of novel tests, building on the classic Diebold-Mariano test, to provide formal statistical guidance on forecast accuracy in the case of functional data. We derive the asymptotic properties of the tests, including their self-normalised versions, and demonstrate the validity of analytic or bootstrap-based critical values. We investigate the finite sample performance of the tests using Monte Carlo simulations. We illustrate the use of our method by comparing forecasts of the U.S. yield curve based on forward rates with a random walk benchmark.

JEL Classification: C22, C12, C53

1 | Introduction

Forecast evaluation plays a central role in economics and finance, where policymakers and practitioners rely heavily on forecasts to guide decisions. The Diebold-Mariano (DM) test, introduced by Diebold and Mariano (1995), remains one of the most widely used tools for comparing the accuracy of competing forecasts. Subsequently, a large body of extensions to the DM test have focussed on comparing forecasts of scalar time series variables such as GDP growth, inflation rates or asset returns. Elsewhere, there has been an increasing prevalence of functional data in modern forecasting applications. Functional forecasts arise naturally in areas such as yield curve modelling, electricity demand forecasting, and intra-day volatility prediction, where the object of interest is a function observed in every time period over a continuous domain (time to maturity, time of day and so on). However, despite the growing empirical importance, there has been very little formal development of forecast accuracy tests designed specifically for functional data, with existing empirical papers such as Shang and Kearney (2022) instead taking scalar DM tests and performing them at isolated points along the function domain.

This paper addresses this gap by developing a new framework for evaluating and comparing two competing functional forecast methods. Building on the idea of the classic DM test, we propose test statistics that operate on loss differentials defined over both time and the functional domain. In particular, we consider a pointwise loss function on forecast errors across both time periods $t = 1, \ldots, T$ and function arguments $\tau \in \mathcal{T}$, leading to a functional loss differential process. This richer structure raises new challenges relative to the scalar forecast evaluation case. The main issue is that forecast accuracy may vary across regions of the functional domain, such that one method outperforms another only on specific sub-intervals (for instance short versus long maturities of interest rates). This means that the traditional notion of pairwise equal predictive ability tests does not translate directly to functional forecast comparison.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Journal of Time Series Analysis* published by John Wiley & Sons Ltd.

To overcome this, we propose three novel tests which have different interpretations: (1) a uniform test for strict dominance across the entire domain; (2) an average test, which equally aggregates loss differences over the domain; and (3) a weighted test, which allows users to supply application-specific importance weights to different parts of the domain (e.g., allowing a fixed-income investor to place greater emphasis on forecasting short maturities than long maturities). The proposed tests complement each other, and the choice of test depends on the decision problem faced by the user. Together, they provide a comprehensive toolkit for forecast comparison in a functional setting.

We derive test statistics and critical values to perform these hypothesis tests which are straightforward to compute and implement. In the average and weighted case, the test statistics are simple (weighted) averages of pointwise DM tests and the critical values can be obtained from the quantiles of the standard normal distribution. On the other hand, the uniform test is based on the supremum of pointwise DM statistics and a bootstrap procedure is outlined for obtaining the quantiles of its limiting distribution. To give large sample justifications for our statistical methodology, we develop some new probabilistic tools. First, we extend the temporal dependence measure known as L^p-m-approximability to Banach spaces of (Hölder-)continuous functions. By developing a theory for L^p-m-approximable functional time series in such spaces, we automatically cover a wide set of relevant time series dependence structure. Second, we develop new weak invariance principles for the partial sum process on function spaces. In contrast to known results, our convergence holds in a very strong sense (in a Hölder-topology, specified below) in both time and functional component. Consequently, we expect these results to have independent use beyond this paper, for instance in change point analysis, where convergence with respect to stronger norms often allows the use of more effectively weighted statistics.

In addition to our main test statistics and testing procedures, we also derive a self-normalised test in the spirit of Shao (2015). The standard DM-test requires the estimation of the long-run variance of the mean loss differential, typically using a heteroskedasticity and autocorrelation consistent (HAC) estimator such as Newey and West (1987). The self-normalised statistic uses a normaliser which has the effect of cancelling the long-run variance, meaning we can avoid computing the long run variance entirely.

We assess the finite sample performance of these testing procedures through Monte Carlo simulations. We find that the tests have desirable size and power properties, in line with our theory, for a range of sample sizes and functional data generating processes. As a practical illustration of our method, we then consider forecasting the U.S. yield curve. Specifically, across the entire maturity spectrum of overnight index swaps (OIS), we compare a random walk benchmark with forecasts based on forward rates. In terms of three-month ahead forecast of yield curves, our tests find no evidence that the forward rates provide superior forecasts, though there is some evidence of superiority at very long forecast horizons. We also break the analysis down into yearly subsamples and find that the forward rates did provide superior forecasts in some years such as 2022 when there were systematic increases in the U.S. federal funds rate.

This paper adds to a long and extensive body of research on forecast evaluation tests dating back to Diebold and Mariano (1995). Extensions of the DM test have been developed to accommodate: model-based comparisons (West 1996; McCracken 2000), nested model comparisons (Clark and McCracken 2001, 2005; Pitarakis 2025), small sample considerations (Coroneo and Iacone 2020; Coroneo et al. 2024), overlapping models (Clark and McCracken 2014; Corradi et al. 2024), generated regressors (Gonçalves et al. 2017), multiple models (White 2000; Romano and Wolf 2005; Hansen et al. 2011) and so on. Meanwhile, there have been empirical forecasting studies using functional data (for instance Horváth et al. 2020) which report forecast evaluation metrics like MSFE across models but without any measure of statistical significance. Our paper builds on this literature by providing the first extension of the DM paradigm to the context of functional forecasts.

The rest of the paper is organised as follows. Section 2 outlines the set-up of the problem and provides details of the hypothesis tests, test statistics and test procedure. Section 3 provides the theoretical analysis and establishes the asymptotic properties of the test statistics. Sections 4 and 5 show implementations of the tests in simulated and empirical settings, respectively. Finally, Section 6 concludes the paper. The proofs of the theoretical results can be found in the Supporting Information.

2 | Set-Up and Tests

The variable we wish to forecast is the functional object $Y_t(\tau)$ for $\tau \in \mathcal{T}$, where the index t refers to the time period (e.g., day). The observations of this function could be "sparse", for example in yield curve analysis where data can only be readily obtained for certain maturities such as 3-month, 6-month and so forth. Such maturities are in the domain $\mathcal{T}$. On the other hand, observations could be "dense" like in financial high frequency data. The index $\mathcal{T}$ then denotes the intraday time.

Suppose that, after shifting time t, we observe a realisation of the target variable $\{Y_t(\tau), t = 1, 2, \ldots, T\}$ for which two competing sets of functional forecasts $Y_t^{(1)}(\tau)$ and $Y_t^{(2)}(\tau)$ are observed across time periods $t = 1, \ldots, T$. These forecasts are assumed to be directly observed and model-free, which could include predictions like the no-change forecast, rolling window average forecasts with different window lengths, professional forecasts, or forecasts based on futures prices in contexts like interest rate derivatives or commodity price prediction.

In the following, without loss of generality, we assume that the first set of forecasts, $Y_t^{(1)}(\tau)$, are labelled as the "benchmark" method, and the second set of forecasts, $Y_t^{(2)}(\tau)$, results from an alternative method which is in competition with the benchmark. We would like to establish superiority of the benchmark or the alternative method with statistical confidence.

To do this, following Diebold and Mariano (1995), we consider a general loss function ℓ and compare $\ell\left(Y_t^{(1)}(\tau), Y_t(\tau)\right)$ to $\ell\left(Y_t^{(2)}(\tau), Y_t(\tau)\right)$ via the functional loss differential:

$$d_t(\tau) = \ell\left(Y_t^{(1)}(\tau), Y_t(\tau)\right) - \ell\left(Y_t^{(2)}(\tau), Y_t(\tau)\right).$$

14679892, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jtsa.70087, Wiley Online Library on [05/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

This is a pointwise loss function that depends on τ and does not reflect the behaviour of functions across the entire domain $\mathcal{T}$. In the functional setting we may want to study regions where one forecast may be superior to another. For example, one method may predict short term interest rates (maturities within one year) better, but not the whole yield curve. Later, we will demonstrate different ways of combining the loss differential across τ to perform a single test of predictive ability.

There are many possible candidate functions one can consider for $\ell(.,.)$, which is a choice based on the type of forecasts under study. The most popular is perhaps the squared error loss function, which leads to

$$d_t(\tau) = \left(Y_t^{(1)}(\tau) - Y_t(\tau)\right)^2 - \left(Y_t^{(2)}(\tau) - Y_t(\tau)\right)^2.$$

However, there are many other possibilities including the absolute error loss function, the check loss function for quantile forecast evaluation or other asymmetric loss functions like linear exponential loss (Linex).

We make assumptions directly on the loss differential, in line with the original idea of Diebold and Mariano (1995). We thus assume that the functional loss differential admits the following decomposition:

$$d_t(\tau) = \mu(\tau) + e_t(\tau), \qquad t = 1, \ldots, T, \tag{2.1}$$

where $\mu(\tau)$ is the mean difference between the forecast losses at argument τ, and the errors $e_t(\tau)$ form a mean-zero weakly dependent functional time series.

The addition of the functional dimension in this pairwise forecast comparison is a crucial point of departure from the existing forecast evaluation literature for scalar target variables. In the standard DM approach, the typical hypothesis test is for pairwise equal predictive ability (EPA) where expected forecast losses are conjectured to be strictly equal under the null hypothesis. The extension of this notion to an entire function is delicate because there may be arguments τ where EPA locally holds ($\mu(\tau) = 0$) and others where it does not hold ($\mu(\tau) \neq 0$). As a result, we will subsequently investigate a variety of functionals that aggregate the entire loss differential into a single quantity. There are a number of such formulations which can be considered within model (2.1) to take account of the functional nature of the loss differentials. The choice of the functional is problem driven and reflects the emphasis put on different aspects of prediction (e.g., emphasis on long or short time horizons in various term structures).

2.1 | The Uniform Hypothesis

The first test assesses whether the benchmark method is not worse than the competitor method at any τ:

$$H_0(U): \qquad \sup_{\tau \in \mathcal{T}} \mu(\tau) \leq 0. \tag{2.2}$$

Under the alternative hypothesis, the competitor improves over the benchmark for some part of the interval $\mathcal{T}$:

$$H_A(U): \qquad \mu(\tau) > 0, \quad \text{for some } \tau \in \mathcal{T}.$$

This test assesses whether the benchmark forecasting method is uniformly superior, or at least no worse, than the competing method across the entire functional domain under the null. The alternative allows the competing method to outperform the benchmark on some subset of the domain. The hypotheses therefore have a similar logical structure to the superior predictive ability tests commonly used for scalar forecast comparison with multiple models (see White 2000), although our framework remains a pairwise comparison between a benchmark and a single competing forecasting method. In our setting, it is the functional dimension, rather than the presence of multiple competing forecasting methods, that motivates the use of a uniform superiority hypothesis rather than an EPA hypothesis.

A natural test statistic for the hypothesis $H_0(U)$ is

$$S_T^U = \sup_{\tau \in \mathcal{T}} \sqrt{T}\,\bar{d}(\tau), \bar{d}(\tau) = \frac{1}{T}\sum_{t=1}^{T} d_t(\tau), \tag{2.3}$$

which makes use of the sample average loss differential $\bar{d}(\tau)$.

One potential drawback of the uniform test is that it may be considered too "strict" in the sense that a small portion of the domain could drive a rejection of the null. This may not necessarily be a drawback, depending on user preferences and the specific problem. Nevertheless, it is also useful to consider other test formulations which could be considered less strict.

2.2 | The Mean Hypothesis

The second test uses the mean performance of the two forecasting methods across the function domain. Under the null hypothesis, the mean performance of the benchmark is superior or equal to that of the competitor:

$$H_0(M): \qquad \int_{\mathcal{T}} \mu(\tau) d\tau \leq 0.$$

Under the alternative hypothesis, the competitor improves over the benchmark on average across the domain $\mathcal{T}$:

$$H_A(M): \qquad \int_{\mathcal{T}} \mu(\tau) d\tau > 0.$$

A natural test statistic for the hypothesis $H_0(M)$ is:

$$S_T^M = \int_{\mathcal{T}} \sqrt{T}\,\bar{d}(\tau) d\tau,$$

where $\bar{d}(\tau)$ is the sample average loss differential, defined in (2.3).

Evidently, the uniform hypothesis $H_0(U)$ implies the less restrictive mean hypothesis $H_0(M)$ since if the benchmark is superior or equal at every point τ, then it is also superior or equal on average. Both hypotheses have different interpretations and the corresponding tests proposed in this paper have distinct uses. For instance, it might be true that method 2 beats the benchmark method 1 in some point $\tau^+ \in \mathcal{T}$, leading to a violation of $H_0(U)$ with $\mu(\tau^+) > 0$. Yet, this local violation may be averaged out over the entire domain, leading to no violation of $H_0(M)$ if

14679892, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jtsa.70087, Wiley Online Library on [05/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$\int_{\mathcal{T}} \mu(\tau)d\tau \leq 0$. Which hypothesis is more appropriate will depend on the specific application in hand. In some applications, any superiority of forecasting method 2 will be important and in this case the user should test $H_0(U)$. On the other hand, if average performance over the entire domain is of importance then $H_0(U)$ can be too restrictive and testing $H_0(M)$ may be preferred.

Finally, there might be cases in between these two approaches, where all parts of the domain matter, but some matter more than others. For instance, if a fixed-income investor is more focused on the short range of the term structure (e.g., less than 1 year), this priority should be reflected in the hypothesis under investigation. To capture priorities in the domain, we introduce the class of weighted integrals.

2.3 | The Weighted Hypothesis

We now introduce the weighted hypothesis, which provides a flexible compromise between the uniform and mean hypotheses. In many applications, all parts of the functional domain are relevant, but some regions are naturally more important than others from an economic or decision-making perspective. The choice of the weight function $w(\tau)$ therefore allows the researcher to encode these priorities directly into the hypothesis being tested. For example, in yield curve forecasting, a fixed-income investor with a short investment horizon may place substantially greater weight on maturities below one year than on long-dated securities, leading to a weight function that declines with maturity. Conversely, a pension fund or insurance company managing long-duration liabilities may choose a weight function that increases with maturity. Elsewhere, in applications like electricity price forecasting, where τ indexes the time of day, a system operator may wish to assign greater weight to peak-demand periods than to overnight hours, reflecting the larger economic costs associated with forecasting errors during periods of network congestion. The framework developed below accommodates any non-negative weighting function, allowing the hypothesis under investigation to be tailored to the forecasting objective.

To develop a weighted test, we note that the mean test from the previous section is based on the linear functional

$$\Lambda_M : f \mapsto \int_{\mathcal{T}} f(\tau)d\tau,$$

which maps a function f to its integral. Thus, the hypothesis $H_0(M)$ can be formulated as: $H_0(M) : \Lambda_M[\mu] \leq 0$. We also notice that Λ_M is a positive functional, in the sense that for any non-negative function f it yields a non-negative value, that is,

$$f(\tau) \geq 0 \forall \tau \in \mathcal{T} \Rightarrow \Lambda_M[f] = \int_{\mathcal{T}} f(\tau)d\tau \geq 0.$$

Formally, we can now substitute Λ_M by any alternative, positive linear functional Λ_W, which yields a new class of hypotheses of the form

$$H_0(W) : \Lambda_W[\mu] \leq 0 \quad \text{versus} \quad H_A(W) : \Lambda_W[\mu] > 0. \tag{2.4}$$

According to Riesz representation theorem on function spaces, Λ_W can be expressed as an integral over f w.r.t. to a unique, positive Borel measure W, such that:

$$\Lambda_W[f] \coloneqq \int_{\mathcal{T}} f(\tau)dW(\tau).$$

An important special case arises if W is characterised by a continuous or discrete weight function w. For the continuous case, consider $w : \mathcal{T} \to [0, \infty]$ and therewith define the weighted integral:

$$\Lambda_W[f] \coloneqq \int_{\mathcal{T}} f(\tau)w(\tau)d\tau,$$

where $w(\tau) \equiv 1$ yields the mean functional $\Lambda_W = \Lambda_M$ as a special case. A choice of w tailored to the application in hand will allow users to prioritise prediction performance in some parts of $\mathcal{T}$ over others. In the discrete case, W can be characterised by values at K points in $\mathcal{T}$, leading to

$$\Lambda_W[f] = \sum_{k=1}^{K} f(\tau_k)w(\tau_k).$$

As a natural statistic to test linear hypotheses of the form (2.4), we consider

$$S_T^W \coloneqq \sqrt{T}\Lambda_W[\overline{d}].$$

2.4 | Test Decisions

In this section, we provide a practical outline of how to implement the test procedures. In the next section, we will derive the asymptotic behaviour of the three test statistics. We start with the case of the weighted hypothesis which includes the mean hypothesis as a special case. For the hypotheses in (2.4), we first consider the test decision:

$$\text{Decision (1)}: \quad \text{Reject if} \quad \frac{\sqrt{T}\Lambda_W[\overline{d}]}{\widehat{\sigma}} > \Phi^{-1}(1-\alpha).$$

We use a studentised statistic where $\hat{\sigma}^2$ is an estimator of the long-run variance (LRV) of the mean of $\{\Lambda_W[d_t] : 1 \leq t \leq T\}$ and Φ^{-1} refers to the quantile function of the standard normal distribution.

We will show that it is possible to avoid the problem of estimating $\hat{\sigma}$, by using a self-normalised test statistic (see Shao 2015). In the context of time series analysis, self-normalisation refers to the division of a statistic of interest (in our case $\Lambda_W[\overline{d}]$) by a "normaliser" that cancels out the long-run variance and hence makes the normalised statistic asymptotically pivotal. It is important to emphasise that self-normalisation is not equivalent to long-run variance estimation, since the normaliser does not converge to a fixed number, but to a non-degenerate random variable that is proportional to the long-run variance. Self-normalisation eliminates the need to estimate the LRV.

In our case, we define the normalised version of S_T^W as:

$$S_T^{SN} \coloneqq \frac{\sum_{t=1}^{T}\Lambda_W[d_t]}{\left\{\frac{1}{T}\sum_{t=1}^{T}\frac{t^2}{T^2}\left(\sum_{s=1}^{t}\Lambda_W[d_s] - \frac{t}{T}\sum_{r=1}^{T}\Lambda_W[d_r]\right)^2\right\}^{1/2}}.$$

14679892, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jtsa.70087, Wiley Online Library on [05/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

We now define the test decision:

$$\text{Decision (2)}: \quad \text{Reject if} \quad S_T^{SN} > \Psi_T^{-1}(1-\alpha)$$

where $\Psi_T^{-1}(1-\alpha)$ is the upper α quantile of the distribution of:

$$\tilde{S}_T^{SN} := \frac{\sum_{t=1}^T Z_t}{\left\{\frac{1}{T}\sum_{t=1}^T \frac{t^2}{T^2}\left(\sum_{s=1}^t Z_s - \frac{t}{T}\sum_{r=1}^T Z_r\right)^2\right\}^{1/2}}.$$

The terms $Z_1, \ldots, Z_T$ are i.i.d. standard normal, so the distribution can be readily simulated. We will show in the next section that $\tilde{S}_T^{SN} \xrightarrow{d} \tilde{S}_\infty^{SN}$ and thus, we can use the quantiles of $\tilde{S}_T^{SN}$ based on a very large number of simulation runs, given in Table S1.

In the context of the uniform hypothesis, the limit of the test statistic S_T^U is not pivotal, but depends on the entire long-run covariance operator of the time series and the set argmax(μ). Indeed it holds that,

$$\sqrt{T}\{\sup_{\tau\in\mathcal{T}}[\overline{d}(\tau)] - \sup_{\tau\in\mathcal{T}}[\mu(\tau)]\} \to \sup_{\tau\in\text{argmax}(\mu)} G(\tau), \tag{2.5}$$

where G is a centred Gaussian process on the interval $[0,1]$, governed by the asymptotic covariance:

$$C_G(\tau,\tau') := \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \{d_t(\tau)-\mu(\tau)\}\sum_{t=1}^T \{d_t(\tau')-\mu(\tau')\}\right].$$

We propose to use a parametric bootstrap to obtain the quantiles of the limiting distribution. More precisely, let $\hat{C}_G$ be an estimator of the long-run variance function and $\hat{G}$ a centred Gaussian process with covariance function $\hat{C}_G$. Then, we define the bootstrap version:

$$\sup_{\tau\in\mathcal{T}}\left\{a_T\cdot\overline{d}(\tau)+\hat{G}(\tau)\right\} - \sup_{\tau\in\mathcal{T}}\left\{a_T\cdot\overline{d}(\tau)\right\}. \tag{2.6}$$

Here, a_T is a sequence satisfying $a_T\to\infty$ and $a_T/\sqrt{T}\to 0$. Using many replications of the bootstrap statistic, we obtain an approximate $(1-\alpha)$ quantile $q_{1-\alpha}^U$ and decide to reject whenever $S_T^U > q_{1-\alpha}^U$. One advantage of the above bootstrap is that it avoids the explicit estimation of the maximising set argmax(μ). An intuition for boostrap validity is given in Remark 2.2 below.

Remark 2.1. For the bootstrap to work, we require, of course, consistent estimation of the long-run covariance. A concrete HAC estimator $\hat{C}_G$ is detailed in Algorithm 2 of the Supporting Information, and for theoretical analyses of related estimators we refer to Horváth et al. (2013). The HAC estimator then depends on the choice of a bandwidth parameter, essentially the number of lags. Too small a bandwidth may underestimate serial dependence and can therefore lead to more liberal tests, while too large a bandwidth may increase estimation variability. In our simulations, bandwidths have been chosen as $T^{1/2}$ (a choice consistent with the theory in Horváth et al. 2013). The second parameter that is important for practical testing is a_T: theoretically, it has to diverge sufficiently slowly, in the sense that $a_T\to\infty$ and $a_T/\sqrt{T}\to 0$. In practice, we have noticed that slower growth leads to better nominal approximations, and in our simulations $a_T = \log(\log(T))$ yielded the best results. The reason seems to be that for larger values of a_T, the bootstrap statistic tends to (implicitly) underestimate the size of argmax(μ) leading to slightly too small quantiles.

Remark 2.2. The intuition behind the weak convergence (2.5) and the corresponding bootstrap is as follows: According to the central limit theorem in the function space, it holds that $\overline{d}(\tau)\approx \mu(\tau)+G(\tau)/\sqrt{T}$. Since the random component $G(\tau)/\sqrt{T}$ tends to 0, the largest values of $\overline{d}(\tau)$ will asymptotically be attained at those arguments τ where $\mu(\tau)$ is maximised. This implies

$$\sup_{\tau\in\mathcal{T}}\overline{d}(\tau) \approx \sup_{\tau\in\text{argmax}(\mu)}\{\mu(\tau)+G(\tau)/\sqrt{T}\} = \sup_{\tau\in\mathcal{T}}\{\mu(\tau)\} + \sup_{\tau\in\text{argmax}(\mu)}\{G(\tau)/\sqrt{T}\}.$$

Then scaling the above by $\sqrt{T}$ explains the convergence (2.5). For the bootstrap, recall that, again by the CLT, $\overline{d}(\tau)=\mu(\tau)+\mathcal{O}_P\left(T^{-1/2}\right)$. This implies

$$\sup_{\tau\in\mathcal{T}}\left\{\overline{d}(\tau)+\hat{G}(\tau)/a_T\right\} = \sup_{\tau\in\mathcal{T}}\left\{\mu(\tau)+\hat{G}(\tau)/a_T+\mathcal{O}_P\left(T^{-1/2}\right)\right\}. \tag{2.7}$$

The $\mathcal{O}_P\left(T^{-1/2}\right)$ term is negligible compared to the noise term $\hat{G}(\tau)/a_T$ because $a_T \ll T^{1/2}$, and thus we obtain that the right-hand side of (2.7) is approximately

$$\sup_{\tau\in\mathcal{T}}\left\{\mu(\tau)+\hat{G}(\tau)/a_T\right\} \approx \sup_{\tau\in\mathcal{T}}\{\mu(\tau)\} + \sup_{\tau\in\text{argmax}(\mu)}\left\{\hat{G}(\tau)/a_T\right\}. \tag{2.8}$$

Since, for the bootstrap Gaussian process, $\hat{G}(\tau)\overset{d}{\approx} G(\tau)$, the approximation (2.8), after scaling by a_T, explains the validity of the bootstrap approach in (2.6).

3 | Theoretical Analysis

We derive large sample justifications for the statistical tests presented in the previous section. It shows that the critical values must be as specified in Section 2. To formulate our results, we first introduce relevant notation and concepts.

3.1 | Mathematical Preliminaries

3.1.1 | Function Spaces

Let $I\subset\mathbb{R}^d$ be a compact, non-empty Borel set. In our analysis, $d=1,2$ and $I=\mathcal{T}$ (a compact interval) or $I=[0,1]\times\mathcal{T}$, respectively. We define the space

$$C(I,\mathbb{R}) := \{f : I\to\mathbb{R} \mid f \text{ continuous}\},$$

which, when equipped with the supremum norm $\|f\| := \sup_{u\in I}|f(u)|$, is a separable Banach space. To measure the smoothness of a continuous function, for $\xi\in[0,1]$, we define the ξ-Hölder norm

$$\|f\|_\xi = \|f\| + \kappa_\xi(f), \tag{3.1}$$

14679892, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jtsa.70087, Wiley Online Library on [05/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

where κ_ξ is a Hölder modulus of continuity, defined as

$$\kappa_\xi(f) := \sup_{u,v\in I} \frac{|f(u) - f(v)|}{|u - v|^\xi}.$$

We subsequently require Hölder continuity of the functional loss differentials, d_t, expressed through moment assumptions on their Hölder norms, $\|d_t\|_\xi$. This condition ensures that the d_t are, with high probability, concentrated within a set of sufficiently smooth functions. In turn, such concentration is crucial when establishing the central limit theorem in a Banach space, particularly for proving tightness.

In the above supremum for κ_ξ we interpret the ratio as equal to 0 when $u = v$. Obviously, for $\chi > \xi$ and $C_I = diam(I)$ (diameter of the set I), the following relations hold for any continuous function f:

$$\|f\| \le \|f\|_\xi \le \big(C_I + 1\big)\|f\|_\chi.$$

In this sense, Hölder norms are nested and become more restrictive as the Hölder index becomes larger. Finally, we define the Hölder-space $\mathcal{C}_\xi(I, \mathbb{R})$ as the collection of all functions f with $\|f\|_\xi < \infty$ that also satisfy the additional smoothness condition

$$\lim_{\delta \downarrow 0} \sup_{|u-v|\le\delta} \frac{|f(u) - f(v)|}{|u - v|^\xi} = 0.$$

Equipped with the norm $\|\cdot\|_\xi$, the space $\mathcal{C}_\xi(I, \mathbb{R})$ is again a separable Banach space (Račkauskas and Suquet 2009).

3.1.2 | Random Functions

A random function X is a random element with values in a function space. In this work, we consider exclusively random functions with values in $\mathcal{C}(I, \mathbb{R})$ and its subspaces $\mathcal{C}_\xi(I, \mathbb{R})$. If $\mathbb{E}\|X\| < \infty$, the expected value of X exists (in a Bochner sense) $\mathbb{E}X \in \mathcal{C}(I, \mathbb{R})$, and can be defined by the pointwise identity

$$\mathbb{E}[X](u) = \mathbb{E}[X(u)], \qquad u \in I.$$

If $\mathbb{E}\|X\|^2 < \infty$ we can define the covariance kernel c as the continuous function

$$c(u, v) := \mathbb{E}[\{X(u) - \mathbb{E}X(u)\}\{X(v) - \mathbb{E}X(v)\}], \qquad u, v \in I.$$

For further details on random continuous functions, we refer to Janson and Kaijser (2015).

3.2 | A Weak Invariance Principle for Random Functions

For our theoretical analysis, we impose assumptions on the dependence structure of the observations. Specifically, we use the concept of L^p-m-approximability for functional time series. Studied by Hörmann and Kokoszka (2010), L^p-m-approximability was initially formulated for functional data in the Hilbert space of square-integrable functions. This concept encompasses many standard time series models, including functional autoregressive and functional GARCH processes. Under L^p-m-approximability, weak invariance principles in the space of square-integrable functions have been established by Berkes et al. (2013). However, to the best of our knowledge, there exist no related results for other Banach spaces. Thus, our analysis of L^p-m-approximable time series of continuous functions is of independent interest.

The motivation for modelling observations on the space of continuous rather than square integrable functions comes from our statistical methods. All test statistics we consider can be expressed as transforms of a partial sum process of random functions P_T (defined in (3.3)). These transforms are continuous with respect to the stronger supremum norm, but generally fail to be continuous with respect to the topology of square integrable functions.

Definition 3.1. (L^p-m-approximable sequences) Let $(\eta_t)_{t\in\mathbb{Z}}$ be a sequence of i.i.d. random variables taking values in a measurable space S, $g : S^\infty \to \mathbb{B}$ be a measurable function and $(d_t)_{t\in\mathbb{Z}}$ a sequence of random variables in a separable Banach space $(\mathbb{B}, \|\cdot\|_\mathbb{B})$ admitting the representation:

$$d_t = g(\eta_t, \eta_{t-1}, \dots), \quad t \in \mathbb{Z}.$$

Moreover, assume that $(\eta'_t)_{t\in\mathbb{Z}}$ is an independent copy of $(\eta_t)_{t\in\mathbb{Z}}$ defined on the same probability space, and set

$$d_t^{(m)} = g(\eta_t, \eta_{t-1}, \dots, \eta_{t-m+1}, \eta'_{t-m}, \eta'_{t-m-1}, \dots), \quad m \in \mathbb{N}.$$

Further, for some $p \ge 1$, suppose that $\mathbb{E}\|d_1\|_\mathbb{B}^p < \infty$ and that for a function $\rho : \mathbb{N} \to [0, \infty)$,

$$\Big[\mathbb{E}\|d_1 - d_1^{(m)}\|_\mathbb{B}^p\Big]^{1/p} \le \rho(m). \tag{3.2}$$

Then, if $\rho(m) \to 0$ as $m \to \infty$, we call the sequence $(d_t)_{t\in\mathbb{Z}}$ L^p-m-approximable with rate function ρ.

Notice that L^p-m-approximability entails strict stationarity of the sequence $(d_t)_{t\in\mathbb{Z}}$. We can now formulate the central assumption for our analysis. Recall that in this work $\mathcal{T}$ is a compact, non-degenerate interval.

Assumption 3.1. The time series $(d_t)_{t\in\mathbb{Z}}$ is L^p-m-approximable in $(\mathcal{C}_\xi(\mathcal{T}, \mathbb{R}), \|\cdot\|_\xi)$ with rate function ρ. The parameter $\xi \in (0, 1/2]$ satisfies the relation $p\xi > 2$. The rate function satisfies, for some constants $C_\rho, z > 0$, $\rho(m) \le C_\rho m^{-(1+z)} \quad \forall m \in \mathbb{N}$.

Assumption 3.1 requires L^p-m-approximability with respect to a Hölder norm. The Hölder norm is a composite object, involving both a supremum norm and a modulus of continuity (see (3.1)). To make it easier to mathematically verify Assumption 3.1, we provide with the following Lemma 3.1 a sufficient criterion for L^p-m-approximability in the Hölder norm. Mathematically, the lemma can be used to check Assumption 3.1 for standard time series models, and we have worked out a proof for functional autoregressive processes in the Supporting Information. In the context of functional data on Hilbert spaces, L^p-m-approximability has also been shown for wider classes of relevant processes, such as functional GARCH, and we refer to Hörmann and Kokoszka (2010) for further examples. Intuitively, Lemma 3.1 states that the time series $(d_t)_{t\in\mathbb{Z}}$ has to be weakly

14679892, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jtsa.70087, Wiley Online Library on [05/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

dependent and that individual functions d_t need to be sufficiently smooth. In practice, diagnostic plots can help to determine weak dependence of $(d_t)_{t\in\mathbb{Z}}$. A standard approach is plotting the autocorrelation function for a small number of leading fPC scores, where fast decay indicates weak dependence. A related measure is the decay of the norm of lagged empirical autocovariance operators. Smoothness typically follows from arguments inherent in the application (the functional variable is known to evolve continuously), from preprocessing (using smoothing techniques when reconstructing d_t from discrete data), or both.

Lemma 3.1. *Suppose the time series $(d_t)_{t\in\mathbb{Z}}$ is L^p-m-approximable in $(C(\mathcal{T},\mathbb{R}),\|\cdot\|)$ with rate function $\tilde{\rho}$ that satisfies for some constants $C_{\tilde{\rho}}>0$ and $\tilde{z}>1$ the inequality $\tilde{\rho}(m)\le C_{\tilde{\rho}}m^{-(1+\tilde{z})}$. Furthermore, suppose that $\mathbb{E}\|d_1\|_\chi^p<\infty$ for some $\chi>\tilde{z}^{-1}$. Then, it follows for any $0<\xi<\frac{\tilde{z}\chi}{\tilde{z}+1}$ that $(d_t)_{t\in\mathbb{Z}}$ is L^p-m-approximable in $\left(C_\xi(\mathcal{T},\mathbb{R}),\|\cdot\|_\xi\right)$ with rate function ρ that satisfies, with some constants $C_\rho, z>0$, $\rho(m)\le C_\rho m^{-(1+z)}$.*

We notice that many processes considered in the literature, such as autoregressive processes, have an exponentially decaying rate function $\tilde{\rho}$ and thus $\tilde{z}$ in Lemma 3.1 can be chosen arbitrarily large. In this case any choice $\xi<\chi$ is permissible.

We now turn to the main probabilistic result of this work, the weak convergence of the partial sum process

$$P_T(x,\tau):=\frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor Tx\rfloor}\left[d_t(\tau)-\mu(\tau)\right]+[Tx-\lfloor Tx\rfloor]\frac{\left[d_{\lfloor Tx\rfloor+1}(\tau)-\mu(\tau)\right]}{\sqrt{T}},\quad 0\le x\le 1. \tag{3.3}$$

Notice that by construction P_T is a continuous function in both arguments $\tau\in\mathcal{T}$ and $x\in[0,1]$. The second term on the right-hand side of (3.3) ensures continuity of P_T in x. Convergence of the process $\{P_T(1,\tau):\tau\in\mathcal{T}\}$ is the central limit theorem in a Banach space of Hölder functions. Convergence of the entire process $\{P_T(x,\tau):x\in[0,1],\tau\in\mathcal{T}\}$ is the corresponding weak invariance principle (with respect to a Hölder norm in time and functional component). Classical convergence results on Banach spaces typically impose some type of mixing condition on the data (e.g., Dehling (1983), Samur (1987)). However, even for one dimensional data, many standard processes fail to be mixing. For instance, Andrews (1984) constructs a simple example of a discrete autoregressive process, which fails to be mixing. In contrast, the concept of L^p-m-approximability used in this paper covers most commonly used scalar and functional processes. Incidentally, the example constructed by Andrews (1984) even has an exponentially decaying rate function.

Theorem 3.1. *Suppose that Assumption 3.1 holds. Then, there exists a Gaussian process $\{G(x,\tau):x\in[0,1],\tau\in\mathcal{T}\}$ with continuous sample paths, such that the weak convergence*

$$\{P_T(x,\tau):x\in[0,1],\tau\in\mathcal{T}\}\xrightarrow{d}\{G(x,\tau):x\in[0,1],\tau\in\mathcal{T}\} \tag{3.4}$$

holds in $C_{\xi'}([0,1]\times\mathcal{T},\mathbb{R})$, for any

$$\xi'\in\left(0,\xi\left[\frac{p\xi-2}{p\xi+2}\right]\right). \tag{3.5}$$

The Gaussian process G is centred and characterised by the long-run covariance kernel

$$c(x_1,x_2,\tau_1,\tau_2):=\min(x_1,x_2)\sum_{z\in\mathbb{Z}}\mathbb{E}\left[\{d_0(\tau_1)-\mu(\tau_1)\}\{d_z(\tau_2)-\mu(\tau_2)\}\right].$$

Remark 3.1.

1. The interval in (3.5) is non-empty because by Assumption 3.1 we have $p\xi>2$. For our subsequent statistical theory, convergence on $C_{\xi'}([0,1]\times\mathcal{T},\mathbb{R})$ for any value of $\xi'>0$ is sufficient. However, there exist important applications in statistics, where the size of ξ' is of interest and that can be based on our theorem. Examples include multi-scale change point tests for epidemic changes (Račkauskas and Suquet (2006)), multiple changes (Kutta et al. (2025)) and sequential monitoring (Kutta and Dörnemann (2025)).
2. Theorem 3.1 implies in particular, that the random function $\{G(x,\tau):x\in[0,1],\tau\in\mathcal{T}\}$ is almost surely an element of $C_{\xi'}([0,1]\times\mathcal{T},\mathbb{R})$.
3. The process G in Theorem 3.1 is a Banach space valued Brownian motion, which follows from its covariance structure. For a general introduction to such processes, we refer the reader to the work of Kuelbs (1973). By the continuous mapping theorem, it follows for any continuous, linear functional $\Lambda_W:C(\mathcal{T},\mathbb{R})\to\mathbb{R}$ that where is a standard Brownian motion and the long run variance.

$$\left\{\Lambda_W[G(x,\cdot)]:x\in[0,1]\right\}\overset{d}{=}\{\sigma B(x):x\in[0,1]\},\qquad \sigma^2:=\sum_{z\in\mathbb{Z}}\mathbb{E}\left[\Lambda_W\left[d_0-\mu\right]\cdot\Lambda_W\left[d_z-\mu\right]\right]. \tag{3.6}$$

3.3 | Main Statistical Results

We apply Theorem 3.1 to study the weak convergence of the test statistics S_T^W, S_T^{SN}, S_T^U introduced in Section 2. We recall that S_T^M is a special case of S_T^W and will therefore not receive a separate treatment in our theory. Both statistics S_T^W, S_T^{SN} test the null hypothesis $H_0(W)$ (Equation (2.4)) and S_T^U the null hypothesis $H_0(U)$ (Equation (2.2)). For the formulation of our results, we distinguish three cases: The interior of the null hypothesis, the boundary of the null hypothesis, and the alternative.

	Interior of H_0	**Boundary of H_0**	**Alternative H_1**
$H_0(W)$	$\Lambda_W(\mu)<0$	$\Lambda_W(\mu)=0$	$\Lambda_W(\mu)>0$
$H_0(U)$	$\sup_{\tau\in\mathcal{T}}\mu(\tau)<0$	$\sup_{\tau\in\mathcal{T}}\mu(\tau)=0$	$\sup_{\tau\in\mathcal{T}}\mu(\tau)>0$

For each of the test statistics S_T^W, S_T^{SN}, S_T^U we can derive weak convergence on the boundary of the null hypothesis and convergence to a degenerate distribution else. To state these results, we define the limiting distributions as follows:

$$S^W\overset{d}{=}\sigma B(1), \tag{3.7}$$

14679892, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jtsa.70087, Wiley Online Library on [05/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$$S^{SN} \overset{d}{=} \frac{B(1)}{\left\{\int_0^1 x^2(B(x)-xB(1))^2dx\right\}^{1/2}}, \tag{3.8}$$

$$S^{U} \overset{d}{=} \sup_{\tau:\mu(\tau)=0} G(1,\tau), \tag{3.9}$$

where B is a standard Brownian motion, σ^2 is defined in (3.6) and G is the Gaussian process defined in Theorem 3.1. For the next theorem, we assume tacitly that the covariance kernel c is non-degenerate and that the variance σ^2 is positive.

Theorem 3.2. *Suppose that Assumption 3.1 holds. Then, for any* $\boldsymbol{x} \in \{W, SN, U\}$ *it follows that*

$$S_T^{\boldsymbol{x}} \overset{d}{\to} \begin{cases} -\infty, & \text{interior of } H_0 \\ S^{\boldsymbol{x}}, & \text{boundary of } H_0 \\ \infty, & \text{alternative } H_1. \end{cases}$$

In other words, the theorem states that if $q_{1-\alpha}^{\boldsymbol{x}}$ is the upper $1-\alpha$ quantile of a non-degenerate limiting distribution $S_T^{\boldsymbol{x}}$ for some $\alpha \in (0,1)$, it follows that

$$\mathbb{P}\left(S_T^{\boldsymbol{x}} > q_{1-\alpha}^{\boldsymbol{x}}\right) \to \begin{cases} 0, & \text{interior of } H_0 \\ \alpha, & \text{boundary of } H_0 \\ 1, & \text{alternative } H_1. \end{cases}$$

Approximating the limiting quantile is straightforward in the case of the statistic S_T^{SN}, where the limit is pivotal. Approximating the limit S^W requires the estimation of the long-run variance of the transformed differences $\Lambda_W\left[d_t\right]$, and we provide some details in our finite sample analysis.

For the case of the uniform hypothesis, we have proposed a parametric bootstrap in Section 2 (see (2.6)). There, an estimate of the long-run variance operator is used to create a centred Gaussian process that converges weakly to the limit $G(1,\tau)$ (as the estimated covariance convergence to the true covariance). The following lemma provides general large sample justification for this approach. We denote the Gaussian processes in this lemma by $G_T^{|}, G^{|}$ rather than G_T, G, because they do not live on $C([0,1]\times\mathcal{T},\mathbb{R})$ but on the space $C(\mathcal{T},\mathbb{R})$.

Lemma 3.2. *Consider a sequence of Gaussian processes* $\left(G_T^{|}\right)_T$ *that converge to a non-degenerate limit Gaussian process* $G^{|}$ *in* $C(\mathcal{T},\mathbb{R})$. *Furthermore let* $\left(a_T\right)_T$ *be a sequence of positive numbers such that* $a_T \to \infty$ *but* $a_T/\sqrt{T} \to 0$. *Then,*

$$\sup_{\tau\in\mathcal{T}}\left\{a_T\cdot\overline{d}(\tau)+G_T^{|}(\tau)\right\}-\sup_{\tau\in\mathcal{T}}\left\{a_T\cdot\overline{d}(\tau)\right\}\overset{d}{\to}\sup_{\tau\in\mathcal{T}_{max}}G^{|}(\tau).$$

The set $\mathcal{T}_{max}$ *on the right-hand side is defined as*

$$\mathcal{T}_{max} := \{\tau\in\mathcal{T} : \mu(\tau)=\sup_{s\in\mathcal{T}}\mu(s)\}.$$

The parametric bootstrap proposed in this section relies on a functional CLT (implied by Theorem 3.1) together with a continuous mapping argument for the difference in (2.6). Of course, alternative bootstrap methods may also be used to construct a process $\hat{G}\approx G$ that serves the same purpose. One possible alternative is a dependent multiplier bootstrap, in the spirit of Dette et al. (2024).

4 | Monte Carlo Simulations

In this section, we evaluate the finite-sample performance of the proposed tests through Monte Carlo simulations. We begin by examining the performance of the LRV-based and self-normalised (SN) statistics for testing the mean and weighted hypotheses, and then investigate the performance of the parametric bootstrap procedure for testing the uniform hypothesis.

4.1 | Finite Sample Performance: Mean and Weighted Hypotheses

Within model (2.1), we consider the mean function $\mu(\tau)$ that covers three cases: the boundary of H_0, the interior of H_0, and the alternative H_1. We provide some clarification on the interpretation of our empirical rejection rates in relation to conventional hypothesis testing, in particular for size and power. Specifically, rejection rates under the boundary case correspond to empirical sizes and are expected to be close to the nominal significance levels; rejection rates under the interior case are expected to be less than the nominal levels; rejection rates under the alternative represent empirical power, where higher values are preferable.

To investigate finite-sample performance under these three cases, we consider five different mean functions $\mu(\tau)$:

- Flat at zero: $\mu(\tau)=0$, representing the boundary case in the domain $[0,1]$.
- Flat at a negative constant: $\mu(\tau)=-c_1$, representing the interior case in the domain $[0,1]$.
- Flat at a positive constant: $\mu(\tau)=c_2$, representing the alternative in the domain $[0,1]$.
- Slope function: $\mu(\tau)=-c_3+2c_3\tau$, representing the interior case in the domain $[0,0.5]$ and the alternative in $[0.5,1]$.
- Sine function: $\mu(\tau)=c_4\sin(2\pi\tau)$, resulting in the alternative in the domain $[0,0.5]$ and the interior case in $[0.5,1]$.

Figure S1 illustrates the shapes of these mean functions. The values of constants c_1, c_2, c_3, c_4 are calibrated to ensure that the mean functions are in a similar scale as the average loss differentials observed in our empirical application in Section 5.

Next, we consider three specifications of the error term $e_t(\tau)$, which is a mean-zero functional time series. The first specification is a sequence of independent and identically distributed (IID) random functions:

$$e_t(\tau)=\sum_{m=1}^{M}c_{t,m}\phi_m(\tau),\quad \tau\in[0,1],\ t=1,\ldots,T, \tag{4.1}$$

where $c_{t,m}\sim i.i.d.\mathcal{N}\left(0,0.1^2\right)$, and $\phi_m(\tau), m=1,\ldots,M$, are the B-spline basis functions. We use $M=13$ basis functions,

14679892, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jtsa.70087, Wiley Online Library on [05/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

corresponding to order-four B-splines with nine equally spaced interior knots over the interval $[0, 1]$.

The second specification is a functional autoregressive (FAR) process:

$$e_t(\tau) = \int \psi(\tau, s) e_{t-1}(s) ds + \varepsilon_t(\tau), \quad \tau \in [0, 1], \ t = 1, \ldots, T,$$

with the error functions ε_t equal to the e_t in Equation (4.1). The kernel is specified as $\psi(\tau, s) = 2.25 s\tau$, yielding $\|\psi\| = \left\{\int\int \psi^2(\tau, s) d\tau ds\right\}^{1/2} = 0.75$. This specification induces weak dependence in the functional time series process.

The third specification is a functional GARCH (FGARCH) process:

$$e_t(\tau) = \sigma_t(\tau)\varepsilon_t(\tau), \tau \in [0, 1], t = 1, \ldots, T,$$
$$\sigma_t^2(\tau) = \omega_t(\tau) + \int \alpha(\tau, s) e_{t-1}^2(s) ds + \int \beta(\tau, s)\sigma_{t-1}^2(s) ds,$$

where the innovation $\varepsilon_t(\tau)$ follows an Ornstein-Uhlenbeck process, the constant coefficient is $\omega_t(\tau) = 0.1\tau(1 - \tau)$, and the functional ARCH and GARCH coefficients are $\alpha(\tau, s) = \beta(\tau, s) = 12\tau(1 - \tau)s(1 - s)$. This specification captures conditional heteroskedasticity in the functional time series.

To implement the test procedure, an additional key input is the choice of the weight function $w(\tau)$ used in the test statistic. To examine the sensitivity of the weighted procedure to the specification of $w(\tau)$, we consider three economically distinct weighting schemes:

Equal-weight: $w(\tau) = 1$.

$$\text{First-half: } w(\tau) = \begin{cases} 2, & \tau \in [0, 1/2], \\ 0, & \tau \in [1/2, 1]. \end{cases}$$

$$\text{Second-half: } w(\tau) = \begin{cases} 0, & \tau \in [0, 1/2], \\ 2, & \tau \in [1/2, 1]. \end{cases}$$

The equal-weight specification corresponds to the mean hypothesis, while the "first-half" and "second-half" specifications represent situations where the researcher places all emphasis on different regions of the functional domain. These deliberately contrasting choices provide a simple assessment of how the finite-sample performance of the procedure varies with the selected weighting function.

We summarise our step-by-step test procedure in Algorithm 1 provided in the Supporting Information. Based on 5000 Monte Carlo replications, we generate data from our DGPs and apply Algorithm 1 to obtain the empirical rejection rates under various cases. Table 1 reports the empirical rejection rates under the boundary case, which correspond to the empirical sizes. As expected, the rejection rates converge to their nominal significance levels as the sample size increases. Both the LRV-based and SN tests exhibit good size control across the three mean functions when the errors are IID, at the 1%, 5%, and 10% significance levels. When errors follow FAR and FGARCH processes, the SN test achieves better size control than the LRV-based test, particularly in small samples. Across the three shapes of the mean function, the size performance is broadly similar. Moreover, testing the mean hypothesis in the domain $[0, 1]$ using the equal-weight function yields results comparable to those obtained when testing the weighted hypotheses with the first-half weight function for the domain $[0, 0.5]$ and with second-half weight function for the $[0.5, 1]$.

Table 2 shows the empirical rejection rates at the 5% significance level under the interior case, where the rejection rates are theoretically expected to be below the nominal significance level. This pattern is consistently observed for both the LRV-based and SN tests across all settings, including different mean functions, error processes, and weight functions. In line with our theory in Section 3.3, the rejection rates under the interior case converge to zero as the sample size increases. Notably, the convergence is fastest under IID errors and slowest under FGARCH errors, with the FAR process exhibiting an intermediate rate.

Table 3 presents the empirical rejection rates at the 5% significance level under the alternative, which represents the empirical power of the tests. As suggested by our theory, the rejection rate is expected to increase to one with the sample size, and this is observed universally for both the LRV-based and SN tests with all settings. Again, the convergence is fastest under IID errors and slowest under FGARCH errors. In particular, under IID errors the power approaches 100% even for relatively small samples (e.g., $T = 100$), whereas the power under FGARCH requires a larger sample size (such as $T = 500$) and a larger magnitude of the deviation from the null in order to achieve power above 90%. Across most settings, the LRV-based test exhibits slightly higher power than the SN test. Overall, the finite-sample behaviour of the proposed procedure is also stable across the alternative weighting schemes, suggesting that the test is not unduly sensitive to reasonable changes in the specification of $w(\tau)$.

4.2 | Finite Sample Performance: Uniform Hypothesis

We now turn to the evaluation of the parametric bootstrap procedure for testing the uniform hypothesis $H_0(U)$. To facilitate its implementation by empirical researchers, we summarise the bootstrap procedure in Algorithm 2 which can be found in the Supporting Information. For the kernel choice in Step 4, we follow Horváth et al. (2014) and use a flat top kernel:[1]

$$K(x) = \begin{cases} 1, & 0 \le x < 0.1 \\ 1.1 - |x|, & 0.1 \le x < 1.1 \\ 0, & |x| > 1.1, \end{cases}$$

with bandwidth $h = T^{1/2}$. For the choice of a_T in Step 5, we set $a_T = \log(\log(T))$.

We focus on the three shapes of the mean function. Under the boundary case, we use $\mu(\tau) = 0$ (flat at zero) and under the interior case, we employ $\mu(\tau) = -c_1$ (flat at a negative constant), as in Section 4.1. For the alternative, we construct a special mean function,

$$\mu(\tau) = -| - 0.2 + 0.4\tau| + 0.05,$$

14679892, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jtsa.70087, Wiley Online Library on [05/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 | Empirical rejection rates under the boundary case.

	Weight: Equal $\mu(\tau)=0$			Weight: First half $\mu(\tau)=0$			Weight: Second half $\mu(\tau)=0$			Weight: Equal $\mu(\tau)=-0.2+0.4\tau$			Weight: Equal $\mu(\tau)=0.2\sin(2\pi\tau)$		
	10%	**5%**	**1%**	**10%**	**5%**	**1%**	**10%**	**5%**	**1%**	**10%**	**5%**	**1%**	**10%**	**5%**	**1%**
LRV-based															
IID															
$T=100$	10.9%	5.8%	2.0%	11.9%	6.5%	1.9%	12.0%	7.2%	1.9%	11.0%	6.3%	1.9%	11.5%	6.5%	1.7%
$T=200$	10.7%	5.5%	1.4%	10.1%	5.6%	1.4%	10.8%	5.5%	1.2%	11.6%	6.2%	1.5%	10.6%	5.6%	1.5%
$T=500$	10.7%	5.5%	1.0%	10.9%	5.6%	1.3%	10.0%	5.2%	1.3%	10.9%	5.8%	0.9%	10.8%	5.9%	1.3%
FAR															
$T=100$	16.7%	10.7%	4.2%	17.0%	11.2%	4.4%	14.5%	9.8%	3.8%	15.4%	9.5%	3.5%	16.1%	10.6%	4.5%
$T=200$	15.4%	9.8%	3.3%	16.1%	10.0%	3.4%	12.8%	7.2%	2.8%	14.6%	8.8%	3.3%	15.1%	9.6%	3.5%
$T=500$	12.9%	7.7%	2.3%	13.7%	8.2%	2.9%	12.8%	7.3%	2.0%	13.7%	7.9%	2.7%	13.2%	7.4%	2.5%
FGARCH															
$T=100$	12.4%	6.8%	1.8%	12.3%	6.9%	1.8%	12.4%	6.6%	2.1%	11.3%	6.6%	1.8%	12.9%	7.4%	2.0%
$T=200$	10.6%	5.5%	1.0%	10.6%	5.8%	1.0%	11.9%	6.0%	1.3%	10.6%	5.4%	1.2%	11.8%	6.3%	1.4%
$T=500$	11.4%	5.5%	1.2%	11.2%	6.0%	1.3%	10.4%	5.4%	1.3%	10.4%	5.2%	1.1%	10.1%	5.3%	1.1%
Self-normalised															
IID															
$T=100$	9.6%	4.6%	1.0%	9.5%	5.0%	1.1%	10.5%	5.4%	1.1%	9.9%	4.8%	1.0%	9.5%	4.4%	0.7%
$T=200$	9.5%	4.7%	1.1%	9.4%	4.6%	1.0%	9.4%	4.7%	0.9%	8.8%	4.9%	1.5%	9.9%	4.8%	1.1%
$T=500$	10.0%	4.7%	0.8%	10.1%	4.8%	0.7%	9.4%	4.7%	0.8%	9.9%	5.1%	0.8%	10.4%	5.4%	1.1%
FAR															
$T=100$	13.1%	7.8%	2.5%	12.0%	6.9%	1.9%	12.0%	6.9%	2.1%	12.1%	7.1%	1.9%	12.6%	7.6%	1.9%
$T=200$	11.7%	6.1%	1.5%	10.2%	5.4%	1.0%	11.6%	5.8%	1.5%	11.0%	5.8%	1.4%	10.9%	5.6%	1.3%
$T=500$	10.1%	5.7%	1.2%	10.9%	5.9%	1.2%	10.2%	5.3%	1.0%	10.7%	5.8%	1.4%	11.5%	6.1%	1.4%
FGARCH															
$T=100$	9.5%	4.8%	0.9%	10.2%	5.1%	1.0%	10.3%	4.8%	1.1%	9.6%	4.4%	0.8%	9.7%	4.9%	0.9%
$T=200$	9.7%	4.1%	0.7%	9.8%	4.6%	0.9%	10.5%	5.1%	1.0%	10.9%	5.2%	0.8%	10.4%	4.9%	1.0%
$T=500$	10.8%	5.4%	1.0%	10.6%	4.6%	0.8%	10.2%	5.4%	1.2%	9.5%	4.6%	0.8%	10.1%	4.8%	0.9%

Note: The empirical rejection rates under the boundary case represent the empirical sizes, which are expected to be close to the nominal levels.

which we refer to as the "iceberg tip," reflecting that only a small proportion of the function lies above zero. All other DGP settings are the same as those in Section 4.1.

To compute the empirical rejection rates, we generate the data and run Algorithm 2 for 5000 replications. We use larger sample sizes to illustrate convergence. Table 4 presents the rejection rates at the 5% significance level for the boundary, interior, and alternative cases. Under the boundary case, the size is slightly above the nominal size which is not unusual in bootstrap tests involving time series dependence. In terms of the interior case, rejection rates are below the nominal level and converge towards zero as the sample size increases. Under the alternative, the empirical power converges to one with increasing sample sizes and as the deviation from the null increases in the DGP. As observed previously, the convergence rate is fastest with IID errors and slowest with FGARCH errors.

4.3 | Additional Simulations

We conclude this section by summarising three additional simulations, with full details and results reported in Section S5.

First, the bootstrap procedure for the uniform test requires only that $a_T \to \infty$ and $a_T/\sqrt{T} \to 0$, and there is a wide range of choices of a_T. In Section S5, we repeat the simulation of Section 4.2 for two other choices, $a_T = \log(T)$ and $a_T = T^{1/4}$. Rejection rates in the interior of the null and under the alternative are at similar levels across the three choices of a_T, while

14679892, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jtsa.70087, Wiley Online Library on [05/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 | Empirical rejection rates under the interior case at the 5% significance level.

	Weight: Equal $\mu(\tau) = -c_1$			Weight: First half $\mu(\tau) = -c_2 + 2c_2\tau$			Weight: Second half $\mu(\tau) = c_3 \sin(2\pi\tau)$		
	$c_1 = 0.01$	$c_1 = 0.02$	$c_1 = 0.05$	$c_2 = 0.05$	$c_2 = 0.10$	$c_2 = 0.20$	$c_3 = 0.05$	$c_3 = 0.10$	$c_3 = 0.20$
LRV-based									
IID									
$T = 100$	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
$T = 200$	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
$T = 500$	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
FAR									
$T = 100$	1.3%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
$T = 200$	0.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
$T = 500$	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
FGARCH									
$T = 100$	3.8%	2.6%	1.0%	2.2%	0.6%	0.1%	1.6%	0.4%	0.0%
$T = 200$	3.1%	1.8%	0.3%	1.4%	0.1%	0.0%	1.0%	0.0%	0.0%
$T = 500$	2.0%	0.6%	0.0%	0.5%	0.0%	0.0%	0.2%	0.0%	0.0%
Self-normalised									
IID									
$T = 100$	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
$T = 200$	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
$T = 500$	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
FAR									
$T = 100$	1.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
$T = 200$	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
$T = 500$	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
FGARCH									
$T = 100$	3.5%	2.3%	0.5%	2.2%	0.7%	0.0%	1.4%	0.4%	0.0%
$T = 200$	3.0%	2.1%	0.1%	1.5%	0.3%	0.0%	0.9%	0.0%	0.0%
$T = 500$	2.5%	0.9%	0.0%	0.5%	0.1%	0.0%	0.3%	0.0%	0.0%

Note: This table shows empirical rejection rates under the interior case, and the rejection rates are expected to be less than to the nominal level.

the empirical size at the boundary deteriorates with $a_T = \log(T)$ and $a_T = T^{1/4}$. Therefore, $a_T = \log(\log(T))$ is best for size control, without noticeable compromise in power.

Second, we treat the loss differentials as densely observed curves in our DGPs. In some applications, however, functional data are only sparsely observed on a discrete grid, and the curves need to be reconstructed before applying our tests (Cai and Hu 2025). In Section S5.2, we repeat the above designs with curves generated on a grid of 21 equally spaced points and reconstructed by B-splines.[2] The reconstruction step has marginal effect on the mean and weighted tests. For the uniform test, it has similar size at the boundary, while there is some minor reduction in power in small samples. These results indicate robustness to the sparse observation and reconstruction of functional data in the setting considered here.

Third, to further highlight the advantage of self-normalised tests, Section S5.3 repeats the simulation for a FAR process with a stronger serial dependence of $\|\psi\| = 0.9$, compared with $\|\psi\| = 0.75$ used before. The size distortion of the LRV-based test at the boundary increases substantially, while that of the self-normalised test remains much closer to the nominal level. The result showcases that the self-normalised test has advantage of good size control, in particular for functional data with strong serial dependence.

5 | Empirical Application

This section is dedicated to an illustration of our new methodology using a real-world dataset where we compare two forecasting methods for the U.S. yield curve. We are interested in whether

14679892, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jtsa.70087, Wiley Online Library on [05/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 3 | Empirical rejection rates under the alternative at the 5% significance level.

	Weight: Equal $\mu(\tau) = c_1$			Weight: Second half $\mu(\tau) = -c_2 + 2c_2\tau$			Weight: First half $\mu(\tau) = c_3 \sin(2\pi\tau)$		
	$c_1 = 0.02$	$c_1 = 0.05$	$c_1 = 0.10$	$c_2 = 0.05$	$c_2 = 0.10$	$c_2 = 0.20$	$c_3 = 0.05$	$c_3 = 0.10$	$c_3 = 0.20$
LRV-based									
IID									
$T = 100$	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
$T = 200$	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
$T = 500$	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
FAR									
$T = 100$	70.9%	99.9%	100.0%	57.5%	95.3%	100.0%	100.0%	100.0%	100.0%
$T = 200$	90.0%	100.0%	100.0%	78.2%	99.9%	100.0%	100.0%	100.0%	100.0%
$T = 500$	99.7%	100.0%	100.0%	98.3%	100.0%	100.0%	100.0%	100.0%	100.0%
FGARCH									
$T = 100$	12.9%	28.3%	62.0%	14.3%	27.2%	59.5%	17.7%	34.6%	73.4%
$T = 200$	14.0%	39.2%	80.3%	18.1%	36.3%	79.4%	19.6%	48.0%	90.9%
$T = 500$	20.5%	62.3%	98.2%	26.6%	62.9%	97.9%	34.3%	77.0%	99.6%
Self-normalised									
IID									
$T = 100$	99.9%	100.0%	100.0%	99.5%	100.0%	100.0%	100.0%	100.0%	100.0%
$T = 200$	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
$T = 500$	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
FAR									
$T = 100$	53.4%	97.7%	100.0%	45.7%	84.9%	99.9%	97.9%	100.0%	100.0%
$T = 200$	71.7%	99.9%	100.0%	62.8%	96.7%	100.0%	100.0%	100.0%	100.0%
$T = 500$	94.5%	100.0%	100.0%	87.5%	99.9%	100.0%	100.0%	100.0%	100.0%
FGARCH									
$T = 100$	9.4%	20.8%	46.3%	10.9%	19.6%	44.8%	12.4%	25.0%	58.9%
$T = 200$	11.2%	28.9%	66.9%	14.4%	28.6%	65.3%	16.2%	39.0%	79.4%
$T = 500$	16.3%	50.1%	89.9%	21.2%	49.2%	89.6%	27.7%	63.1%	96.4%

Note: The rejection rates under the alternative represent the empirical power.

using forward rates can outperform a random walk benchmark method in forecasting the entire yield curve (mean hypothesis), a specifically chosen part of yield curves (weighted hypothesis), and an arbitrary portion of yield curves (uniform hypothesis). Specifically, if the trading day observations of yield curves are $Y_t(\tau)$ then the h-step ahead random walk predictions are given as

$$\widehat{Y}_t^{(1)}(\tau) = Y_{t-h}(\tau) \tag{5.1}$$

whereas with h-day forward rates $f_t^{(h)}(\tau)$ then the forecasts implied by the forward rates are

$$\widehat{Y}_t^{(2)}(\tau) = f_{t-h}^{(h)}(\tau). \tag{5.2}$$

The specific interest rates we use are the U.S. *Overnight Index Swaps* (OIS), which are based on over-the-counter derivatives in which two agents agree to exchange fixed and floating interest payments over a given period. There are two advantages to using OIS rates. First, they are available across a wide range of maturities, which aligns well with the nature of functional data. Second, there are forward contracts (at different contract horizons) of *Interest Rate Swaps* (IRS) that provide forward rates which are commonly used to measure interest rate expectations and serve as the market-based (model-free) forecasts of the OIS rates.

Our data are downloaded from Refinitiv EIKON (previously Datastream). The OIS rates are available for maturities of 3–12 months, 1–10, 12, 15, and 20 year.[3] The forward rates are available at maturities of 3, 6, 9 month, 1, 2, 3, 5, 7, 10, 15, and 20 year. Given that the maturities of the OIS rates and the forward rates do not coincide, modelling and forecasting the yield curve as a functional data object (continuous curve) is the most natural approach. A further motivation to consider the functional data framework is that financial practitioners are also trading fixed-income products that are based on non-standard maturities,

14679892, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jtsa.70087, Wiley Online Library on [05/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 4 | Empirical rejection rates of testing the uniform hypothesis at the 5% significance level.

	Boundary	Interior			Alternative		
		$\mu(\tau) = -c_1$			$\mu(\tau) = -\lvert -0.2 + 0.4\tau \rvert + c_4$		
	$\mu(\tau) = 0$	$c_1 = 0.01$	$c_1 = 0.02$	$c_1 = 0.05$	$c_4 = 0.02$	$c_4 = 0.05$	$c_4 = 0.10$
IID							
$T = 200$	8.5%	0.1%	0.0%	0.0%	98.6%	100.0%	100.0%
$T = 500$	7.9%	0.0%	0.0%	0.0%	100.0%	100.0%	100.0%
$T = 2000$	6.8%	0.0%	0.0%	0.0%	100.0%	100.0%	100.0%
FAR							
$T = 200$	8.3%	1.6%	0.2%	0.0%	71.6%	100.0%	100.0%
$T = 500$	6.9%	0.3%	0.0%	0.0%	96.5%	100.0%	100.0%
$T = 2000$	6.2%	0.0%	0.0%	0.0%	100.0%	100.0%	100.0%
FGARCH							
$T = 200$	7.6%	5.2%	3.3%	0.7%	9.0%	22.1%	54.6%
$T = 500$	6.6%	3.5%	1.4%	0.1%	11.9%	36.3%	83.6%
$T = 2000$	6.5%	1.6%	0.1%	0.0%	24.0%	82.9%	99.9%

Note: This table shows the empirical rejection rates of testing the uniform hypothesis for the boundary, interior and alternative cases. On the boundary, the rejection rates are expected to be close to the nominal levels; under the interior cases, the rejection rates are expected to be less than the nominal level; under the alternative, the rejection rates are expected to approach one.

and thus any arbitrary domain of yield curves could be practically relevant. We observe and forecast yield curves on a daily basis. Due to data availability, the sample period is chosen to be between September 28, 2011 and September 30, 2024. There are different contract horizons of the forward contracts.

5.1 | Main Results

In our main results, we focus on the forecast horizon of 63 trading days (i.e., a 3-month horizon). Consequently, the out-of-sample forecasts can be evaluated from December 26, 2011 to September 30, 2024. Figure 1 shows the raw data of OIS rates (upper panel) and 3-month forward rates (lower panel).

The actual observations of yield curves $Y_t(\tau)$ are available for monthly maturities $\tau \in [3, 240]$, and we denote the forward rates at 3-months (63 days) as $f_t^{(63)}(\tau)$. The benchmark forecasting method (Forecast 1) is a random walk method corresponding to Equation (5.1) which is given by $\hat{Y}_t^{(1)}(\tau) = Y_{t-63}(\tau)$ in this application. The competing forecasting method (Forecast 2) takes the 3-month forward rates as the forecasts of yield curves at a 3-month horizon, that is, $\hat{Y}_t^{(2)}(\tau) = f_{t-63}^{(63)}(\tau)$ corresponding to Equation (5.2). Forecast performance is evaluated using the squared error loss. Figure 2 displays the loss differentials (upper panel) and their average (lower panel). Notably, the average loss differential is above zero for maturities between 3 months and 1 year, but below zero beyond 1 year.

To test the mean hypothesis, we set $w(\tau) = 1, \tau \in [3, 240]$. The resulting test statistic is -1.71, which is not significant at any conventional significance level.[4] Thus, the competing forecasting method does not outperform the benchmark on average across the full domain of maturities.

The weighted test we propose above illustrates how different economic objectives can be incorporated into the hypothesis through the choice of $w(\tau)$. We investigate this by assessing whether the competing predictions perform better over a domain between the 3-month to 1-year maturity, and thus we set $w(\tau) = 1, \tau \in [3, 12]; w(\tau) = 0, \tau \in [12, 240]$ to test this weighted hypothesis. The corresponding test statistic is 0.30, which is again not significant. Therefore, the competing method does not have statistically superior performance in this selected domain of maturity, despite the positive average loss differential observed between 3 month and 1 year in Figure 2 (lower panel). We further compute the self-normalised statistics. The SN statistic of the mean hypothesis is -6.28, and that for the weighted hypothesis is 1.26. Both statistics are not significant,[5] and thus the same conclusion holds. Lastly, the uniform hypothesis is tested using the bootstrap procedure, and the test statistic is 2.74, which is not statistically significant. Overall, there is no evidence to conclude that the competing forecasting method based on 3-month forward rates outperforms the benchmark random walk method in terms of mean, weighted, and uniform hypotheses.

5.2 | Extended Analysis

In our extended application, we use the same framework as in the previous section, but for different forecasting horizons, including 6-month, 9-month, 1-year, 2-year, 3-year, 4-year, 5-year and 10-year. It is important to note that both the forecasting horizon and the two forecasting methods are adjusted accordingly for each specific horizon. For instance, the forecasting horizon is set to be 126 trading days ($h = 21 \times 6$) for 6-month forward rates and 252 trading days ($h = 21 \times 12$) for 1-year forward rates. Then the forecasting methods are given by those in Equations (5.1) and (5.2) for a general horizon h.

14679892, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jtsa.70087, Wiley Online Library on [05/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

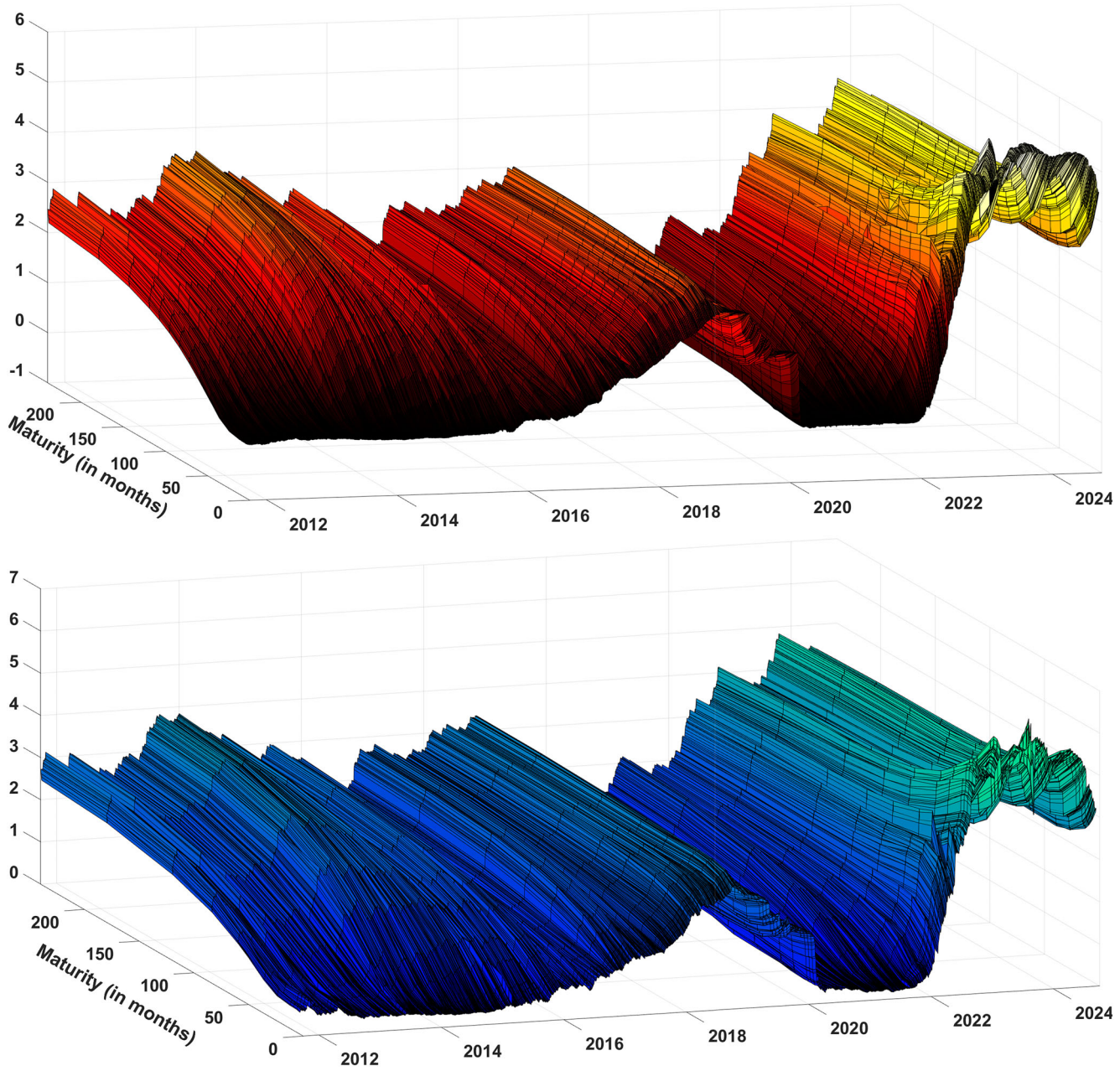


FIGURE 1 | Upper panel: OIS rates; Lower panel: 3 month forward rates.

Given data availability, we always use the longest possible period as the out-of-sample period for testing. Table 5 reports the test results of the mean, weighted, uniform hypothesis for different horizons. Based on the mean and the weighted hypotheses, the benchmark method is superior or equal to the competitor for most horizons. Nevertheless, it is noteworthy that we find significant evidence at 1% level for the 10-year horizon that the competitor improves over the benchmark according to both the LRV-based test and the SN test. This result suggests that using 10-year forward rates provides superior forecasting performance relative to the benchmark method, both on average across the full curve ($\tau \in [3, 240]$) and within the 1-year maturity range ($\tau \in [3, 12]$). In terms of the uniform hypothesis, we observe statistically significant differences for horizons between 6-month and 2-year at the 5% level, and for the 10-year horizon at the 1% level. These findings indicate that the benchmark forecasting method does not uniformly outperform the competing forecasting method on the entire yield curve for those horizons.

It is possible that the forecasting performance may vary from year to year. To investigate this, we apply our tests to annual subsamples. To conserve space,[6] we present the year-by-year result for the 3-month horizon in Table 6, and the 10-year horizon in Table 7. For the 3-month horizon, there is no significant evidence that the competing method outperforms the benchmark in most years, except for 2022 when all three hypotheses are rejected at 1% level. By examining Figure 2, the loss differentials are predominately above zero for most days in 2022, particularly for relatively short maturities. This significant result can be attributed to the systematic increase in the federal funds rate during that

14679892, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jtsa.70087, Wiley Online Library on [05/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

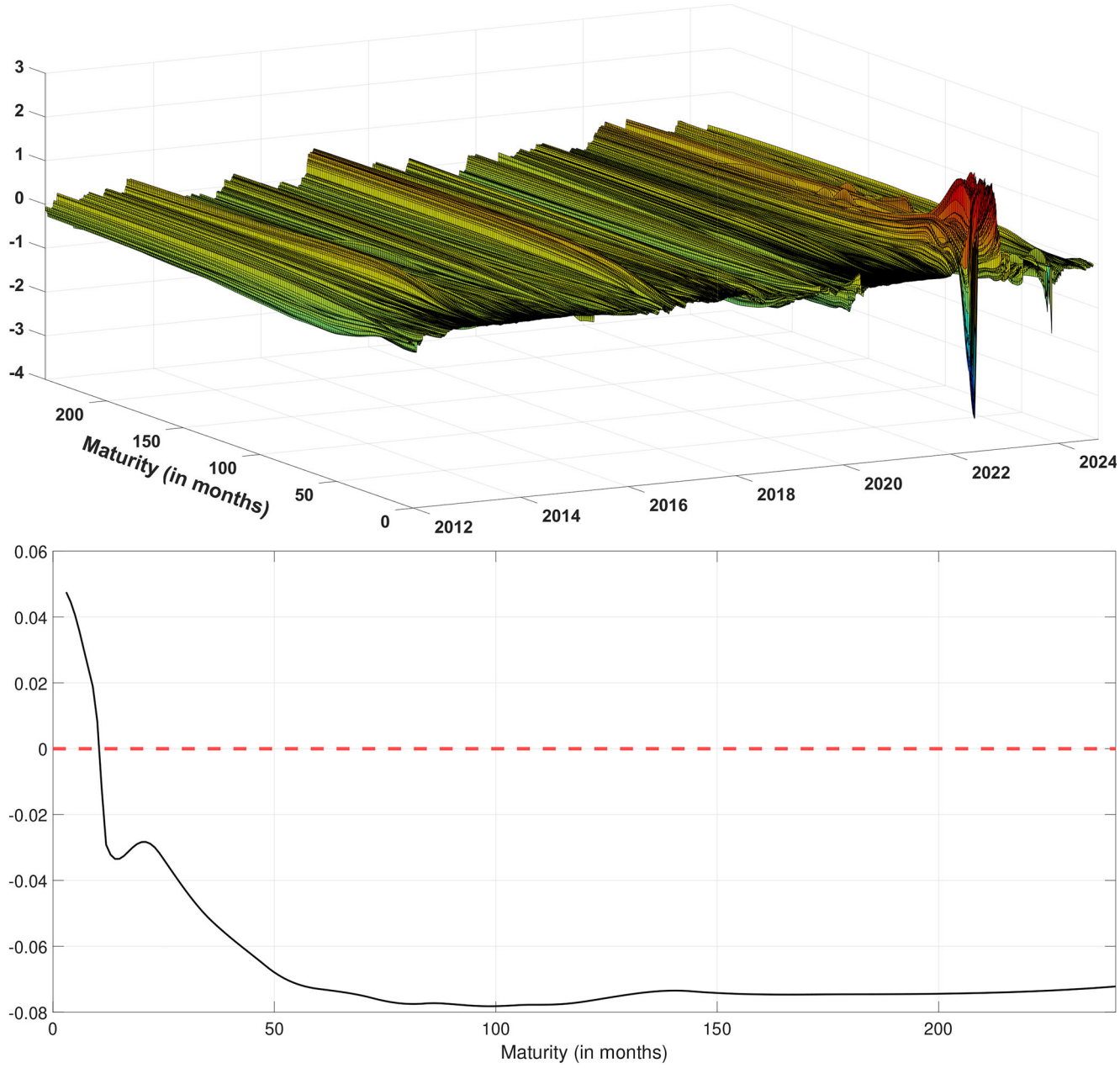


FIGURE 2 | Upper panel: Loss differentials between the two yield curve forecasts; Lower panel: Average loss differentials.

TABLE 5 | Test results of different forecasting horizons.

	Out-of-sample period			Mean Hypo.		Weighted Hypo.		
Horizon	**Start**	**End**	**#Obs.**	**Standard**	**SN**	**Standard**	**SN**	**Unif. Hypo.**
3-month	2011-12-26	2024-09-30	3331	−1.71	−6.28	0.30	1.26	2.74
6-month	2013-06-13	2024-09-30	2948	−0.16	−0.69	0.38	3.76	18.57**
9-month	2013-09-10	2024-09-30	2885	−0.10	−0.44	0.38	3.81	32.80**
1-year	2013-12-06	2024-09-30	2822	−0.06	−0.27	0.48	3.64	37.44**
2-year	2014-11-25	2024-09-30	2570	−0.03	−0.17	0.64	2.71	67.77**
3-year	2015-11-12	2024-09-30	2318	−0.10	−0.61	0.09	1.76	41.32
4-year	2016-10-31	2024-09-30	2066	−0.38	−1.89	0.05	0.39	10.07
5-year	2017-10-18	2024-09-30	1814	−0.83	−3.14	0.07	0.93	9.25
10-year	2022-08-17	2024-09-30	554	9.05***	98.87***	12.65***	83.72***	540.08***

Note: *** denotes significance at the 1% level, ** at the 5% level, and * at the 10% level.

14679892, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jtsa.70087, Wiley Online Library on [05/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 6 | Test results of 3-month forecasting horizon by different years.

	Mean Hypo.		Weighted Hypo.		
Year	**Standard**	**SN**	**Standard**	**SN**	**Unif. Hypo.**
2012	−2.46	−14.85	−1.83	−12.98	0.00
2013	0.10	0.75	−11.29	−31.06	0.85
2014	−7.71	−27.57	−9.67	−57.18	0.00
2015	−2.48	−19.02	−4.34	−19.47	0.00
2016	−0.17	−3.19	−3.88	−29.31	0.00
2017	−0.93	−6.57	0.60	4.01	0.22
2018	−0.08	−1.04	−0.05	−0.62	0.32
2019	−1.64	−10.65	−2.08	−10.65	0.00
2020	−1.41	−5.78	−1.09	−9.80	0.00
2021	0.18	1.36	−2.09	−5.92	0.53
2022	2.42***	24.94***	2.75***	18.54***	18.82***
2023	−0.08	−1.14	−0.60	−1.37	0.13
2024	−1.55	−7.32	−2.82	−20.68	0.00

Note: *** denotes significance at 1% level, ** at 5% level, and * at 10% level.

TABLE 7 | Test results of 10-year forecasting horizon by different years.

	Mean Hypo.		Weighted Hypo.		
Year	**Standard**	**SN**	**Standard**	**SN**	**Unif. Hypo.**
2022	11.65***	55.14***	8.63***	31.41***	183.29***
2023	4.99***	25.23***	29.02***	97.48***	414.83***
2024	7.17***	38.19***	19.33***	39.31***	327.09***

Note: *** denotes significance at 1% level, ** at 5% level, and * at 10% level.

year. Following the Federal Open Market Committee (FOMC) meeting on March 16, 2022, the Federal Reserve released a statement detailing its decision to raise the target range for the federal funds rate to the range 0.25%–0.5% and anticipated that ongoing increases in the target range would be appropriate.[7] In terms of the 10-year horizon, all three hypotheses are rejected at 1% level in 2022, 2023, and 2024. This indicates that the superior performance of the competing forecasting method is not driven by a single year but is consistent across the three years of the out-of-sample period.

6 | Conclusion

We have developed a framework for evaluating two sets of competing functional forecasts. Building on the classic Diebold-Mariano test, we propose several test variants which aim to provide statistical guidance about whether a specific forecast method improves on simple forecasts or not. We have derived these tests using new asymptotic theory and demonstrated their performance by Monte Carlo simulations. An empirical application to forecasting the term structure of interest rates was used to show how the test works in practice, and revealed somewhat unexpected properties of forecast based on forward rates.

By developing the classic DM-type test for functional data, there are a variety of possible future developments. Firstly, we have focussed on the model-free context for functional forecasts. In order to accommodate model-based functional forecasts, new theory to establish that the asymptotic distribution of the DM test is not affected by parameter estimation, as West (1996) and subsequent papers establish in the scalar context, is required. Another potentially useful extension would be to accommodate the comparison of many functional forecasting models. We leave these to be addressed in future work, and also hope that this paper will motivate other researchers to pursue such extensions.

Acknowledgements

We are grateful for comments from Valentina Corradi, Ilias Chronopoulos, Jean-Yves Pitarakis, Malte Knüppel, Matteo Fontana, Xiaochun Meng, Rob Taylor, and participants of the "*Econometrics Workshop on Model Selection, Predictability and High Dimensionality: Theory and Applications*" at the University of Southampton; the "*Quarterly Forecasting Forum*" in London, May 2025; CFE-CMStatistics 2025, London; the ECMFE Seminar, University of Essex; the Brown Bag Seminar, Department of Finance and Accounting, University of Reading; and for departmental seminars at the University of Leicester, Shanghai Maritime University and Xiamen University. Piotr Kokoszka acknowledges support of the United States National Science Foundation.

14679892, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jtsa.70087, Wiley Online Library on [05/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Funding

This work was supported by the U.S. National Science Foundation grant DMS-2412408.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data in this paper are available from Refinitiv EIKON (formerly Datastream). Restrictions apply to the availability of these data, which were used under licence for this study. Data are available from the authors with the permission of Refinitiv.

Endnotes

[1]We also experimented with alternative kernels and bandwidth selections, such as the Bartlett kernel with bandwidth $h = \log(T)$, but the performance is not as good as the chosen kernel.

[2]We do not introduce additional measure error on the sparse observations.

[3]Additional tenors are also available for U.S. OIS rates, including 1 month, 2 month, 25 year, 30 year, 35 year, 40 year, 45 year, and 50 year. We exclude these maturities because the forward rates are available only from 3 months to 20 years.

[4]The critical values are 1.28 at 10%, 1.64 at 5%, and 2.33 at 1%.

[5]The critical values of the SN statistics of the sample size of 3331 are 7.24 at 10%, 10.013 at 5%, and 16.398 at 1%.

[6]The year-by-year result for other horizons are available upon a request.

[7]https://www.federalreserve.gov/newsevents/pressreleases/monetary20220316a.htm, assessed on 20th September, 2025.

References

Andrews, D. 1984. "Non-Strong Mixing Autoregressive Processes." *Journal of Applied Probability* 21, no. 4: 930–934.

Berkes, I., L. Horváth, and G. Rice. 2013. "Weak Invariance Principles for Sums of Dependent Random Functions." *Stochastic Processes and their Applications* 123: 385–403.

Cai, L., and Q. Hu. 2025. "From Sparse to Dense Functional Data: Phase Transitions From a Simultaneous Inference Perspective." *Statistics and Computing* 35: 146.

Clark, T. E., and M. W. McCracken. 2001. "Tests of Equal Forecast Accuracy and Encompassing for Nested Models." *Journal of Econometrics* 105, no. 1: 85–110.

Clark, T. E., and M. W. McCracken. 2005. "Evaluating Direct Multistep Forecasts." *Econometric Reviews* 24, no. 4: 369–404.

Clark, T. E., and M. W. McCracken. 2014. "Tests of Equal Forecast Accuracy for Overlapping Models." *Journal of Applied Econometrics* 29, no. 3: 415–430.

Coroneo, L., and F. Iacone. 2020. "Comparing Predictive Accuracy in Small Samples Using Fixed-Smoothing Asymptotics." *Journal of Applied Econometrics* 35, no. 4: 391–409.

Coroneo, L., F. Iacone, and F. Profumo. 2024. "Survey Density Forecast Comparison in Small Samples." *International Journal of Forecasting* 40, no. 4: 1486–1504.

Corradi, V., J. Fosten, and D. Gutknecht. 2024. "Predictive Ability Tests With Possibly Overlapping Models." *Journal of Econometrics* 241, no. 1: 105716.

Dehling, H. 1983. "Limit Theorems for Sums of Weakly Dependent Banach Space Valued Random Variables." *Zeitschrift für Wahrscheinlichkeitstheorie Und Verwandte Gebiete* 63: 393–432.

Dette, H., G. Dierickx, and T. Kutta. 2024. "Testing Covariance Separability for Continuous Functional Data." *Journal of Time Series Analysis* 46: 402–420.

Diebold, F. X., and R. S. Mariano. 1995. "Comparing Predictive Accuracy." *Journal of Business and Economic Statistics* 13, no. 3: 253–263.

Gonçalves, S., M. W. McCracken, and B. Perron. 2017. "Tests of Equal Accuracy for Nested Models With Estimated Factors." *Journal of Econometrics* 198, no. 2: 231–252.

Hansen, P. R., A. Lunde, and J. M. Nason. 2011. "The Model Confidence Set." *Econometrica* 79, no. 2: 453–497.

Hörmann, S., and P. Kokoszka. 2010. "Weakly Dependent Functional Data." *Annals of Statistics* 38, no. 3: 1845–1884.

Horváth, L., P. Kokoszka, and R. Reeder. 2013. "Estimation of the Mean of Functional Time Series and a Two-Sample Problem." *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 75, no. 1: 103–122.

Horváth, L., P. Kokoszka, and G. Rice. 2014. "Testing Stationarity of Functional Time Series." *Journal of Econometrics* 179, no. 1: 66–82.

Horváth, L., Z. Liu, G. Rice, and S. Wang. 2020. "A Functional Time Series Analysis of Forward Curves Derived From Commodity Futures." *International Journal of Forecasting* 36, no. 2: 646–665.

Janson, S., and S. Kaijser. 2015. "Higher Moments of Banach Space Valued Random Variables." *American Mathematical Society* 238, no. 1127.

Kuelbs, T. W. 1973. "The Invariance Principle for Banach Space Valued Random Variables." *Journal of Multivariate Analysis* 3: 161–172.

Kutta, T., H. Dette, and S. Wang. 2025. "Multiscale Change Point Detection for Functional Time Series. arXiv preprint arXiv:2511.06870."

Kutta, T., and N. Dörnemann. 2025. "Monitoring Time Series With Short Detection Delay." *Electronic Journal of Statistics* 19, no. 1: 2239–2275.

McCracken, M. W. 2000. "Robust Out-Of-Sample Inference." *Journal of Econometrics* 99, no. 2: 195–223.

Newey, W. K., and K. D. West. 1987. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica* 55, no. 3: 703–708.

Pitarakis, J.-Y. 2025. "A Novel Approach to Predictive Accuracy Testing in Nested Environments." *Econometric Theory* 41, no. 1: 35–78.

Račkauskas, A., and C. Suquet. 2006. "Testing Epidemic Changes of Infinite Dimensional Parameters." *Statistical Inference for Stochastic Processes* 9: 111–134.

Račkauskas, A., and C. Suquet. 2009. "Hölderian Invariance Principle for Hilbertian Linear Processes." *ESAIM Probability and Statistics* 13: 261–275.

Romano, J. P., and M. Wolf. 2005. "Stepwise Multiple Testing as Formalized Data Snooping." *Econometrica* 73, no. 4: 1237–1282.

Samur, J. D. 1987. "On the Invariance Principle for Stationary φ-Mixing Triangular Arrays With Infinitely Divisible Limits." *Probability Theory and Related Fields* 75: 245–259.

Shang, H. L., and F. Kearney. 2022. "Dynamic Functional Time-Series Forecasts of Foreign Exchange Implied Volatility Surfaces." *International Journal of Forecasting* 38, no. 3: 1025–1049.

Shao, X. 2015. "Self-Normalization for Time Series: A Review of Recent Developments." *Journal of the American Statistical Association* 110, no. 512: 1797–1817.

West, K. D. 1996. "Asymptotic Inference About Predictive Ability." *Econometrica* 64, no. 5: 1067–1084.

14679892, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jtsa.70087, Wiley Online Library on [05/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

White, H. 2000. "A Reality Check for Data Snooping." *Econometrica* 68, no. 5: 1097–1126.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Table S1:** Simulated quantiles of $\tilde{S}_T^{SN}$. **Figure S1:** Representative functional mean difference. **Table S2:** Empirical rejection rates of testing the uniform hypothesis with other choices of a_T at the 5% significance level. **Table S3:** Empirical rejection rates under reconstruction from discretely observed curves. **Table S4:** Empirical rejection rates under stronger temporal dependence ($\|\psi\| = 9/10$).

14679892, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/jtsa.70087, Wiley Online Library on [05/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License