ROC approach to forecasting recessions using daily yield spreads

Kajal Lahiri1 · Cheng Yang2

Published online: 5 November 2022 © National Association for Business Economics 2022

Abstract
Even though many studies have established the existence of structural breaks and declining predictability in the relationship between GDP growth and yield spreads, business analysts continue to watch for the inversion of the spread as one of the leading indicators for recessions. We use the Receiving Operating Characteristics (ROC) approach, to reevaluate the enduring power of spread to forecast recessions, notwithstanding the temporal instabilities. We identify the value of the spread that produces the highest discriminatory power as measured by different functionals of the ROC curve e.g., the hit rate, false alarm rate, and the Youden’s index. Based on data starting from January 2, 1962, we find that the optimal threshold has drifted upwards from zero since the early 1980s, and the deteriorating power of yield spread can largely be restored once the optimal cut-off values are used to issue recession forecasts.

1 Introduction
The term spread, calculated as the difference between the yields on the ten-year Treasury note and the 3-month Treasury bill rate, became inverted in the U.S. in the last week of May 2019, and stayed that way till the first week of October 2019. This inversion followed a steady stream of policy hikes in the federal funds rate from early 2016 till January 2019, during time which the targeted policy rate was increased from about 0 to 2.5%. Similar to the last seven recessions, the inversion of the yield curve slope preceded the downturn that started in February 2020, with a lead time of nearly 12 months. According to the National Bureau of Economic Research (NBER) Dating Committee announcement of June 8, 2020, the recession began distinctly ahead of the pandemic-related free fall of the economy after the COVID-19 lockdown.

Since the slope of the Treasury yield curve has an almost unblemished record of forecasting recessions, there has been a lot of research exploring the pathways through which the effects operate, see Cooper et al. (2020) for a recent summary. The spread reflects the current stance of monetary policy and its complex interactions with expected future monetary policies that, in turn, are linked to expectations of future business cycle outcomes. The yield curve slope summarizes many different unobserved and observed variables in one composite indicator.

Concurrent to the above developments, however, there has been a parallel body of research on the hypothesis that the yield spread has lost its mojo since the 1980s. The failure of the experimental recession index developed by Stock and Watson (1993) was attributed to their reliance on these spread variables, see Dotsey (1998) and Jardet (2004). Many factors including unconventional monetary policies, financial innovations, deregulation, deepening of the commercial paper market, increasing globalization, structural breaks, and inflation targeting have been proposed as reasons for the loss in forecasting power, see Pažicky (2021) and Giacommini and Rossi (2006). Chauvet and Potter (2005) utilize a number of alternative models to accommodate some of the documented instabilities in yield curve prediction models. It is interesting to note that the yield spread got included in the list of ten leading indicators of The Conference Board (TCB) only in 1996, and since then it has maintained its coveted position in the list, see Levanon et al. (2015). These considerations bring to the forefront an unresolved puzzle.
first put forward by Rudebusch and Williams (2009): despite the overwhelming econometric evidence on the usefulness of the yield spread, professional forecasters have not incorporated this information in projecting recessions. Lahiri and Wang (2013) have shown that probability forecasts for real GDP declines issued by the Survey of Professional Forecasters have no statistical value at the four-quarter horizon, where the yield spread has its maximum power.

In this paper we evaluate the value the Treasury yield spread has to predict business cycle peaks from the perspective of a recurring binary event that is relatively rare or uncommon. After all, in the U.S. recessions have occurred only seven times during last 60 years. In several ways our approach is novel:

First, the standard workhorse in this sphere has been the probit model, e.g., Estrella and Hardouvelis (1991) and Estrella and Mishkin (1996). The probit link function is symmetric and may not be consistent with the loss function of the forecast user. Building on Berge and Jordà (2011), we use the Receiving Operating Characteristics (ROC) approach that directly focuses on the identification of a binary event, and evaluates the trade-offs between missed signals and false alarms that may help to resolve the Rudebusch–Williams puzzle. The commonly used goodness-of-fit criteria like the concordance index or the mean squared forecast errors tend to get overwhelmed by the dominant event of the sample, see Stephenson (2000) and Lahiri and Yang (2013).

Second, like Bauer and Mertens (2018), we translate the yields directly into binary predictions without going through an intermediate step of estimating the probabilities from a probit or other models. Liu and Moench (2016) and Miller (2019) have used the ROC approach to evaluate selected recession indicators using the Area under the Curve (AUC) statistic associated with a ROC. But they obtained the ROC curves using generated probabilities from underlying probit models. As we will argue later, because the AUC statistic is obtained by evaluating over the whole probability space, it may fail to identify the goodness of a predictor in the region where it matters, cf. Elliott and Lieli (2013) and Zhou et al. (2011). Thus, our emphasis is somewhat different from Berge and Jordà (2011) and others in that we focus on specific points of the ROC curve to examine the associated hit and false alarm rates. More importantly, since financial and market analysts directly monitor yield spreads and watch when they cross the zero threshold (i.e., inversion of the yield curve), our approach of generating spreads and watch when they cross the zero threshold (i.e., the inversion of the yield curve), our approach of generating spread curves using generated probabilities from underlying parametric models, is a difference of two variables and is stationary. Business analysts do not wait till the end of the month to get overwhelmed by the dominant event of the sample, see Stephenson (2000) and Lahiri and Yang (2013).

Finally, and importantly, since the market analysts and the policy makers monitor indicators like yield spreads on a continuous basis, we evaluate the predictive value using daily yields. Analysts do not wait till the end of the month or quarter to monitor these values. Also, a larger data base helps us to derive more precise inference.

The rest of the paper is structured as follows: In Sect. 2, we introduce the ROC approach, and present the data analysis to delineate the underlying type I and type II errors. Finally, conclusions are summarized in Sect. 3.

## 2 Forecast evaluation methods using ROC

### 2.1 Interest rate spread as threshold

Unlike many of other leading indicators, interest rate spread is a difference of two variables and is stationary. Business analysts can make predictions about recessions by directly looking at the spread values. This approach is often preferred by practitioners in that it is straightforward and free from various parametric models. Instead of having thresholds based on estimated probabilities from parametric models, we can use the interest rate spread itself as the threshold. With daily data, our objective is to link the spread directly to the binary recession forecasts.

Given a chosen threshold $\hat{\delta}$ for the daily spread $x_{s,t}$, a recession is expected to start in any day of a pre-selected future period if the spread value in day $s$ of month $t$ falls to or below $\hat{\delta}$.

$$\hat{y}_{s,t}(x_{s,t};\hat{\delta}) = 1(x_{s,t} \leq \hat{\delta}).$$

(1)

In our application, the left-hand-side variable $y_{s,t}$ is an indicator that is equal to 1 if a new recession starts any time during the next 12 months and 0 otherwise.

There are four possible outcomes that describe the joint distribution of the binary forecasts and outcomes: (a) forecast is recession and outcome is recession ($\hat{y}_{s,t} = 1, y_{s,t} = 1$); (b) forecast is not recession and outcome is recession ($\hat{y}_{s,t} = 0, y_{s,t} = 1$); (c) forecast is recession and outcome is not recession ($\hat{y}_{s,t} = 1, y_{s,t} = 0$); and (d) forecast is not recession and outcome is not recession ($\hat{y}_{s,t} = 0, y_{s,t} = 0$). The proportion of each case is $\pi_{ij}$ with $i,j \in \{0, 1\}$ and is summarized in a $2 \times 2$ contingency table, viz., see Table 1.
We define $H$ as hit rate, the proportion of recession periods that are correctly predicted: $H = \pi_{11}/(\pi_{11} + \pi_{00})$. $H$ is also called true positive rate or sensitivity. False alarm rate ($F$) is defined as the proportion of non-recession periods that are mistakenly considered to be recessions: $F = \pi_{10}/(\pi_{10} + \pi_{00})$. $F$ is also called false positive rate or one minus specificity. Both $H$ and $F$ are functions of threshold $\delta$. By plotting $F$ against $H$ for all possible spread thresholds $\delta$, we have the Receiver Operating Characteristic (ROC) curve. $F$ is depicted on the horizontal axis and $H$ on the vertical axis, both ranging from 0 to 1. If $\delta$ is smaller than even the minimum value of the spread, all forecasts $\hat{y}_{t,i}$ are 0. Both hit rate and false alarm rate are 0. Thus points with smaller spread thresholds are closer to bottom-left corner. If $\delta$ is larger than the maximum value of the spread, all forecasts $\hat{y}_{t,i}$ are 1. Both hit rate and false alarm rate are 1. Points with larger spread thresholds are closer to upper-right corner. There is a trade-off between $H$ and $F$, where $H$ increases with $F$. The ROC curve gives a complete summary of these trade-offs given the past history of a predictor or a forecasting system. The ROC curves of better forecasts are closer to the upper-left corner. Perfect forecasts have the ROC curve connecting bottom-left, upper-left and upper-right corners. Points on the 45° line represent thresholds generating forecasts which are not informative and are same as random guesses. Points below the 45° line represent thresholds which give forecasts even worse than random guesses.

Another measure for forecast evaluation is concordance ($C$), which tells how much two binary outcome series $I_t$ and $S_t$ move together:

$$C = \frac{1}{T} \sum_{i=1}^{T} (S_i I_t + (1 - S_i)(1 - I_t))$$

Concordance has been used to determine if the series is pro-cyclical or counter-cyclical (Harding and Pagan 2002). When $I_t$ and $S_t$ are recession indicator and forecast, concordance measures the fraction of correct forecasts. Clearly concordance is a function of the threshold $\delta$. It can also be written as a function of hit and false alarm rates, which also depend on the threshold:

$$C(\delta) = \frac{T_1}{T} H(\delta) - \frac{T_0}{T} F(\delta) + \frac{T_0}{T}$$

where $T_0$ is the number of observations in expansions and $T_1$ is the same in recessions. Compared with $KS$, concordance gives more weight to the false alarm rate since $T_0/T$ is much larger than $T_1/T$, as in our case. Thus the concordance criterion is such that the loss or dis-utility from mispredicting a recession is less than that from giving a false signal. As a result, we tend to get lower hit rate and lower false alarm rate if the threshold is chosen to maximize the concordance. Thus, in forecasting rare events, the maximized value of this criterion may look impressive, yet the fitted model may fail to identify adequately the rare event, which is often the target for identification.

### 2.2 Data, implementation and results

The monthly NBER recession indicator is used as the basis for the left-hand-side outcome variable. The daily data of the spread for evaluation is the 10-year rate minus the 3-month rate, from January 2, 1962 to September 22, 2022 (Fig. 1).
Not only does the 3-month rate reflect the market rate better, Bauer and Mertens (2018) show that the difference between the 10-year and 3-month rates is the most effective term spread to forecast recessions without any adjustments for term premium or quantitative easing.\footnote{Fendel et al. (2021) suggest a modification of the traditional yield curve due to the zero lower bound to restore its predictive power of the Euro area business cycle using AUC. See also Pažický (2021) and Duarte et al. (2005).}

In order to simplify our analysis when defining horizon, we assume the number of trading days in each month is constant and equal to 22. If there are fewer than 22 values and there are holidays on weekdays in a month, they are treated as missing values and are linearly interpolated. If there are still fewer than 22 values, missing values are created between two months and are also linearly interpolated. In either case, missing values are linearly interpolated so that there are 22 values in each month. If there are more than 22 values in a month, following Ghysels et al. (2020), we replace the last several observations in that month with their average such that every month has 22 values.

Our target variable takes value 1 if a new recession starts any time during the next 12 months, and 0 otherwise. Consistent with the original NBER definition, this target does not pre-specify a fixed horizon and has been used in a number of recent papers, cf. Wright (2006), Ergungor (2016), Johansson and Meldrum (2018), Bauer and Mertens (2018), and others.

Historically, the causes and mechanisms of recessions have been different over time. In addition, the optimal threshold for the spread variable can change due to trends in financial innovations in the economy. Therefore, we allow different optimal thresholds over time by working with different sub-samples. We start from January 1962 and add several years each time in such a way that one more recession is included in the sample that ends exactly 12 months before a recession begins. By doing this cumulatively and recursively, we are able to find how the optimal threshold changed over time at the time recession forecasts were made. Specifically, we take seven sub-samples: sample 1 (1/2/1962–1/31/1979), sample 2 (1/2/1962–7/31/1980), sample 3 (1/2/1962–7/31/1989), sample 4 (1/2/1962 to 3/31/2000), sample 5 (1/2/1962–12/29/2006), sample 6 (1/2/1962–2/28/2019), and finally sample 7 (1/2/1962–9/22/2022), which is our latest full sample. Each sub-sample ends precisely 12 months before the beginning of the 1980, 1982, 1990, 2001, 2008 and the 2020 recessions because the spread has the maximum predictive power around this horizon.

We experimented with thresholds that span the whole sample of spread values, but only few relevant ones (seven for each case with the fourth one with maximized Kuipers score) are reported in six panels. From Table 3 we find that the optimum threshold that maximized KS was never close to zero, with the lowest value of 0.2 in early sub-samples 2 and 3. In recent years it has steadily increased to 0.91. It may be noted that Kuipers score, defined as \( H - F \), gives equal weight to missed signals (\( 1 - H \)) and false alarms. If this loss function is not compatible with the utility function, the forecast user can inspect the associated pairs of hit rates and false alarm rates separately and chose a particular threshold that is acceptable. Thus one may insist having at least a 90\% (or 80\%) hit rate and accept the implied threshold; the false alarm will automatically follow given the quality of the predictor. Given the shape of the ROC curve, the false alarm rate will necessarily be higher with a higher hit rate. By inspecting the last three sub-samples in Table 3, we find that enjoying a hit rate of 90\% will impose the disutility of having a false alarm rate of around 25–30\%. The ROC analysis sheds light of the intrinsic quality of the spread variable as a predictor of recession—it defines the frontier of trade-offs between \( H \) and \( F \). For a chosen threshold, it quantifies the underlying uncertainty of a recession call in terms of the associated hit and false alarm rates. If a false alarm rate of 25–30\% is too high, the practitioner should possibly choose a different threshold which would give a lower false alarm rate (and hit rate), although it may no longer maximizes KS. For example, a threshold value of 0.51\% in Table 4 would have reduced the false alarm rate to 15\%, but at the cost of reducing the hit rate to 80\% from 94\%.

Using ROC analysis, we can determine if the forecastability of the spread has fallen over time—an issue that has attracted a great deal of attention over last three decades. If we chose Kuipers score as our maximizing metric and allow thresholds to adjust accordingly over time, we find no evidence of such decay. In Fig. 2, we have plotted the ROC curves for each sub-sample with the associated AUC values.

Details on the hit rate, false alarm rate and the associated Kuipers score for selected values of the threshold for different sub-samples are presented in Table 2.\footnote{In what follows, we do not focus on the results using Concordance. Given that the percentage of days in recession is only 14\%, the Concordance index was very high around 0.88 quite independent of the threshold values along the ROC curve. The resultant hit rates were close to 0.50 with a false alarm rates around 0.05 at zero threshold; see Table 3. Unless the relative preference for \( H \) and \( F \) for a forecast client is close to these, Concordance is not an appropriate criterion to pursue in evaluating forecasts for relatively rare or uncommon events like a recession. Given the tremendous societal cost of an unanticipated recession, a hit rate of only 50% may be considered to be unacceptably low by many forecast users. Kuipers score is well-known to be more appropriate in these cases, Stephenson (2000).}
Table 2  Daily spread forecast performance by threshold

| Sample 1 (1/2/1962 to 1/31/1979) | Sample 2 (1/2/1962 to 7/31/1980) |
|----------------------------------|----------------------------------|
| Threshold | $H$ | $F$ | $KS$ | Threshold | $H$ | $F$ | $KS$ |
| 0.37 | 0.75 | 0.27 | 0.48 | 0.12 | 0.56 | 0.09 | 0.47 |
| 0.45 | 0.79 | 0.30 | 0.49 | 0.01 | 0.69 | 0.11 | 0.57 |
| 0.52 | 0.82 | 0.31 | 0.50 | 0.14 | 0.74 | 0.15 | 0.59 |
| 0.57 | 0.84 | 0.32 | 0.51 | 0.20 | 0.77 | 0.17 | 0.60 |
| 0.60 | 0.84 | 0.34 | 0.51 | 0.26 | 0.79 | 0.20 | 0.58 |
| 0.68 | 0.85 | 0.39 | 0.47 | 0.39 | 0.84 | 0.28 | 0.56 |
| 0.75 | 0.86 | 0.42 | 0.44 | 0.52 | 0.88 | 0.32 | 0.56 |

| Sample 3 (1/2/1962 to 7/31/1989) | Sample 4 (1/2/1962 to 3/31/2000) |
|----------------------------------|----------------------------------|
| Threshold | $H$ | $F$ | $KS$ | Threshold | $H$ | $F$ | $KS$ |
| −0.18 | 0.56 | 0.05 | 0.50 | 0.19 | 0.64 | 0.09 | 0.55 |
| −0.04 | 0.65 | 0.07 | 0.58 | 0.37 | 0.72 | 0.15 | 0.57 |
| 0.10 | 0.73 | 0.10 | 0.63 | 0.56 | 0.79 | 0.20 | 0.59 |
| 0.21 | 0.77 | 0.13 | 0.64 | 0.91 | 0.90 | 0.30 | 0.60 |
| 0.37 | 0.82 | 0.20 | 0.63 | 1.10 | 0.93 | 0.36 | 0.57 |
| 0.51 | 0.87 | 0.23 | 0.64 | 1.29 | 0.97 | 0.43 | 0.54 |
| 0.65 | 0.88 | 0.27 | 0.61 | 1.47 | 0.98 | 0.48 | 0.50 |

| Sample 5 (1/2/1962 to 12/29/2006) | Sample 6 (1/2/1962 to 2/28/2019) |
|----------------------------------|----------------------------------|
| Threshold | $H$ | $F$ | $KS$ | Threshold | $H$ | $F$ | $KS$ |
| 0.37 | 0.73 | 0.15 | 0.58 | 0.37 | 0.71 | 0.12 | 0.59 |
| 0.56 | 0.81 | 0.20 | 0.61 | 0.56 | 0.79 | 0.16 | 0.63 |
| 0.74 | 0.86 | 0.26 | 0.60 | 0.74 | 0.85 | 0.21 | 0.64 |
| 0.91 | 0.92 | 0.30 | 0.62 | 0.91 | 0.91 | 0.24 | 0.67 |
| 1.10 | 0.94 | 0.35 | 0.59 | 1.10 | 0.94 | 0.29 | 0.65 |
| 1.29 | 0.97 | 0.41 | 0.57 | 1.29 | 0.98 | 0.35 | 0.62 |
| 1.47 | 0.98 | 0.45 | 0.53 | 1.47 | 0.99 | 0.40 | 0.58 |

| Sample 7 (1/2/1962 to 9/20/2022) |
|----------------------------------|
| Threshold | $H$ | $F$ | $KS$ |
| 0.37 | 0.74 | 0.12 | 0.62 |
| 0.56 | 0.82 | 0.16 | 0.65 |
| 0.74 | 0.87 | 0.22 | 0.65 |
| 0.91 | 0.92 | 0.25 | 0.67 |
| 1.10 | 0.95 | 0.30 | 0.65 |
| 1.29 | 0.98 | 0.36 | 0.62 |
| 1.47 | 0.99 | 0.41 | 0.58 |

$H$ and $F$ are hit rate and false alarm rate at selected thresholds. Each sample end 12 months before the beginning of a recession.

Given on the right lower part of each square. These values, which are calculated over all thresholds, increased slowly from 0.83 in the 1/2/1962–1/31/1979 period to 0.91 in the 1/2/1962–9/22/2022 period. Clearly the evidence does not suggest an overall deterioration in the forecasting prowess of the spread. On the other hand, if we had stuck to a zero threshold (i.e., wait for the yield spread to invert) over the sub-samples, the hit rate would have deteriorated significantly from 68% to 50% and Kuipers score from 60% to 45% with a modest improvement in the false alarm rate from 10% to 5% (see the right panel of Table 4). This scenario would have clearly suggested a deterioration in the forecasting value of yield spread under reasonable utility functions.

Since, compared with the recessions of early 1980s, the estimated probabilities of recession from the probit model showed muted increases before the 1990, 2000 and 2007 recessions (see Fig. 3), economists presumed that yield spreads are not signaling forthcoming slowdowns...
as before. Several types of structural breaks in the economy were proposed as reasons for the loss of forecasting power of the spread. In Fig. 4, we have presented the same probabilities, but now excluding the early 1980s sample. We find that the recession probabilities are no longer muted when the extraordinary gyrations in the spread in the early 1980s are excluded. The early 1980s was a period of a known and well accepted structural break due to the change in the operating policy of the Fed. Note that an important aspect of forecasting is timeliness or lead time. Table 5 shows that the lead times for peak turning points with the Kuipers-maximizing thresholds in each of the sub-samples are considerably better compared with the lead times with zero threshold, particularly during last 10 years.

We plot the conditional densities for the spread values in Fig. 5, one distribution for recessions and the other for the non-recession period using the whole sample. AUC can be interpreted as the amount of overlap between the two conditional densities. If they overlap completely, it means the predictor has no discriminatory power between recessions and non-recessions, and hence AUC will be 0.5. We see little overlap, consistent with the AUC value of 0.913 reported in Fig. 2. It is interesting to note that theoretically, under some simplifying assumptions, the intersection of the two conditionals give the threshold spread that maximizes the Kuipers score, cf. Manzato (2007). Indeed it is 0.91 in Fig. 5 consistent with what we reported in Table 2 for the whole sample.

The NBER dating Committee has determined that a recession in the U.S. had started in February 2020, and in the 4th quarter of 2019 in its quarterly chronology. The coincident indicators like the household employment, real personal income and personal income less transfers all peaked in February, ahead of the COVID-19 economic shutdown that began in late March. Even though the shutdown created an unprecedented decline in economic activity including unemployment claims after March 2020, the economy clearly started to slow down before the Covid-19 shutdown. Whether the economy would have been classified by NBER retrospectively into the recession without the pandemic of early 2020 can be debated, but if our signals for a recession from the yield spread from February 2019 (see Table 4) are taken as counterfactual predictions without the pandemic, then we can conclude that yield spread predicted the recession again this time.

The latest recession, like other occasions, defied numerous media commentaries and expert judgments during the latter half of 2019 that even though the yield curve inverted, ‘it’s different this time’. In a December 2017 FOMC meeting, the possibility of an economic downturn was ruled out

| Sample | Optimal threshold | AUC | Zero threshold |
|--------|-------------------|-----|----------------|
|        | Threshold | $KS$ | $H$ | $F$ | $KS$ | $H$ | $F$ |
| 1      | 0.57       | 0.51 | 0.84 | 0.32 | 0.83 | 0.43 | 0.53 | 0.10 |
| 2      | 0.20       | 0.60 | 0.77 | 0.17 | 0.87 | 0.57 | 0.68 | 0.11 |
| 3      | 0.21       | 0.64 | 0.77 | 0.13 | 0.90 | 0.60 | 0.68 | 0.08 |
| 4      | 0.91       | 0.60 | 0.90 | 0.30 | 0.89 | 0.49 | 0.55 | 0.06 |
| 5      | 0.91       | 0.62 | 0.92 | 0.30 | 0.89 | 0.47 | 0.53 | 0.06 |
| 6      | 0.91       | 0.67 | 0.91 | 0.24 | 0.91 | 0.47 | 0.51 | 0.05 |
| 7      | 0.91       | 0.67 | 0.92 | 0.25 | 0.91 | 0.45 | 0.50 | 0.05 |

Sample 1: 1/2/1962 to 1/31/1979; Sample 2: 1/2/1962 to 7/31/1980; Sample 3: 1/2/1962 to 7/31/1989; Sample 4: 1/2/1962 to 3/31/2000; Sample 5: 1/2/1962 to 12/29/2006; Sample 6: 1/2/1962 to 2/28/2019; Each sub-sample ended 12 months before the beginning of the recessions of 1980, 1981, 1990, 2001, 2008 and 2020, respectively. Sample 7 is our full sample.

| Threshold | $H$ | $F$ | $C$ | $KS$ |
|-----------|-----|-----|-----|-----|
| − 2.24    | 0.02 | 0.00 | 0.86 | 0.02 |
| − 1.99    | 0.03 | 0.00 | 0.86 | 0.03 |
| − 1.74    | 0.05 | 0.00 | 0.87 | 0.04 |
| − 1.49    | 0.07 | 0.00 | 0.87 | 0.06 |
| − 1.24    | 0.09 | 0.00 | 0.87 | 0.09 |
| − 0.99    | 0.12 | 0.01 | 0.87 | 0.11 |
| − 0.74    | 0.15 | 0.01 | 0.87 | 0.14 |
| − 0.49    | 0.20 | 0.02 | 0.88 | 0.18 |
| − 0.24    | 0.34 | 0.03 | 0.89 | 0.31 |
| 0.00      | 0.50 | 0.04 | 0.89 | 0.46 |
| 0.26      | 0.67 | 0.09 | 0.88 | 0.58 |
| 0.51      | 0.80 | 0.15 | 0.85 | 0.65 |
| 0.76      | 0.88 | 0.22 | 0.79 | 0.66 |
| 1.01      | 0.94 | 0.28 | 0.75 | 0.66 |
| 1.26      | 0.98 | 0.36 | 0.69 | 0.62 |
| 1.51      | 0.99 | 0.43 | 0.63 | 0.56 |
| 1.76      | 0.99 | 0.51 | 0.57 | 0.48 |

$H$ and $F$ are hit rate and false alarm rate at selected thresholds. $C$ is concordance. Sample is from 1/2/1962 to 9/20/2022.
even though some of the members foresaw a flattening of the yield curve, see Johansson and Meldrum (2018). In our analysis, the conventional spread signaled a recession holistically with an underlying hit rate and false alarm rates of 92% and 25%, respectively (Table 4). These probabilities remained almost the same around 92% and 24% till February 2020 using the optimized threshold value of 0.91. Thus, the signal for a recession remained strong throughout the year before pandemic hit the economy in March 2020. Importantly, the ROC analysis makes explicit the associated type I and II probabilities of the binary forecast for a forthcoming recession. The possibility that NBER Dating Committee wouldn’t have announced the latest recession if the COVID-19-related lockdown had not occurred is consistent with the 25% false alarm.

It has been argued that using the spread alone may not be optimal in terms of recession forecasts. For example, Cooper et al. (2020) suggested using time-varying neutral federal funds rates for a better assessment of the monetary policy stance. As mentioned in the introduction, the approach by Iqbal et al. (IBS; 2019) is to compare the short-term rate with the lowest long-term rate during the rising period of the short-term rate. Their new definition decouples the effect of the current monetary policy stance on the long run rate, thereby capturing the stance of the current monetary policy better. Interestingly, a similar idea has been used by Crump et al. (2020) and Sahm (2019), where an effective unemployment rate is optimally defined as a deviation from a threshold based on past values. In the interest rate spread case, this approach is equivalent to creating a modified spread that is equal to the minimum long-term rate minus the current short-term rate. One can then check if the modified spread falls below zero. Although, as stated by the authors, a rising short-term rate environment is what is important for predicting recessions, we created the modified spread for declining periods also. We do this for the purpose of having a complete spread series over the whole period and being able to do evaluations using both true positives and false alarms. The same evaluation methodology of considering all thresholds was applied to the modified spread. Since these spread values are necessarily smaller than the original spread, the optimal threshold values were found to be closer to, but significantly larger than, zero and increased over time. Before the February 2020 peak the modified IBS maximized the Kuiper’s score at threshold 0.22 with associated hit and false alarm rates of 0.88 and 0.22, respectively, and were very close to the trade-off implied by the original spread (see Table 4, sample 6).3

Finally, with spread data till the end of September 2022, the optimal threshold has been breached again as of 07/11/2022 and a signal for another recession has crept up as the Fed started to increase its federal funds rate aggressively to combat inflation. This can also be seen in the generated recession probabilities from probit models; see Figs. 1, 3 and 4.

### 3 Robustness

In this section we show the robustness of our main conclusions with respect to two important modeling approaches used in the paper. First is the use of daily rather than monthly or quarterly data which has been the standard in the literature. Second is the use of a definition of the timing of peak turning points as happening any time during next 12 months. One may wonder if the zero threshold may still work with either longer or shorter horizons than 12 months.

We regenerated the results for the seven separate sub-samples reported in Table 4 using both monthly and quarterly observations. These results are reported in Tables 6 (monthly) and 7 (quarterly), respectively. In these tables we report the optimal threshold values that maximized the Kuipers score and also the corresponding hit and false alarm rates. The same statistics with assumed threshold value of zero are also reported. We see that the optimal threshold has increased to values close to one.4 With higher thresholds close to 1.0, the hit rates are close to 90%, and the Kuipers scores are uniformly higher, but with the caveat that the false alarm rate is slightly higher. These results are almost the same as those with daily data.

We also experimented with two alternatives where the horizon was set at 6 and 18 months. The optimal thresholds and corresponding statistics are reported in Tables 8 and 9. As expected, forecasting deteriorates as the horizon

---

3 These results are available from the authors on request.

4 With quarterly data in Table 7, although the optimal threshold is 0.30% for Sample 5, the KS is only slightly lower (0.63 rather 0.64) when a threshold of 0.94 is chosen as found in adjoining sub-samples 4, 6 and 7. This is because with quarterly data KS is rather flat between 0.3 and 1.0. At threshold of 0.94 the hit rate was much higher (H = 92%) than that with threshold 0.30% (H = 52%), but a higher F = 29%.
Table 6 Monthly data: 12-month-ahead

| Sample | Optimal threshold | AUC | Zero threshold |
|--------|-------------------|-----|---------------|
|        | Threshold | KS | H | F | KS | H | F |
| 1      | 0.14      | 0.55 | 0.67 | 0.12 | 0.84 | 0.36 | 0.46 | 0.09 |
| 2      | 0.14      | 0.64 | 0.78 | 0.13 | 0.87 | 0.44 | 0.64 | 0.11 |
| 3      | 0.14      | 0.69 | 0.79 | 0.10 | 0.90 | 0.51 | 0.67 | 0.07 |
| 4      | 0.49      | 0.61 | 0.78 | 0.17 | 0.89 | 0.42 | 0.53 | 0.05 |
| 5      | 0.49      | 0.63 | 0.81 | 0.17 | 0.90 | 0.46 | 0.51 | 0.06 |
| 6      | 1.11      | 0.67 | 0.96 | 0.29 | 0.91 | 0.40 | 0.49 | 0.04 |
| 7      | 0.65      | 0.68 | 0.86 | 0.18 | 0.92 | 0.39 | 0.48 | 0.04 |

See Table 4 for the definition of sub-samples

Table 7 Quarterly data: 12-month-ahead

| Sample | Optimal threshold | AUC | Zero threshold |
|--------|-------------------|-----|---------------|
|        | Threshold | KS | H | F | KS | H | F |
| 1      | 0.21      | 0.59 | 0.78 | 0.18 | 0.81 | 0.36 | 0.44 | 0.08 |
| 2      | 0.22      | 0.58 | 0.77 | 0.19 | 0.79 | 0.44 | 0.54 | 0.10 |
| 3      | 0.30      | 0.65 | 0.82 | 0.17 | 0.87 | 0.51 | 0.59 | 0.07 |
| 4      | 0.94      | 0.61 | 0.90 | 0.29 | 0.87 | 0.42 | 0.48 | 0.05 |
| 5      | 0.30      | 0.64 | 0.76 | 0.12 | 0.88 | 0.46 | 0.52 | 0.06 |
| 6      | 0.94      | 0.70 | 0.93 | 0.23 | 0.90 | 0.40 | 0.45 | 0.04 |
| 7      | 0.94      | 0.69 | 0.94 | 0.24 | 0.90 | 0.39 | 0.44 | 0.04 |

See Table 4 for the definition of sub-samples

Table 8 Optimal interest rate spread thresholds: 6-month-ahead

| Sample | Optimal threshold | AUC | Zero threshold |
|--------|-------------------|-----|---------------|
|        | Threshold | KS | H | F | KS | H | F |
| 1      | 0.06      | 0.79 | 0.93 | 0.14 | 0.93 | 0.80 | 0.89 | 0.13 |
| 2      | 0.04      | 0.79 | 0.95 | 0.16 | 0.93 | 0.83 | 0.92 | 0.15 |
| 3      | 0.06      | 0.82 | 0.93 | 0.12 | 0.95 | 0.84 | 0.91 | 0.11 |
| 4      | 0.06      | 0.66 | 0.75 | 0.09 | 0.90 | 0.68 | 0.73 | 0.08 |
| 5      | 0.06      | 0.63 | 0.72 | 0.10 | 0.90 | 0.65 | 0.70 | 0.09 |
| 6      | 1.02      | 0.64 | 0.96 | 0.32 | 0.90 | 0.56 | 0.61 | 0.07 |
| 7      | 1.02      | 0.65 | 0.97 | 0.32 | 0.90 | 0.53 | 0.57 | 0.07 |

See Table 4 for the definition of sub-samples

Table 9 Optimal interest rate spread thresholds: 18-month-ahead

| Sample | Optimal threshold | AUC | Zero threshold |
|--------|-------------------|-----|---------------|
|        | Threshold | KS | H | F | KS | H | F |
| 1      | 0.56      | 0.43 | 0.72 | 0.30 | 0.74 | 0.27 | 0.36 | 0.09 |
| 2      | 0.29      | 0.48 | 0.68 | 0.20 | 0.79 | 0.39 | 0.50 | 0.09 |
| 3      | 0.56      | 0.56 | 0.76 | 0.20 | 0.84 | 0.45 | 0.53 | 0.06 |
| 4      | 0.87      | 0.55 | 0.82 | 0.27 | 0.85 | 0.37 | 0.43 | 0.05 |
| 5      | 0.91      | 0.55 | 0.81 | 0.26 | 0.86 | 0.36 | 0.41 | 0.04 |
| 6      | 1.00      | 0.62 | 0.85 | 0.22 | 0.89 | 0.38 | 0.43 | 0.03 |
| 7      | 0.95      | 0.64 | 0.85 | 0.22 | 0.89 | 0.36 | 0.41 | 0.03 |

See Table 4 for the definition of sub-samples
increases from 6 to 18 months across all sub-samples. What is more important to notice is that the optimal thresholds have increased to around 1.0% at both horizons in recent years. Interestingly, at 6-month horizon, the optimal threshold was indeed very close to zero till about the 2008 recession. At this horizon, the deterioration in the forecasting power of yield spread is apparent very clearly after the end of sample 3, i.e., after 1990 when the Kuipers scores and the hit rates fell substantially. The discovery seems to be that the horizon at which yield spread is most effective predictor with the conventional threshold of zero became much longer than 6 months in later years. After the 1980s, the optimal threshold even at the 6-month horizon has increased to 1.02% in order to optimize on the Kuipers score. Further, no obvious deterioration of predictive power is found in terms of AUC. These results of high hit rates and Kuipers scores but at the cost of elevated false alarm rates with threshold values near one in recent recession episodes are very similar to what we found in our main results.

4 Conclusions

We forecast recessions using daily interest rate spreads directly, and focus on the threshold that can be used optimally to issue forecasts. No models such as probit or logit are used. By extending the sample gradually and recursively, we find that the optimal threshold of the daily spread has increased over time and is close to 0.9 in recent years. Our forecasts are evaluated more in terms of Kuipers score than on global measures like AUC or Concordance. Because a recession is a relatively rare event, the Concordance index is not found to be useful in our context. We also find evidence on the usefulness of working with daily spread data. Importantly, the mix of hit and false alarm rates over seven sub-samples show that the predictive power of the spread has not deteriorated in recent decades, provided the optimum values of the threshold are used. This may be a piece of useful information for practitioners who forecast recessions by directly monitoring the spread value in real time. Our results are robust even when we worked with monthly or quarterly data and with alternative forecast horizons.

Based on our optimal thresholds, we found evidence in favor of an upcoming recession long before the NBER announcement of the latest peak of February 2020. With data til the end of September 2022, another upcoming recession is signaled as the Fed has been increasing its federal funds rate aggressively to combat inflation. Although the binary forecasts are useful and often needed, ROC analysis allows practitioners to be aware of the associated trade-off between hit rates and false alarm rates. Forecast statements of this type make the job of forecasting binary events more credible because it quantifies the associated uncertainty in forecasting. In our analysis based on maximizing Kuipers score, the forecast for the current recession was issued a year ago with 91% hit rate and an associated false alarm rate of 24%. The ROC analysis allows a forecast user to choose from a combination of hit and false alarm rates constrained by the ROC frontier: one can choose a reduced false alarm rate, but only at the cost of reduced hit rate. For instance, if we had followed the conventional inverted yield threshold of 0, the

![Fig. 1 Optimal threshold and daily spread. (Color figure online)](image)
(a) Sample 1 (1/2/1962 to 1/31/1979)
(b) Sample 2 (1/2/1962 to 7/31/1980)
(c) Sample 3 (1/2/1962 to 7/31/1989)
(d) Sample 4 (1/2/1962 to 3/31/2000)
(e) Sample 5 (1/2/1962 to 12/29/2006)
(f) Sample 6 (1/2/1962 to 2/28/2019)
(g) Sample 7 (1/2/1962 to 9/20/2022)

Note: The red asterisk in each graph highlights the optimum threshold that maximizes $KS$ in that sample. They are 0.57 20 21 0.91 0.91 0.91 0.91 for Sample 1 to Sample 7. The diagrams include 95% confidence intervals computed using R package pROC.

Fig. 2 ROC curves with daily spread as threshold. (Color figure online)
recession forecast would have come with an underlying hit rate of 50% and a false alarm rate of 5%. Coincidentally, maximizing the concordance would have given the same trade-off. Since the conventional wisdom on the ground is to forecast recession when the yield spread inverts, the extraordinarily low hit rate of 50% may also explain why the professional forecasters have been hesitant to take the signals from the yield spread seriously—suggesting a possible resolution of the Rudebusch–Williams (2009) puzzle.
The number of observations is 2104 for days with recessions (red line) in the next 12 months and 13670 otherwise (black dashed line).

Fig. 5 Conditional densities for the spread. (Color figure online)

Acknowledgements We thank Ulrich Hounyo, Zhongwen Liang and the participants at 64th. Annual NABE Meeting in Chicago, 2021 Asian Meeting of the Econometric Society in Malaysia, and the 41st International Symposium on Forecasting for many helpful comments.

References

Bauer, Michael D., and Thomas M. Mertens. 2018. Information in the yield curve about future recessions. Federal Reserve Bank of San Francisco Economic Letter 20: 1–5.

Berge, Travis J., and Òscar. Jordà. 2011. Evaluating the classification accuracy of business cycles. International Symposium on Forecasting for many helpful comments.

Crump, Richard K., Domenico Giannone, and David O. Lucca. 2020. Reading the tea leaves of the U.S. business cycle-part one. Federal Reserve Bank of New York, Liberty Street Economics February 10. https://libertystreeteconomics.newyorkfed.org/2020/02/reading-the-tea-leaves-of-the-us-business-cycle-part-one/.

Dotsey, Michael 1998. The predictive content of the interest rate term spread for future economic growth. FRB Richmond Economic Quarterly 84 (3): 31–51.

Duarte, Augustin, Ioannis A. Venetis, and IvanPaya. 2005. Predicting real growth and the probability of recession in the euro area using the yield spread. International Journal of Forecasting 21 (2): 261–277.

Elliott, Graham, and Robert P. Lioi. 2013. Predicting binary outcomes. Journal of Econometrics 174 (1): 15–26.

Ergungor, O. Emre. 2016. Recession probabilities. Federal Reserve Bank of Cleveland, Economic Commentary 2016–09. https://doi.org/10.26509/frbc-ec-201609

Estrella, Arturo, and Gikas A. Hardouvelis. 1991. The term structure as a predictor of real economic activity. The Journal of Finance 46 (2): 555–576.

Estrella, Arturo and Frederic S. Mishkin. 1996. The yield curve as a predictor of U.S. recessions. Current Issues in Economics and Finance 2(7): 1–6.

Fendel, Ralf, Nicola Mai, and Oliver Mohr. 2021. Recessions probabilities for the eurozone at the zero lower bound: Challenges to the term spread and rise of alternatives. Journal of Forecasting 40(6): 1000–1026.

Ghysels, Eric, Jonathan B. Hill, and Kaiji Motegi. 2020. Testing a large set of zero restrictions in regression models, with an application to mixed frequency granger causality. Journal of Econometrics 218(2): 633–654.

Giacomini, Raffaella, and Barbara Rossi. 2006. How stable is the forecasting performance of the yield curve for output growth? Oxford Bulletin of Economics and Statistics 68: 783–795.

Granger, Clive W., and Mohammad H. Pesaran. 2000. Economic and statistical measures of forecast accuracy. Journal of Forecasting 19 (7): 537–560.

Harding, Donald, and Adrian Pagan. 2002. Dissecting the cycle: a methodological investigation. Journal of Monetary Economics 49 (2): 365–381. https://doi.org/10.1016/S0304-3932(01)00108-8.

Iqbal, Azhar, Sam Bullard, and John Silvia. 2019. Are yield-curve/monetary cycles’ approaches enough to predict recessions? Business Economics 54 (1): 61–68.

Jardet, Caroline. 2004. Why did the term structure of interest rates lose its predictive power? Economic Modelling 21 (3): 509–524.

Johansson, Peter, and Andrew C. Meldrum. 2018. Predicting recession probabilities using the slope of the yield curve. FEDs Notes:March 1.

Krzanowski, Wojtek J., and David J. Hand. 2009. ROC curves for continuous data. Boca Raton: CRC Press.

Lahiri, Kajal, and J George Wang. 2013. Evaluating probability forecasts for GDP declines using alternative methodologies. International Journal of Forecasting 29 (1): 175–190.

Lahiri, Kajal, and Liu Yang. 2013. Forecasting binary outcomes, Handbook of economic forecasting, Volume 2, 1025–1106. Elsevier.

Levanon, Gad, Jean-Claude Manini, Ataman Ozylidirim, Brian Schaitkin, and Jennelyn Tanchua. 2015. Using financial indicators to predict turning points in the business cycle: The case of the leading economic index for the united states. International Journal of Forecasting 31 (2): 426–445.

Liu, Weiling, and Emanuel Moench. 2016. What predicts US recessions? International Journal of Forecasting 32 (4): 1138–1150.

Manzato, Agostino. 2007. A note on the maximum Peirce skill score. Weather and Forecasting 22 (5): 1148–1154.

Miller, David S. 2019. There is no single best predictor of recessions. FEDs Notes: May 21.

Pažický, Martin. 2021. Predicting recessions in Germany using the German and the US yield curve. Journal of Business Cycle Research 17(3): 263–291.

Peirce, Charles S. 1884. The numerical measure of the success of predictions. Science 93: 453–454.

Rudebusch, Glenn D., and John C. Williams. 2009. Forecasting recessions: the puzzle of the enduring power of the yield curve. Journal of Business & Economic Statistics 27 (4): 492–503.

Sahm, Claudia. 2019. Direct stimulus payments to individuals. https://www.brookings.edu/research/direct-stimulus-payments-to-individuals/.

Stephenson, David B. 2000. Use of the “odds ratio” for diagnosing forecast skill. Weather and Forecasting 15 (2): 221–232.

Stock, James H., and Mark W. Watson. 1993. Introduction. In Business Cycles, Indicators, and Forecasting (James H. Stock, and Mark W. Watson, editors). Chicago: University of Chicago Press.
Wright, Jonathan H. 2006. *The yield curve and predicting recessions*. FEDS Discussion Paper 2006–07. https://www.federalreserve.gov/pubs/feds/2006/200607/200607pap.pdf.
Zhou, Xiao-Hua, Donna K, McClish, and Nancy A. Obuchowski. 2011. *Statistical Methods in Diagnostic Medicine*, Second Edition, Chapter 2. Inc: New York: Wiley.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.