The Devil in the Details: How Sensitive Are "Pockets of Predictability" to Methodological Choices?

*Nusret Cakici, Christian Fieberg, Tobias Neumaier, Thorsten Poddig, Adam Zaremba**

Abstract

The growing complexity of forecasting models increases the number of decision nodes in the research process, raising the risk of overfitting to specific design choices. We illustrate this issue using the recent concept of "pockets of predictability," which posits that return predictability is time-varying and that short windows of high predictability can be identified ex-ante. In this study, we reassess the robustness and practical applicability of this approach. By analyzing 19,440 variations of the original methodology, we find that its effectiveness depends critically on various seemingly minor methodological decisions. Furthermore, return predictability has declined significantly in recent decades, and the potential economic gains are highly sensitive to trading costs. Overall, strategies based on pockets of predictability should be approached with caution.

Keywords: return predictability, equity risk premium, pockets of predictability, replication, methodological uncertainty, research design, aggregate market returns.

JEL codes: G11, G12, G14, G17

This version: April 16, 2025

* Nusret Cakici is at the Gabelli School of Business, Fordham University. Christian Fieberg is at HSB Hochschule Bremen - City University of Applied Sciences, University of Luxembourg and Concordia University. Tobias Neumaier and Thorsten Poddig are at the University of Bremen. Adam Zaremba (corresponding author: a.zaremba@mbs-education.com or adam.zaremba@ue.poznan.pl) is at MBS School of Business, Poznan University of Economics and Business, and Monash University. Adam Zaremba acknowledges the support of the National Science Center of Poland [Grant No. 2022/45/B/HS4/00451].

1. Introduction

Forecasting the equity premium remains a central challenge in finance. As the literature continues to expand with an increasing array of explanatory variables (Goyal et al., 2024), researchers are continually developing and refining models to enhance accuracy. Unsurprisingly, the prediction methods have become increasingly complex over time. Advanced machine learning techniques (e.g., Rapach & Zhou, 2020; Dong et al., 2022; Masini et al., 2023; Stern, 2024) and extensive alternative datasets (e.g., Adämmer & Schüssler, 2020; Jiang et al., 2023; Hirshleifer et al., 2024) are now commonly employed, often yielding favorable results. As Kelly et al. (2022, 2024) have argued, sophisticated models tend to beat their simpler counterparts.

However, this sophistication comes with a trade-off. Complex models usually require numerous research design choices, many of which can be arbitrary. As a result, findings may be highly sensitive to seemingly minor implementation decisions, which, though initially appearing negligible, can significantly influence outcomes (Soebhag et al., 2024; Walter et al., 2024). Menkveld et al. (2024) coined the term "nonstandard errors" to describe these challenges. While robustness checks can address some of these issues, they are not always straightforward. Appropriate tests may not be immediately apparent, vary between studies, and often hinge on the specific methodology employed. Consequently, even ostensibly robust frameworks may be fundamentally shaped by subjective decisions made by researchers (e.g., Cakici et al., 2024b; Denk & Löffler, 2024).

To illustrate this concern, we explore the concept of "pockets of predictability." In their recent studies, Farmer et al. (2023, 2024) offer a fresh perspective on risk premia forecasting: they argue that stock markets experience prolonged periods of unpredictability punctuated by short windows of heightened predictability. They refer to them as "pockets of predictability." Crucially, these episodes can be identified ex-ante using kernel regressions, providing opportunities for substantial economic gains. This approach enables investors to generate alphas through more precise market timing.

The idea has attracted significant research interest, with the seminal work of Farmer et al. (2023) already amassing over 140 citations.[1] However, timing market returns is far from straightforward. The reliability of any method depends on its robustness under varying assumptions. A system that relies too heavily on specific methodological or parameter choices runs the risk of overfitting historical data. Since complex models involve multiple parameters, they are inherently prone to this issue, and seemingly impressive results may simply reflect statistical artifacts arising from over-optimization.

[1] As of 15 December 2024, based on Google Scholar.

Moreover, parameter validity is not carved in stone. Those that proved effective in the past may lose their relevance in the future. As a result, the observed predictability may diminish over time, undermining even the most promising trading strategies. Finally, transaction costs can erode performance, raising further concerns about a model's practical utility.

To revisit the utility of forecasting methods based on the pockets of predictability, we run a multiverse analysis in the spirit of Steegen et al. (2016). We aim to explore their robustness to underlying assumptions and parameters. Furthermore, we extend our analyses further to consider real-life impediments: changes in the return predictability over time and the influence of trading costs. Our analyses reveal essential weaknesses of pocket-based strategies.

First and foremost, the effectiveness of the pockets methodology is highly susceptible to various seemingly irrelevant research design choices. Farmer et al.'s (2024) approach requires many arbitrary decisions concerning pocket identification, return forecasting, and trading strategies. Whereas ostensibly secondary, they prove critical to the overall performance. In our tests, we identify nine such decision nodes and, by combining various design choices, generate 19,440 distinct implementations. As we demonstrate, many of them may disappoint.

While all considered research design choices influence the results to some extent, certain prove particularly impactful. For instance, the specifics of the pocket classification scheme play a crucial role. Farmer et al. (2023) project "squared error differentials" onto a time trend, whereas Farmer et al. (2024) omit this adjustment. Although the authors regard this choice as unimportant, it significantly affects performance, diminishing the practical benefits of return predictions. Specifically, accounting for the time trend can reduce abnormal returns by even 60%, slashing the annual alpha of their combination forecast strategy (*comb1*) from 4.51% to 1.65%.

Another critical detail lies in the shrinkage procedure used to smooth forecasts. Farmer et al.'s (2023, 2024) studies consider two approaches: one shrinks the forecasts toward a prevailing mean, while the other smooths them toward zero. This distinction is far from trivial. The former method proves considerably less effective, eliminating up to 70% of abnormal returns and trimming the alpha of the *comb1* strategy to just 1.05%.

Even seemingly minor adjustments, such as forecast winsorizing, can have substantial effects. For example, capping forecasts at the 2.5th percentile—as done in Farmer et al. (2024) but not in Farmer et al. (2023)—significantly enhances returns. Omitting this procedure reduces alphas by roughly half.

Overall, across the 19,440 possible implementations examined in our multiverse analysis, the pockets-based methodology consistently underwhelms. Most specifications fail to replicate the strong predictive performance reported by Farmer et al. (2024). Moreover, most implementations produce forecasts no more accurate than the prevailing mean benchmark at the 5% significance level, as measured by Clark and West's (2007) *t*-statistics (CW). This shortfall persists regardless of the choice of prediction variables or their aggregation method.

Besides the role of nonstandard errors, our extended analyses reveal two further limitations of the pockets framework—particularly relevant from a practical perspective. To begin with, the return predictability tends to decrease over time, and even pockets of predictability cannot overcome this trend. Consequently, their potential to identify and forge market patterns into measurable profits has shrunk over the last three decades. Before 1990, most pocket models generated reliable predictions, significantly beating the prevailing mean benchmark forecasts. However, from 1990 to 2016, the CW statistics values roughly halved, typically losing statistical significance. Also, in the broader context of all 19,440 simulations, the significant superiority of the prevailing mean forecast was far more prevalent than the reverse. In other words, the gains in prediction accuracy from applying this methodology can no longer be confirmed.

Moreover, the pockets-based strategies may be sensitive to transaction costs. While market timing typically relies on trading liquid and easily accessible securities, including futures and ETFs, they still incur inevitable implementation drag. To evaluate its impact, we reproduce our baseline allocation strategies using different levels of trading costs, from 0 to 20 basis points (bps). Apparently, their benefits decline quickly with the increasing fees—particularly in recent years. Consequently, at the level of 10 bps, no significant alphas can be recorded over the years 1990-2016. Furthermore, even at the level of 5 bps, the strategies' Sharpe ratios fail to exceed those of a passive buy-and-hold market portfolio. While this does not rule out the utility of pocket identification, it poses a practical hurdle that investors should be aware of.

To conclude, while the concept of pockets of predictability is intriguing, investors should handle it with care. The devil is in the details; even minor methodological decisions may matter. Issues such as declining predictability over time and transaction costs can prove detrimental to potential success.

Our study connects to several key areas of asset pricing research. First, our analyses are most closely related to the concept of pockets of predictability introduced by Farmer et al. (2023). A subsequent paper by Cakici et al. (2024) highlights a potential look-ahead bias due to a coding error, and Farmer et al. (2024) propose a smoothing mechanism to address this issue. However, our study differs significantly from that of Cakici et al.

(2024). While their work focuses on the coding error, which could be potentially corrected, we address broader issues within the pockets methodology that impact this entire line of research. Specifically, our analysis exposes weaknesses that cast serious doubt on the validity and rationality of the approach.

Notably, the concept of pockets of predictability has been rapidly adopted and has inspired further research. Andresen et al. (2023) offer a high-frequency analog to the lower-frequency episodes framework by Farmer et al. (2023). Borup et al. (2024) explore time-varying return predictability in bonds, while Li et al. (2023) extend the concept to cross-sectional factor pricing. Changing regimes of return predictability are also investigated by Harvey et al. (2021), Demetrescu et al. (2022), and Cong et al. (2024).

Second, our findings add to the discussion on the role of methodological uncertainty in asset pricing research. Menkveld et al. (2024) surveyed 164 research teams to demonstrate that even minor research design choices can influence overall conclusions in asset pricing research. Subsequent studies by Coqueret (2023), Soebhag et al. (2024), and Walter (2024) pinpoint the role of individual methodological decisions in stock market portfolio sorts. Fieberg et al. (2024) extend this discussion further to cryptocurrencies. Our work, in turn, is among the first to apply the same framework to time-series return predictability.

Third, our study contributes to the ongoing debate on p-hacking in predicting the equity premium. In recent years, a remarkable proliferation of methods and variables has been witnessed, allegedly predicting aggregate market returns. Not surprisingly, Welch and Goyal (2008) and Goyal et al. (2024) express skepticism about whether some are merely statistical artifacts. In line with this, Hollstein et al. (2024), whose massive work explores various predictors in 81 countries over up to 145 years of data, conclude that most standard variables and techniques generate disappointing out-of-sample results. Similarly, Dichtl et al. (2021), who examine a comprehensive battery of forecasting strategies, assert that most fail to beat naïve forecasts. Our findings echo these conclusions. Individual forecasting methods—especially those based on complex algorithms—may prove susceptible to data snooping and fail to cope well with real-life economic constraints. Consequently, practical implementations require particular caution from market investors.

The remainder of this paper proceeds as follows: Section 2 presents our data and explains the foundations of pockets-based methodology. Section 3 covers the empirical findings concerning the multiverse analysis. Section 4 considers the changes in return predictability over time and the impact of trading costs. Finally, Section 5 concludes.

2. Data and Methodology

2.1. Data and Variables

To ensure comparability with previous studies, we rely on Farmer et al.'s (2023, 2024) original sample and study period, strictly adhering to their data, methodology, and assessment metrics. Specifically, we use the data and code sourced from the original replication package labeled "Replication-code 20190881.zip," available on the Journal of Finance website.[2]

Our research focuses on the U.S. stock market. In almost all our tests, the predicted dependent variable is the excess aggregate market return, calculated as the CRSP U.S. stock market return minus the one-day return on a short T-bill rate. The predictor variables include the dividend-price ratio (*dp*), three-month T-bill rate (*tbl*), term spread (*tsp*), and realized variance (*rvar*). The study period runs from 1926 to 2016, as available for the different series.

To test the forecasting models—as in Farmer et al. (2024)—we use both the four individual predictors listed above (*dp*, *tbl*, *tsp*, and *rvar*), as well as five composite measures. The latter category includes: a recursively computed first principal component of the four individual variables (*pc*), a four-variable multivariate forecast estimated using a product kernel (*mv*), and three distinct averages of univariate forecasts, *comb1*, *comb2*, and *comb3*. *comb1* assigns the forecast from the time-varying coefficient model in-pocket but defaults to the prevailing mean out-of-pocket. It then takes the average from all models. *comb2* takes the average of all models indicating a pocket and the prevailing mean if no model indicates a pocket. *comb3* consistently utilizes the equal-weighted mean of the four standalone models, regardless of whether in-pocket or out-of-pocket. Further details of all the composite measures are available from Farmer et al. (2024).

2.2. Detecting the Pockets of Predictability

The seminal work of Farmer et al. (2023) assumes that market predictability varies over time, and the stages of high predictability—termed as "pockets of predictability"—can be identified ex-ante. To this end, they propose a framework based on kernel regressions. Following the original notation, the baseline prediction model can be described as:

$$r_{t+1} = x_t'\beta_t + \varepsilon_{t+1}, \tag{1}$$

where r_{t+1} denotes the excess stock market return at t+1, β_t represents time-varying regression coefficients, x_t is the vector of predictor variables, and ε_{t+1} is an unobservable

[2] See: https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.13229.

disturbance term. The model allows for general forms of conditional heteroskedasticity so that $\sigma_t^2 \equiv \mathbb{E}[\varepsilon_t^2|x_t] = \sigma^2(x_t)$. The parameters β_t are estimated using a local constant model of the following form:

$$\widehat{\beta_t} = \underset{\beta_0 \in R^d}{\arg\min} \sum_{s=1}^{T} K_{hT}(s-t)[r_{t+1} - x_s'\beta_0]^2, \tag{2}$$

where the weights on local observations are determined by the kernel $K_{hT}(u) \equiv K(u/hT)/(hT)$ with h denoting the bandwidth. Farmer et al. (2023, 2024) examine whether local predictability could have been identified with a one-sided version of the Epanechnikov kernel based on a 2.5-year bandwidth:

$$K(u) = \frac{3}{2}(1-u^2)1\{-1 < u < 0\}. \tag{3}$$

At this point, Farmer et al. (2024) introduce a Bayesian approach to shrink the kernel-based market forecasts, $\hat{r}_{t|t-1}$, towards zero and, ultimately, improve the stability of the model. Assume a regression of market excess returns on the forecast estimation,

$$r_{t+1} = \lambda_{1,t+1}\hat{r}_{t|t-1} + \varepsilon_{\mathrm{rt}}, \tag{4}$$

where $\lambda_{1,t+1}$ is estimated over a rolling window from t–s to t–1, with s indicating the window length. To reduce the estimation error, the model employs the G-prior Bayesian shrinkage from Zellner (1986), which uses data up to time t–1:

$$\hat{\lambda}_{1,t-1}^{\mathrm{G}} = \lambda_0 + \frac{1}{1+g}\left(\hat{\lambda}_{1,t-1} - \lambda_0\right), \tag{5}$$

where $\hat{\lambda}_{1,t-1}$ and λ_0 represent the estimates of $\lambda_{1,t-1}$ and the shrinkage target, respectively, and $\hat{\lambda}_{1,t-1}^{G}$ is the G-prior estimator for λ_1 using return information up to time t–1. In their baseline approach, Farmer et al. (2024) set $\lambda_0 = 0$ and $g = 2$. Furthermore, before estimating $\lambda_{1,t-1}$, they winsorize return forecasts at the 2.5% level, as well as restrict the initial rolling estimate of $\hat{\lambda}_{1,t-1}$ to the range between 0 and 1.

To determine the changes in predictive accuracy over time, Farmer et al. (2023, 2024) propose calculating the squared error difference, which compares the prediction errors on the prevailing mean benchmark forecast, $r_t - \bar{r}_{t|t-1}$, with the error on the examined kernel regression model based on the Bayesian adjustment, $r_t - \hat{r}_{t|t-1}^{G}$:

$$SED_t^G = \left(r_t - \bar{r}_{t|t-1}\right)^2 - \left(r_t - \hat{r}_{t|t-1}^{G}\right)^2. \tag{6}$$

Positive SED_t^G indicates that the kernel regression model performs better than the benchmark, producing more accurate predictions with reduced forecast errors.

Conversely, negative SED_t^G suggests the kernel model's forecasts are less precise and do not measure up to the benchmark's performance.

In the next step, the original Farmer et al. (2023) approach assumed projecting the SED values on a constant and a time trend to pinpoint the pockets of predictability:

$$\widehat{SED_t} = \hat{\gamma}_{0,t} + \hat{\gamma}_{1,t}t. \tag{7}$$

where the parameters $\hat{\gamma}_{0,t}$ and $\hat{\gamma}_{1,t}$ should be estimated with a one-sided Epanechnikov kernel with a one-year bandwidth. Farmer et al. (2024) revised this approach to drop the time-trend component and retain only the constant.

$$\widehat{SED_t^G} = \hat{\gamma}_{0,t}$$

Lastly, to avoid excessively short pockets, the pockets are defined as periods when SED_t^G is positive for at least 21 trading days (i.e., one month).

2.3. Implementation Variants

To scrutinize this issue closely, we reproduce the original methodology of Farmer et al. (2023, 2024) with a range of minor modifications. These design choices can be broadly categorized into three main deployment stages: forecasting (steps [1] to [6] below), pocket identification (steps [7] to [8]), and trading (step [9]). Below, we overview nine specific cases that have not been given significant attention in the former research. Notably, the last category, trading, is not concerned with forecasts *per se* but rather their transformation into a market timing strategy.

(1) Forecast winsorization: Before estimating the parameter $\hat{\lambda}_{1,t-1}$ in Eq. (1), Farmer. et al. (2024) winsorizes the return forecasts at the 2.5% level. Notably, this approach differs from Farmer et al. (2023), where no winsorization was applied. Return winsorization does not belong to the most common procedures in return forecasting, and there is no broadly accepted consensus considering the winsorization threshold. To examine the impact of this particular choice, we consider five alternative thresholds: 2%, 1.5%, 1%, 0.5%, and 0% (effectively no winsorization, as in Farmer et al., 2023).

(2) G-prior Bayesian shrinkage: Farmer et al. (2024) employ a G-prior Bayesian scheme, a mechanism representing an innovation relative to Farmer et al. (2023), which aims at smoothing the forecasts. The parameter g is set to 2, effectively putting one-third weight on the kernel forecasts and two-thirds weight on the shrinkage forecast. However, any value of the parameter g is largely arbitrary; hence, to explore its impact on the overall return predictability, we examine two additional levels: g = 1 and g = 3. Notably,

Farmer et al. (2024) indicate the robustness of the results with regard to the choice of g; however, no further details of these tests are disclosed.

(3) Minimum and (4) maximum weights: Another restriction in the forecast procedure concerns the weights restrictions to initial rolling window estimates $\hat{\lambda}_{1,t-1}$ in Eq. (5) in Farmer et al. (2024). The original methodology assumes a limit for its values between 0 and 1, citing „standard practice from forecast combinations." However, eventually, both numbers are set arbitrarily and their modifications may potentially affect the overall return predictability. For example, papers by Campbell and Thompson (2008) and Rapach et al. (2010) consider a maximum weight of 1.5, allowing for modest leverage. Moreover, short positions are not permitted, although they are relatively straightforward to implement for the broad U.S. stock market index. In line with this, even Farmer et al. (2024) in the earlier version of their note discussed a broader range of possible $\hat{\lambda}_{1,t-1}$ values, setting the minimum and maximum to -0.5 and 1.5. respectively. Therefore, to explore the impact of these limitations, we examine alternative minimum weights (-0.5 and -1) and maximum weights (1.5 and 2), which allow for moderate short selling and leverage.

(5) Smoothing benchmark: The pockets-based forecasting procedure in Eq. (5) involves shrinkage, which adjusts raw return forecasts by blending them with a benchmark to enhance stability and reduce noise. Specifically, the adjusted forecast is a weighted combination of the raw forecast and the benchmark, where the weight dynamically adjusts based on historical accuracy. If raw forecasts are noisy or unstable, the model assigns more weight to the benchmark, smoothing the predictions and improving predictability. Farmer et al. (2024) assume a zero benchmark ($\lambda_0 = 0$), which shrinks return forecasts toward zero and ensures maximal stability. This choice, consistent with the benchmark used by Gu et al. (2020), is somewhat arbitrary and effectively implies using no prior directional bias in forecasts. However, reasonable alternatives exist, such as the prevailing mean return forecast, which was employed by Farmer et al. (2023) and incorporates historical averages and captures persistent trends.[3] In this analysis, we compare both approaches—the zero benchmark and the prevailing mean forecast—to examine their impact on overall forecasting accuracy.

(6) Window length: The Bayesian shrinkage approach underlying the pockets-based forecasts involves regressing market returns on the one-sided kernel-based forecast. The estimation relies on the rolling window. However, what should be the specific length of this estimation window? This question has no simple answer, and each assumption may seem arbitrary. Farmer et al. (2024) set the window length to one year; however, no

[3] The approach involving the prevailing mean return forecast was actually also employed in one of the earlier versions of the Farmer et al. (2024) study.

specific justification is provided. To explore this issue, we consider the baseline approach as well as alternatives that are six months longer and shorter, i.e., 0.5, 1, and 1.5 years.

(7) SED identification: Having estimated the SED, their values are subsequently plugged into a pocket classification regression. Nevertheless, the specific approach in the works of Farmer et al. (2023) and (2024) differ at this point. Farmer et al. (2023) project them on a time trend, while Farmer et al. (2024) remove this adjustment. We consider both options to shed light on the importance of this methodological choice.

(8) Minimum pocket length: As Cakici et al. (2024) demonstrated, the actual length of the pockets may prove relatively short-lived. For example, this could pose practical consequences in terms of trading costs. To improve the model stability, Farmer et al. (2024) introduce a minimum pocket length of 21 days. The number seems largely arbitrary—therefore, we consider both longer and shorter alternatives: 42, 32, 21, 11, and 0 days (no minimum).

(9) Portfolio weight c: The practical application of the pockets identification strategy assumes forging them into a trading system. This pockets-based model assumes active market timing, allocating capital in the stock market based on return forecast (Section III.A in Farmer et al., 2024). The allocation is based on the time-varying parameter *c,* which scales the leverage based on historical return volatility. To estimate this parameter, Farmer et al. (2023) use the entire time series 1,...,T. In other words, they do not determine it out-of-sample, but at any time t utilize also future information, not available at the time of the forecast. The problem is of similar nature as that identified by Cakici et al. (2024), but of lesser consequence since it affects only a part of this analysis. Nevertheless, we consider both the in-sample c estimation and the out-of-sample approach to examine its impact on the results. The latter assumes that only data available up to day t are used.

Figure 1 illustrates all the research design choices, with arrows indicating alternative implementation „paths." Combining all possible designs in steps (1) through (9) results in 19,440 implementation variants. The original implementation from Farmer et al. (2024) is shown in pink boxes. By reproducing key results using all possible combinations and comparing them to the original, we can comprehensively assess the importance of various seemingly minor methodological decisions in "pockets of predictability" implementation.

[Insert Figure 1 here]

Notably, our selection of potential variants does not encompass the full range of methodological options. For example, kernel regression itself offers significant flexibility, including choices such as the type of function (e.g., Gaussian, box, or triangular kernels)

and the estimation bandwidth. Therefore, our results should be interpreted as a floor, representing the lower bound of possible outcomes, rather than a complete picture.

3. Baseline Findings

3.1. Statistical Performance

How much does the methodological discretion impact the overall results? To illustrate this, we begin by reporting the key measures of statistical performance across all 19,440 implementation variants. To begin with, we focus on reporting t-statistics from the classical tests of Clark and West (2007) (CW). They compare the predictions from a given model—in our case, the pockets methodology—with a prevailing mean benchmark forecast.

Figure 2 depicts the dispersion in potential results due to minor methodological changes, with Panels A through D focusing on individual forecasting models and Panels E to I on aggregate predictions. The body of each box represents the interquartile range, with its bottom and top marking the 25th and 75th percentiles of the CW statistic. We report the prediction performance for the total sample (FS) as well as for the periods classified as in-pocket (IP) and out-of-pocket (OOP). Furthermore, for comparison, we also plot the results of the original implementation by Farmer et al. (2024), as indicated by the red horizontal lines.

A quick overview of the results reveals several interesting patterns. First, the original implementation of Farmer et al. (2024) is exceptionally strong compared to all alternatives. For the FS and IP periods, the level of the CW test statistic is significantly higher than the vast majority of implementations—suggesting a superior prediction performance compared to alternative settings. For certain predictors, such as *mv*, *comb1*, *comb2*, and *comb3*, the original framework falls into the top percentile of our multiverse. On the other hand, in the OOP periods, the typical return predictability is noticeably stronger than documented by Farmer et al. (2024). In other words—depending on the particular implementation details—the disparity in return predictability between IP and OOP periods may be much weaker than commonly thought. The actual spread between these two market states hinges on minor methodological choices, and typically falls behind the pioneering results in Farmer et al. (2023, 2024).

[Insert Figure 2 here]

However, the second intriguing observation is how many of the implementation variants yield significant return predictability at all. Or, more specifically, how much of the in-pocket periods significantly beat the prevailing mean model, as per the CW t-statistics? Apparently, not that many. As seen from Figure 2, most implementations fail to exceed

the standard statistical threshold of 5%. Furthermore, the result distributions for the FS and IP periods do not differ much from each other, suggesting that the benefit of pinpointing the pockets of predictability may be limited.

The results thus far suggest a remarkable dispersion in results due to ostensibly negligible methodological choices. However, which of them matters the most for the results? Put differently, which methodological decisions strengthen (or weaken) the observed return predictability? To answer this question, we funnel all our 19,440 implementations through individual choices. In other words, considering Figure 1, we assume one branch as constant at each decision node and examine the results across all other decision nodes. Next, we calculate the CW *t*-statistics for the in-pocket periods across all the filtered implementations. This way, we observe the typical return predictability for one design choice across all other choices. Figure 3 reports the findings of this experiment, reporting the average values across all nine forecasting models considered. To illustrate the differences in results across the design choices, we present the median CW *t*-statistic.

[Insert Figure 3 here]

Clearly, specific designs prove more favorable than others, frequently overlapping with methodological choices in Farmer et al. (2024). The SED estimation may serve as one of the most striking examples. While the original study by Farmer et al. (2023) projected SED on the time trend, Farmer et al. (2024) abandoned this practice. As the authors argue, accounting for the time trend is „not important to our results" (Farmer et al., 2024, p. 7). However, while seemingly irrelevant, this issue matters a lot: the implementations without time trend typically have twice as high *t*-statistics as those with the time trend included. This substantial discrepancy not only showcases how seemingly minor methodological choices can materially affect the outcomes but also highlights the critical need for transparency in robustness analyses. As the field of finance grapples with the replication crisis, testing alternative specifications in supplementary materials, such as an online appendix, rather than merely mentioning them in a footnote, would enhance methodological rigor and credibility.

Other research design choices follow. Consider winsorizing. Eliminating the extreme returns—employed in Farmer et al. (2024) but not in Farmer et al. (2023) again increases the *t*-statistics by nearly 100%. Also, the minimum and maximum weight restrictions to the initial $\hat{\lambda}_{1,t-1}$ estimation play a role. Narrowing the range of possible values to between 0 and 1 generates better results than, for example, allowing for a modest leverage of 1.5, as suggested by Campbell and Thompson (2008) and Rapach et al. (2010). Lastly, shrinking the returns forecasts towards zero by setting the parameter $\lambda_0 = 0$ improves the prediction accuracy. Should Farmer et al. (2024) use some other shrinkage mechanism, such as aligning the forecasts with the prevailing mean (as done in one of the earlier versions of their study), the results would be worse.

To sum up, our findings indicate that the return predictability by the "pockets" approach depends critically on various implementation details. A range of auxiliary assumptions and parameter values, play a crucial role in obtaining solid results. If set favorably, they may create the impression that identifying pockets of predictability enables superior prediction performance. However, under the majority of alternative, equally reasonable assumptions, the results prove much worse and typically insignificant. This obviously raises concerns about the robustness and generalizability of pocket-based methodology. The ostensibly solid prediction performance may be due to a specific combination of implementation details rather than to the "pockets-of-predictability" phenomenon per se.

3.2. Economic Performance

In this section, we focus on the measures of economic performance of the trading strategies based on the pockets of predictability. To this end, we reproduce the asset allocation strategy from Farmer et al. (2023, 2024). This approach, referred therein also as the mean-variance optimized pocket portfolio, is constructed by dynamically rebalancing between stocks and T-bills based on real-time forecasts of expected excess returns, adjusted using a scaling factor that aligns portfolio variance with the forecasted return variance. To assess its payoffs, we compute two popular statistics also employed by Farmer et al. (2024): annualized Sharpe ratios and CAPM alphas. We run statistical tests on these measures for all 19,440 implementation variants from Section 3.1.1 to scrutinize their robustness in terms of investment performance. Figure 4 reports the results of this exercise, with Panels A and B concerning the distributions of alphas and Sharpe ratios, respectively.

[Insert Figure 4 here]

The outcomes corroborate our earlier observations. The relation between the original implementation in Farmer et al. (2024) and its alternative variants resembles that seen in Figure 2. In other words, the vast majority of the 19,440 specifications underperform the outcomes from the original paper. The Farmer et al. (2024) specification is in the very far tail, representing a vivid outlier rather than a common observation. This pattern consistently holds across all prediction models—for both alphas and Sharpe ratios.

To summarize, our findings thus far cast doubt on the practical value of the pockets-based predictions: once even a few seemingly negligible parameters change, profitability typically deteriorates.

3.3. Further Multiverse Analysis

To further evaluate the impact of different research decisions, we run two additional tests. First, we evaluate the impact of individual decision nodes with the Anderson-Darling test. Second, we look closer at the sensitivity of the baseline results in Farmer et al. (2024) to individual methodological changes.

3.3.1. Fork Sensitivity

In the first exercise, we follow Menkveld et al. (2024) and employ the Anderson-Darling (AD) to test the fork sensitivity across the 19,440 implementations. This robust method allows for a rigorous comparison of the distributions of multiple samples to determine whether they originate from the same underlying distribution. Specifically, the test calculates differences between the empirical cumulative distribution functions (ECDFs) of the samples. By applying it to subsets of our research designs, filtered by specific decisions at each fork, we formally assess the significance of their impact on performance. By transforming the statistic to asymptotically follow a standard normal distribution, the AD k-sample test facilitates direct comparison across various research designs. The magnitude of the resulting test statistics reflects the relative sensitivity of performance measures to particular decision nodes.

Figure 5 presents the results of this analysis. To provide a comprehensive overview, we report the outcomes for the CW statistics, alphas, and Sharpe ratios. The AD statistics are typically large, highlighting the substantial role played by the decision nodes. All reported scores correspond to p-values near zero, affirming the significance of each considered decision fork in shaping the results. However, the importance of these decisions varies. While the magnitude of the AD statistics differs across performance measures, certain decision nodes clearly stand out—most notably, the SED estimation model and the smoothing benchmark. Specifically, decisions on whether to include a time trend in SED estimation or how to shrink forecasts prove to be critically influential.

[Insert Figure 5 here]

Nevertheless—and we emphasize this—all decisions matter, as evidenced by the test statistics that exceed conventional significance thresholds. For instance, even for CW values, seemingly minor considerations, such as the winsorizing level, yield an AD statistic as high as 1891.9.

3.3.2. Impact of Individual Research Decisions

While our analysis in Section 3.3.1 focused on the importance of forks within the entire multiverse, we now shift our attention to the impact of specific research design choices.

More specifically, we aim to determine how individual deviations from the original methodology in Farmer et al. (2024) influence the eventual outcomes. To illustrate this, we use the example of the *comb1* combination forecasts, which aggregate the predictions of the four individual variables considered by Farmer et al. (2024): *dp*, *tbl*, *tsp*, and *rvar*. To capture the marginal impact of different decisions, we conducted 20 reproductions of the original implementation, each time modifying a single assumption at one fork. Figure 6 presents the results of this exercise. As before, we display the established set of performance measures: CW test statistics, alphas, and Sharpe ratios.

[Insert Figure 6 here]

To begin with, nearly all methodological shifts adversely affect the results. While the severity of the impact varies across specific decisions, the measures of statistical and economic performance consistently decline in the majority of specifications. Hardly any modification improves upon the original findings of Farmer et al. (2024).

When examining specific decisions, the conclusions from Figure 6 largely echo those from Section 3.3.1. Two decisions stand out: i) the inclusion of a time trend in SED estimation and ii) the use of prevailing mean returns to smooth the forecasts. Both decisions result in a significant decline in all performance measures, rendering any statistical or economic gains negligible. For instance, including the time trend in SED estimation alone reduces abnormal returns from 4.51% to 1.65%, a drop of 2.86 percentage points. The impact of smoothing the return forecast toward the prevailing mean is even greater, reducing alphas to 1.05%, a decline of 3.46 percentage points.

The impact of other decisions, while more nuanced, still affects the results. For example, abandoning the practice of forecast winsorization—used in Farmer et al. (2024) but absent in Farmer (2023)—is sufficient to lower the CW test statistics and erase more than one-third of abnormal returns. The role of setting the minimum pocket length is also notable. Farmer et al. (2024) impose this restriction to "avoid ultra-short pockets with no economic meaning" (p. 7). Abandoning this restriction entirely renders the CW test statistics insignificant. However, its impact on Sharpe ratios and alphas is more modest, with the latter declining by only 0.3 percentage points per year.

To conclude, our analysis in this section reiterates the key role of individual research decisions, particularly choices such as smoothing methods, time trend inclusion, and return winsorization. This reinforces the call for robustness checks and transparency in finance research to ensure that conclusions remain both reliable and replicable.

4. Further Insights

In this section, we provide further insights into robustness of the pockets-based forecasts in relation to two major factors. First, we explore the changes in performance over time. Second, we investigate the impact of trading costs.

4.1. Performance through Time

Asset pricing research often indicates a decline in return predictability over time as equity markets mature and become more efficient. This phenomenon affects both cross-sectional return predictability (Chordia et al., 2014; McLean & Pontiff, 2016; Linnainmaa & Roberts, 2018) and the time-series patterns in aggregate returns (e.g., Jones & Pomorski, 2017; Baltussen et al., 2019). Can the pockets-based methodology escape this trend? For practical investors, it is crucial not only whether pockets helped forecast returns in the 1970s or 1980s but also whether they have survived the test of time, including in the most recent decades.

Figure 7 presents the application of pockets methodology within subperiods. To offer a comprehensive picture, we report median performance measures funneled over different decision nodes. Furthermore, we focus on the measures of economic performance—alphas and Sharpe ratios—as comparisons of CW statistics may be affected by the relative length of subperiods length. We arbitrarily select 1990 as the cutoff date and reexamine return predictability before this date and during the more recent period from 1990 to 2016.[4] The differences between these two subperiods are remarkable.

[Insert Figure 7 here]

The discrepancy between the two subperiods is particularly remarkable for the alphas (Panels A.1, A.2). While the years 1926 to 1989 witnessed positive—though uneven—alphas across all the implementation nodes, since 1990, almost all of them have disappeared. The typical alphas are close to zero, signaling no abnormal returns generated from market timing. Systematic deviations are rare, occur only for certain forks, and are equally likely to be positive or negative. The most striking example is perhaps the choice of the smoothing benchmark, disregarded as ostensibly irrelevant by Farmer et al. (2024). Yet, specifications shrinking the forecasts towards the prevailing mean commonly generate negative alphas, while replacing it with zero shrinkage vividly boosts market timing gains in the post-1989 period. In fact, while the performance differences for other decision nodes are less striking, the design choices in Farmer et al.

[4] Consistent with all other calculations, the prevailing mean benchmark forecast for the recent years 1990-2016 in Table 3, Panel C, is based on the return time series starting in 1926. Shifting the starting date for the calculation of the prevailing mean to 1990 has no material impact on the results.

(2024) prove generally favorable, including in cases of winsorizing levels, minimum weights, or pocket length.

Table 1, Panels B, illustrates an analogous analysis of Sharpe ratios. In this case, the differences over time are less evident, as the considered strategies are long-only and generally capitalize on the positive equity premium in both subperiods. Nevertheless, a decline in performance is still visible. Before 1990, the median Sharpe ratios generally oscillate within the range of 0.4 to 0.5. Since 1990, however, this figure shrinks to about 0.35–0.4. Again, the research design choices in Farmer et al. (2024) turn out to be particularly favorable for performance in recent years.

To provide additional context, it is worth noting that the market portfolio's Sharpe ratio is 0.38 over the full study period. However, its value changed over time, being slightly lower in the pre-1990 period (0.28) and increasing in the more recent post-1989 years (0.43). This suggests that the pocket-based timing strategies tended to outperform the market before 1990 but underperformed afterward. Hence, the value added by pursuing these strategies over the last three decades appears, at best, debatable.

To sum up, return predictability stems mainly from the early part of the sample. In recent years—perhaps particularly relevant from an investor's perspective—predictability hardly exists. Consequently, the pockets methodology struggles to add any value, casting doubt on its utility for market participants.

4.2. Trading Costs

Trading stocks incurs costs. In consequence, many return patterns and trading strategies that seem robust on paper eventually disappoint (Chordia et al., 2014; Novy-Marx & Velikov, 2016; Chen & Velikov, 2023). This may also apply to models considered by Farmer et al. (2024). Any timing strategy aiming to beat the market must overcome the impact of transaction costs. Do the pockets of predictability survive this hurdle?

The market timing strategy based on the pocket predictions, as described by Farmer et al. (2023), assumes systematic revisions of market exposure—including leveraged positions—and active allocation between equities and T-bills. While turnover may be relatively high, the ultimate transaction costs depend on the implementation framework. The costliest approach would involve buying and selling actual equities. Although trading would concentrate mainly on big and liquid stocks, the total trading costs, encompassing both commissions and implementation shortfall, may exceed 30 basis points (e.g., Virtu Financial, 2022). A cheaper variant would involve using exchange-traded funds (ETFs) or futures contracts. The bid-ask spreads in both cases are below 0.5 basis points (see, e.g., CME Group, 2016), and typical commissions are around two basis points (with certain brokers offering commission-free trades). The implementation shortfall is unlikely

to exceed one or two basis points, even for large trades. Overall, the total execution costs should be below five basis points.

This solution comes with at least two limitations. First, due to basis movements and temporary price fluctuations, both of these investment vehicles—futures and ETFs—may fail to perfectly track the index. Since pockets-based strategies assume daily rebalancing, even small discrepancies could matter. Second, while execution costs are limited, investors would still face holding costs. In particular, the annual net expense ratios for the most popular S&P ETFs range from 3 to 15 basis points. Trading index futures does not entail management fees, but investors still need to roll the contracts regularly, potentially causing a similar implementation drag.

To explore the practical gains from pockets-based predictions, we reproduce the original market timing strategies proposed in Farmer et al. (2024) with different levels of trading costs, ranging from 0 to 20 basis points. Subsequently, we compute the cost-adjusted alphas and Sharpe ratios to pinpoint the net value added for investors. Notably, for the clarity of presentation, we focus solely on the original specification rather than on the entire implementation multiverse. Table 1 reports the results of this exercise.

[Insert Table 1 here]

First, observe Panel A, which presents the alphas. Nearly all strategies generate positive and significant alphas when no trading costs are assumed. This is true not only for the whole study period but also for both subperiods—the pockets-based methodology yields substantial abnormal returns even in the more recent years, from 1990 to 2016. This relatively robust performance in recent decades, especially compared to the results in Section 3.2, is due to the particularly favorable research design choices in Farmer et al. (2024). However, introducing even modest transaction costs changes the situation. For example, when they equal five basis points, the alphas remain essentially statistically significant for the full study period, but this is mainly due to the early years with stronger return predictability. In the post-1989 period, none of the alphas are significant, though their nominal magnitude is still sizable.

Next, when the assumed costs increase to 10 basis points per trade, even in the early years, the alphas become predominantly insignificant, with the raw magnitude of alphas dropping by roughly 80%. Strikingly, their values fall practically to zero, with half of the strategies generating negative values. Naturally, increasing the costs further results in a stronger drag on investment performance. At 20 basis points, all strategies produce substantially negative abnormal returns in the years 1990–2016, with half of these values significantly below zero.

Panel B of Table 1 focuses on Sharpe ratios. To assess the materiality of gains from a given strategy, we examine whether the Sharpe ratio significantly exceeds (or falls below) the Sharpe ratio of a market portfolio, corresponding to a buy-and-hold strategy without any leverage or short selling. Using the Ledoit and Wolf (2016) statistic, we compare the performance of passive and active strategies. To reiterate, the market's portfolio Sharpe ratio was 0.38 for the full study period 1926-2016, as well as 0.28 and 0.43 for the earlier and later periods, respectively.

Basically, the pockets-based timing strategy competes moderately well against the passive market portfolio approach. With no trading costs assumed, the strategies' Sharpe ratios significantly surpass that of the market portfolio in six out of nine cases—though mainly only at the 10% level (in the one-tailed test framework). The situation deteriorates noticeably at the five and ten basis points cost levels. In a nutshell, the pockets-based strategies no longer beat the market portfolio. The Sharpe ratios typically fail to exceed those of passive buy-and-hold strategies, and this pattern extends to both early and recent years. At higher cost levels, performance worsens further. For instance, when transaction costs reach 20 basis points per trade, all strategies significantly underperform the market portfolio in the most recent part of the sample.

To sum up, while implementation costs do not necessarily eliminate profitability, they represent a substantial drag on abnormal returns. Consequently, the pockets-based strategies could potentially be applied in practice, but with at least two soft spots. First, they should use liquid index trackers like ETFs or futures. Second, investors must accept a significant loss of abnormal returns.

5. Conclusion

Predicting the equity premium remains a significant challenge in finance, with increasingly sophisticated models and diverse datasets driving progress in forecasting accuracy. While these complex methods often outperform simpler alternatives, they come at a cost: reliance on numerous subjective design decisions makes these models vulnerable to methodological changes, raising concerns about their stability and reliability. This sensitivity to seemingly minor choices poses challenges for ensuring robustness, with researchers referring to such uncertainties as "nonstandard errors."

To illustrate these issues, we examine the concept of "pockets of predictability" introduced by Farmer et al. (2023, 2024). These authors propose that the magnitude of return predictability fluctuates over time, with long periods of unpredictability punctuated by brief windows of forecastable returns. Crucially, these pockets can be identified ex-ante using kernel regressions, enabling investors to time the market effectively and achieve abnormal returns.

Our analyses reveal three major weaknesses of this approach. First, pocket forecasts are highly sensitive to minor methodological choices, such as return winsorizing or window lengths. Small modifications in research design frequently have a detrimental impact on forecasting accuracy. When subject to seemingly irrelevant changes, most implementations may prove disappointing. Second, return predictability declines over time. Consequently, the CW values over the last three decades are commonly low and insignificant. Third, potential profits are sensitive to trading costs. Even at the level of 10 basis points, transaction cost drag may effectively erase all significant gains from the pockets-based methodology.

All in all, while our results do not entirely reject the concept of pockets of predictability, they highlight a series of considerable weaknesses. This forecasting method comes with caveats, and investors should be aware of them when considering its implementation in practice.

References

Adämmer, Philipp, & Schüssler, Reiner A. (2020). Forecasting the equity premium: mind the news! *Review of Finance*, 24(6), 1313-1355.

Andersen, Torben G., Yingying Li, Viktor Todorov, & Bo Zhou. (2023). Volatility measurement with pockets of extreme return persistence. *Journal of Econometrics*, *237*(2), 105048.

Baltussen, Guido, Sjoerdvan Bekkum, & Zhi Da. (2019). Indexing and stock market serial dependence around the world. *Journal of Financial Economics*, *132*(1), 26-48.

Borup, Daniel, Jonas N. Eriksen, Mads M. Kjær, Martin Thyrsgaard. (2024). Predicting bond return predictability. *Management Science*, *70*(2), 931-951.

Cakici, Nusret, Christian Fieberg, Daniel Metko, & Adam Zaremba. (2024b). Do anomalies really predict market returns? New data and new evidence. *Review of Finance*, 28(1), 1-44.

Cakici, Nusret, Christian Fieberg, Tobias Neumaier, Thorsten, Poddig, & Adam Zaremba. (2024). Pockets of predictability: A replication. *Journal of Finance*, forthcoming.

Chen, Andrew Y., & Mihail Velikov. (2023). Zeroing in on the expected returns of anomalies. *Journal of Financial and Quantitative Analysis*, 58(3), 968-1004.

Chordia, Tarun, Avanidhar Subrahmanyam, & Qing Tong. (2014). Have capital market anomalies attenuated in the recent era of high liquidity and trading activity? *Journal of Accounting and Economics*, 58(1), 41-58.

Clark, Todd E., & Kenneth D. West. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1), 291-311.

CME Group. (2016). REPORT: The Big Picture: A Cost Comparison of Futures and ETFs. Available at https://www.cmegroup.com/trading/equity-index/report-a-cost-comparison-of-futures-and-etfs.html.

Cong, Lin William, Jingyu He, & Yuanzhi Wang. (2024). Mosaics of predictability. *Available at SSRN 4853767*.

Coqueret, Guillaume. (2023). Forking paths in empirical studies. *Available at SSRN 3999379*.

Demetrescu, Matei, Iliyan Georgiev, Paulo M.M. Rodrigues, A.M. Robert Taylor. (2022). Testing for episodic predictability in stock returns. *Journal of Econometrics*, *227*(1), 85–113.

Denk, Sebastian, Gunter Löffler. (2024). Predicting the equity premium with combination forecasts: A reappraisal. *Review of Asset Pricing Studies*, raae009.

Dichtl, Hubert, Wolfgang Drobetz, Andreas Neuhierl, & Viktoria-Sophie Wendt. (2021). Data snooping in equity premium prediction. *International Journal of Forecasting*, *37*(1), 72-94.

Dong, Xi, Yan Li, David E. Rapach, & Guofu Zhou. (2022). Anomalies and the expected market return. *The Journal of Finance*, 77(1), 639-681.

Farmer, Leland E., Lawrence Schmidt, Allan Timmermann. (2023). Pockets of predictability. *Journal of Finance*, 78(3), 1279-1341.

Farmer, Leland E., Lawrence Schmidt, Allan Timmermann. (2024). Comment on Cakici, Fieberg, Neumaier, Poddig, and Zaremba: Pockets of predictability: A replication. Available at SSRN: 4717259.

Fieberg, Christian, Steffen Günther, Thorsten Poddig, Adam Zaremba. (2024). Non-standard errors in the cryptocurrency world. *International Review of Financial Analysis*, *92*, 103106.

Goyal, Amit, Ivo Welch, Athanasse Zafirov. (2024). A comprehensive 2022 look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 37(11), 3490-3557.

Harvey, David I., Stephen J. Leybourne, Robert Sollis, A.M. Robert Taylor. (2021). Real-time detection of regimes of predictability in the US equity premium. *Journal of Applied Econometrics*, *36*(1), 45–70.

Hirshleifer, David, Dat Mai, Kuntara Pukthuanthong. (2024). War discourse and disaster premium: 160 years of evidence from the stock market. *The Review of Financial Studies*, in press.

Hollstein, Fabian, Marcel Prokopczuk, Björn Tharann, Chardin Wese Simen. (2024). Predicting the equity premium around the globe: Comprehensive evidence from a large sample. *International Journal of Forecasting*, *41*(1), 208-228.

Jiang, Jingwen, Bryan Kelly, Dacheng Xiu (2023). (Re-) Imag (in) ing price trends. *The Journal of Finance*, 78(6), 3193-3249.

Jones, Christopher S., Lukasz Pomorski. (2017). Investing in disappearing anomalies. *Review of Finance*, *21*(1), 237-267.

Kelly, Bryan T., Semyon Malamud. & Kangying Zhou. (2022). The virtue of complexity everywhere. Available at SSRN 4166368.

Kelly, Bryan T., Semyon Malamud. & Kangying Zhou. (2024). The virtue of complexity in return prediction. *The Journal of Finance*, 79(1), 459-503.

Ledoit, Oliver, & Wolf, Michael. (2008). Robust performance hypothesis testing with the Sharpe ratio. *Journal of Empirical Finance*, *15*(5), 850-859.

Li, Sophia Zhengzi, Peixuan Yuan, & Guofu Zhou. (2023). Pockets of factor pricing. Available at SSRN 4661444

Linnainmaa, Juhani T., Michael R. Roberts. (2018). The history of the cross-section of stock returns. *Review of Financial Studies*, *31*(7), 2606-2649.

Löffler, Gunter. (2022). Equity premium forecasts tend to perform worse against a buy-and-hold benchmark. *Critical Finance Review*, *11*(1), 65–77.

Masini, Ricardo P., Marcelo C. Medeiros, & Eduardo F. Mendes. (2023). Machine learning advances for time series forecasting. *Journal of Economic Surveys*, 37(1), 76-111.

McLean, R. David, & Pontiff, Jeffrey. (2016). Does academic research destroy stock return predictability? *Journal of Finance*, *71*(1), 5–32.

Menkveld, Albert J., Anna Dreber, Felix Holzmeister, et al. (2024). Nonstandard errors. *Journal of Finance*, 79(3), 2339-2390.

Newey, Whitney K., & Kenneth D. West. (1987). A simple positive definite, heteroscedasticity, and autocorrelation consistent covariance matrix, *Econometrica*, 55(3), 703–708,

Novy-Marx, Robert, & Velikov, Mihail. (2016). A taxonomy of anomalies and their trading costs. *Review of Financial Studies*, 29(1), 104-147.

Rapach, David E., & Guofu Zhou. (2020). Time-series and cross-sectional stock return forecasting: New machine learning methods. In: *Machine learning for asset management: New developments and financial applications*, 1-33.

Soebhag, Amar, Bart Van Vliet, & Patrick Verwijmeren. (2024). Non-standard errors in asset pricing: Mind your sorts. *Journal of Empirical Finance*, 101517.

Steegen, Sara, Francis Tuerlinckx, Andrew Gelman, & Wolf Vanpaemel. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5), 702-712.

Stein, Tobias. (2024). Forecasting the equity premium with frequency-decomposed technical indicators. *International Journal of Forecasting*, 40(1), 6-28.

Virtu Financial (2024). Global Cost Review. Q4 2021. Available from www.virtu.com.

Walter, Dominik, Rüdiger Weber, & Patrick Weiss. (2024). Methodological uncertainty in portfolio sorts. *Available at SSRN 4164117*.

Welch, Ivo, & Amit Goyal. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, *21*(4), 1455-1508.

Zellner, Arnold. (1986). On assessing prior distributions and Bayesian regression analysis with G-prior distributions. *Bayesian Inference and Decision Techniques*.

Figure 1. Research Design Choices for Pockets Identification

The figure shows the decision nodes considered in the robustness analysis. We consider the paths through nine different decision nodes to obtain the prediction performance. In total, all combinations generate 19,440 implementation variants. The pink boxes represent the original decisions made by Farmer et al. (2024).

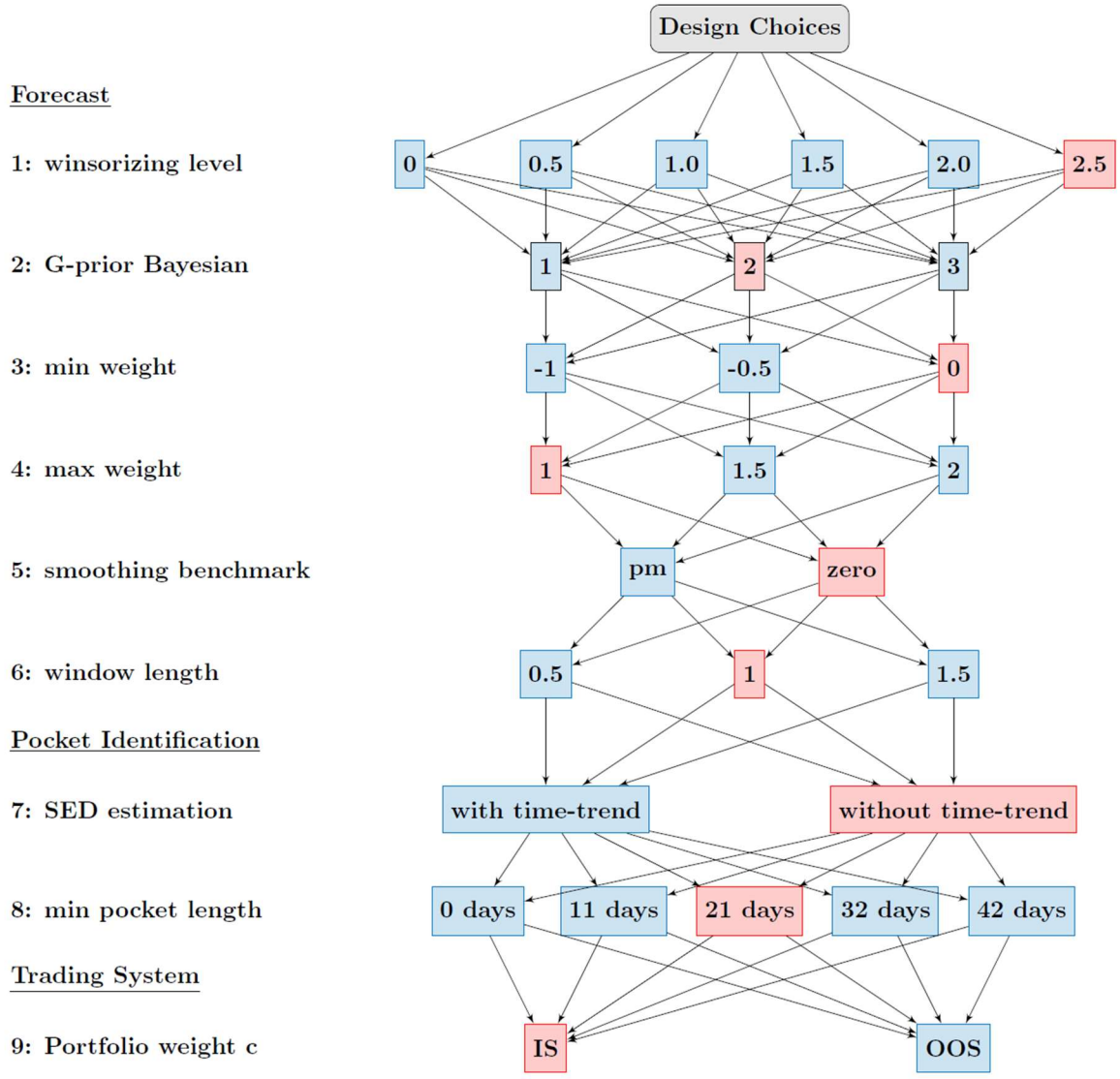


Figure 2. Out-of-Sample Measures of Forecasting Performance – Clark-West Statistics

The figure reports the distribution of 19,440 replications of the results from Farmer et al. (2024), Table III, Panel A. Specifically, it reports the Clark and West (2007) test statistics for out-of-sample return predictability measured relative to a prevailing mean forecast. Positive CW values indicate that return forecasts outperform the prevailing mean forecast, while negative values indicate the opposite and the values of 1.64 (2.33) indicate statistical significance in a one-tailed test at the 5% (1%) level. The body of each box represents the interquartile range, with its bottom and top marking the 25th and 75th percentiles of the *t*-statistics distribution. The horizontal line in the center of a box is the median. The red bar indicates the original specification of Farmer et al. (2024). The dots in the figure represent outliers, defined as values that fall more than 1.5 times the interquartile range above the upper limit or below the lower limit of the box. The results are reported separately for the full sample (FS), the in-pocket (IP), and the out-of-pocket (OOP) periods. Panel A to I report the results for nine prediction models: dividend price ratio (*dp*), treasury bill rate (*tbl*), term spread (*tsp*), and realized variance (*rvar*), and five composite models: principal components (*pc*), four-variable multivariate (*mv*) forecast, and three variants of combination forecasts (*comb1*, *comb2*, *comb3*). The calculations are based on the original data from Farmer et al. (2023) and the modifications of its code available from the replication package.

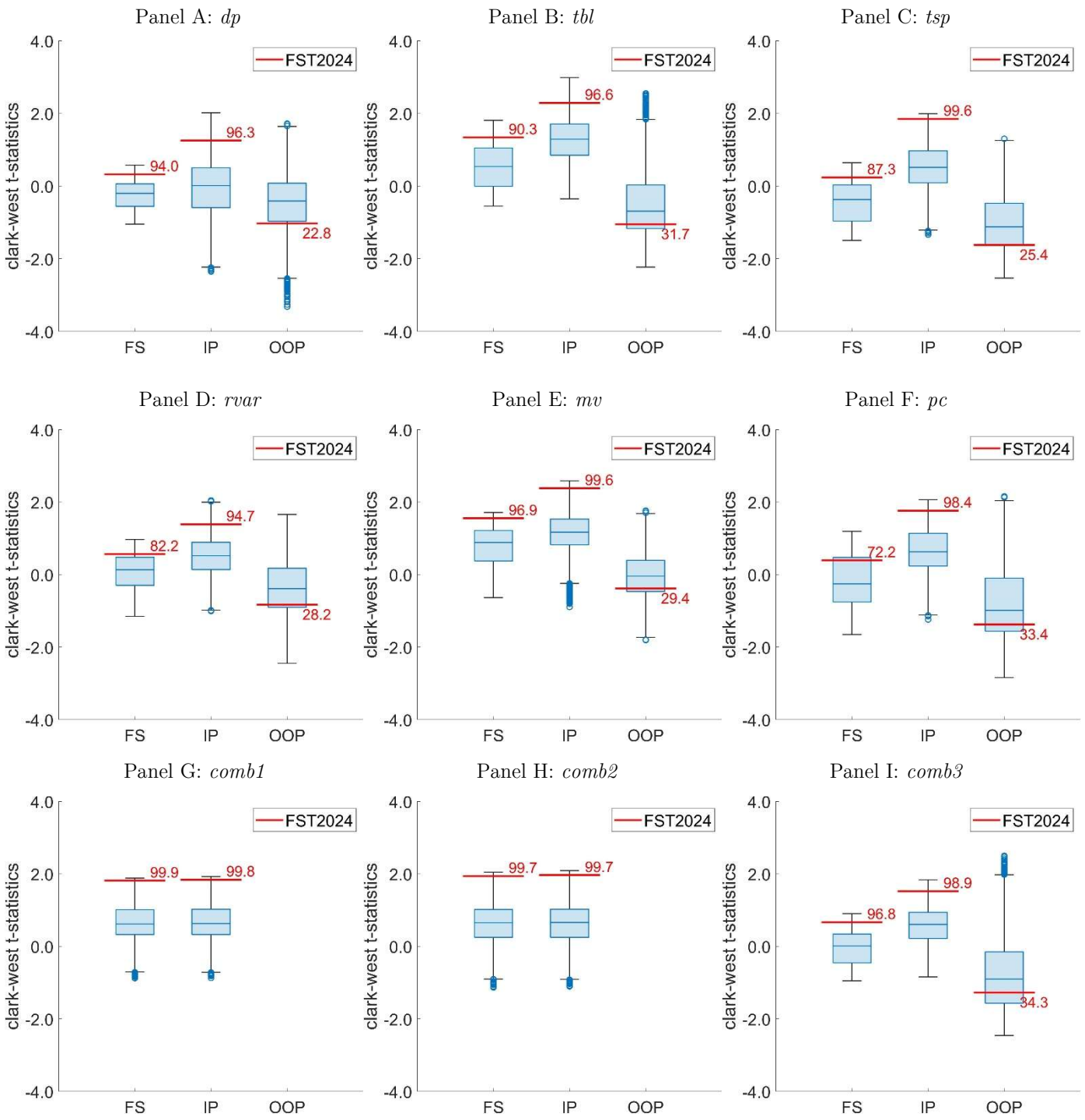


Figure 3. Prediction Efficiency Across Research Design Choices

The table reports the model prediction efficiency measured with the median Clark and West (2007) (CW) test statistic for the in-pocket periods across the distributions of 19,440 replications funneled over individual research design choices. The CW tests compare the out-of-sample return predictability relative to a prevailing mean forecast. Positive CW values indicate that return forecasts outperform the prevailing mean forecast, while negative values indicate the opposite. The figures are based on a pooled set of forecasts from all prediction models from Farmer et al. (2023): dividend price ratio (dp), treasury bill rate (tbl), term spread (tsp), and realized variance (rvar), and five composite models: principal components (pc), four-variable multivariate (mv) forecast, and three variants of combination forecasts (comb1, comb2, comb3). The red bars indicate the original research design choice from Farmer et al. (2024). The calculations are based on the original data from Farmer et al. (2023), and the study period runs from 1926 to 2016, as available for different predictors.

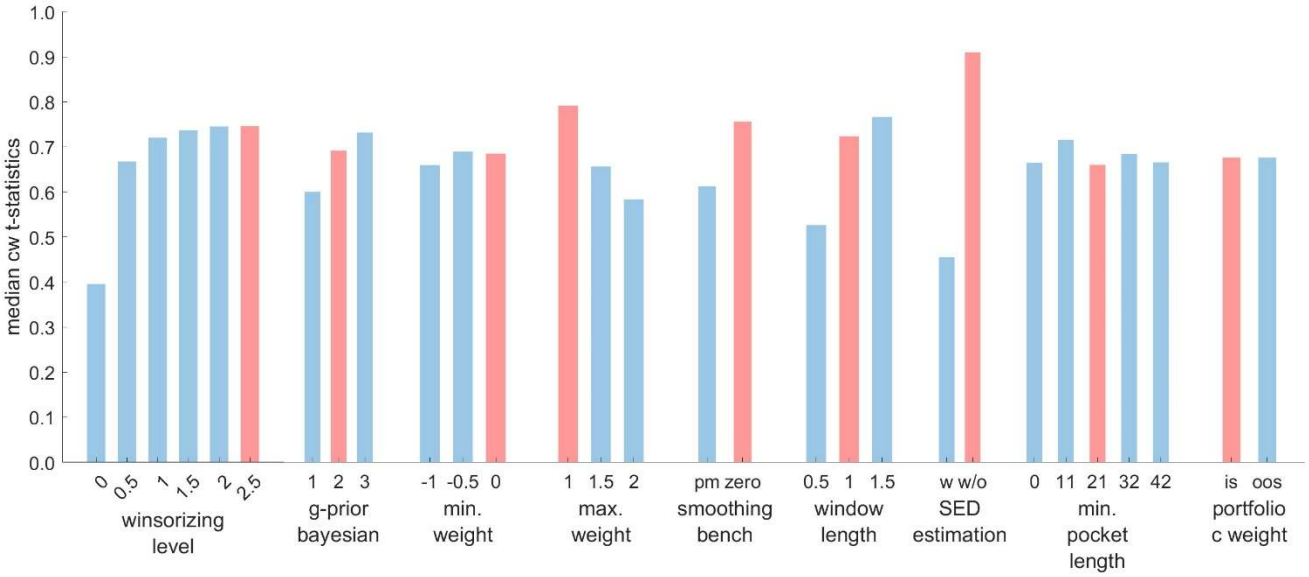
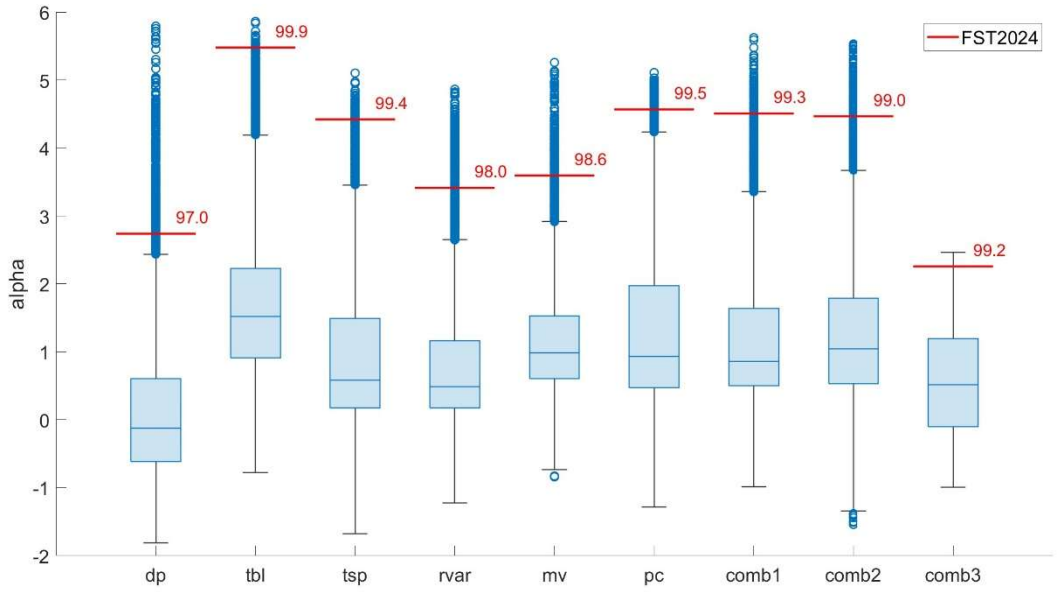


Figure 4. Out-of-Sample Measures of Economic Performance

The figure illustrates the economic gains from the standard forecasting model in Farmer et al. (2024) by reporting two performance measures: the alphas (Panel A) and the annualized Sharpe ratios (Panel B). The candlesticks represent the distribution of 19,440 implementation variants. The body of each box represents the interquartile range, with its bottom and top marking the 25th and 75th percentiles of the return distribution. The horizontal line in the middle of a box is the median. The red bar indicates the original specification of Farmer et al. (2024). The dots in the figure represent outliers, defined as values that fall more than 1.5 times the interquartile range above the upper limit or below the lower limit of the box. The considered models are dividend price ratio (*dp*), treasury bill rate (*tbl*), term spread (*tsp*), and realized variance (*rvar*), and five composite models: principal components (pc), four-variable multivariate (*mv*) forecast, and three variants of combination forecasts (*comb1*, *comb2*, *comb3*). The calculations are based on the original data from Farmer et al. (2023), and the study period runs from 1926 to 2016, as available for different predictors.

Panel A: Alphas



Panel B: Sharpe Ratios

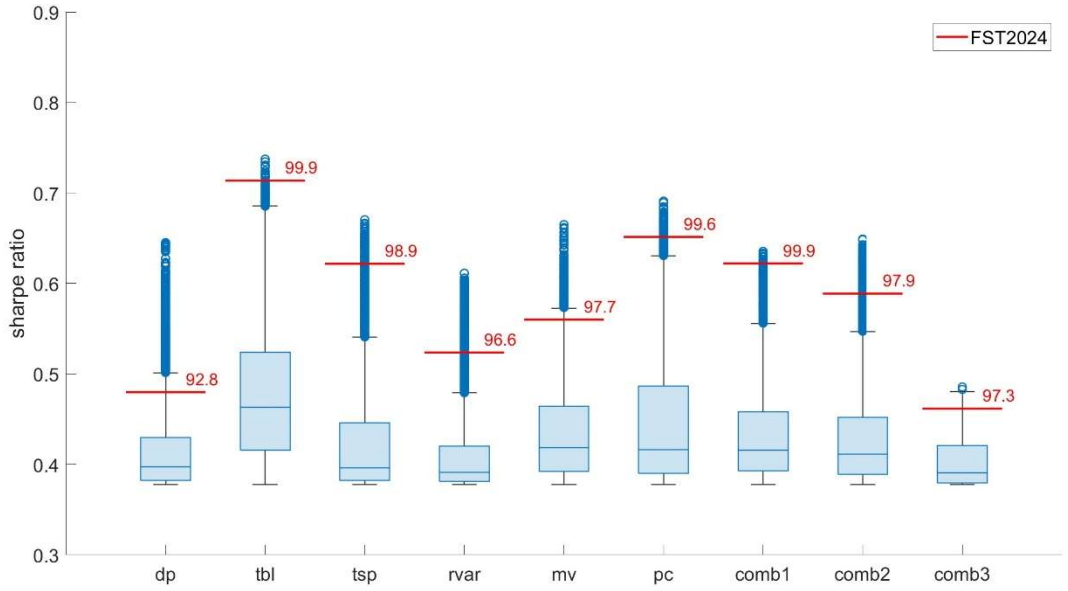


Figure 5. Fork Sensitivity in the Multiverse Analysis

This figure illustrates how sensitive the distribution of results is to the forks in the multiverse analysis associated with different research design choices. The sensitivity is measured with the adjusted standardized Anderson-Darling test statistic. Higher values of the statistic indicate that distributions are more dissimilar across alternative design choices on the fork. We report the results for three different performance measures: Clark and West (2007) test statistics (Panel A), annualized alphas (Panel B), and annualized Sharpe ratios (Panel C). The figures are based on a pooled set of forecasts from all prediction models from Farmer et al. (2023): dividend price ratio (*dp*), treasury bill rate (*tbl*), term spread (*tsp*), and realized variance (*rvar*), and five composite models: principal components (*pc*), four-variable multivariate (*mv*) forecast, and three variants of combination forecasts (*comb1*, *comb2*, *comb3*). The calculations are based on the original data from Farmer et al. (2023), and the study period runs from 1926 to 2016.

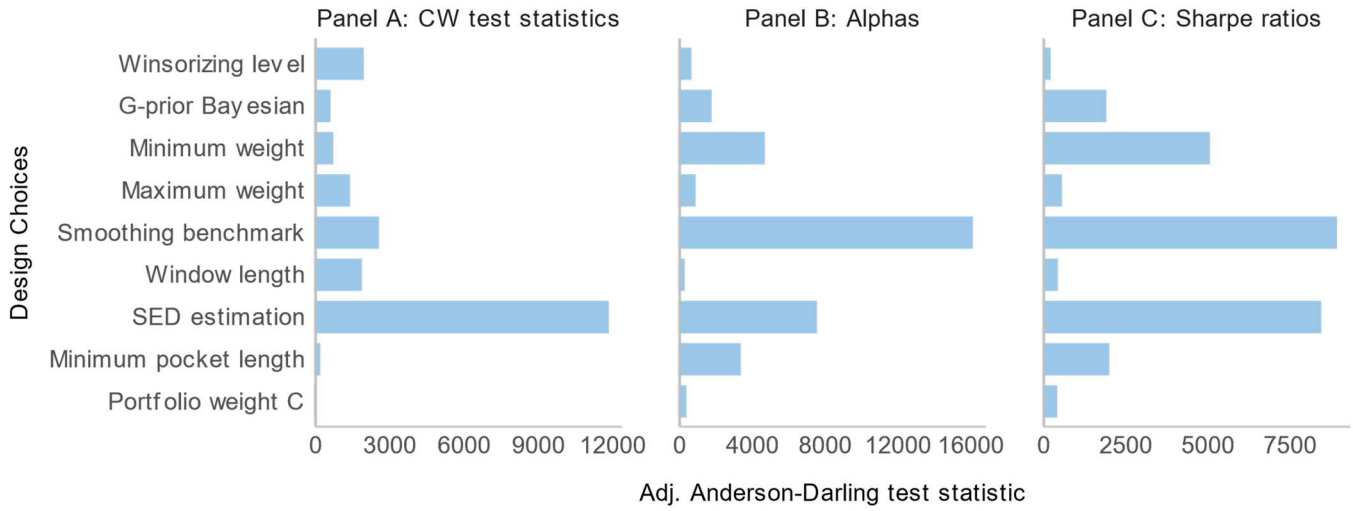


Figure 6. Results Sensitivity to the Modifications of Individual Implementation Parameters

The figure illustrates the impact of the modifications to the original methodology of Farmer et al. (2024). We consider three different performance measures—Clark and West (2007) test statistics (Panel A), annualized alphas (Panel B), and annualized Sharpe ratios (Panel C)—and calculate their values using the original (red dots) and modified (blue dots) methodology. The left axis indicates the methodological modification. The empty (*full*) dots indicate values insignificant (significant) at the 5% level. The forecasts are based on the *comb1* combination method, which aggregates for individual predictors: dividend price ratio (*dp*), treasury bill rate (*tbl*), term spread (*tsp*), and realized variance (*rvar*). The calculations are based on the original data from Farmer et al. (2023), and the study period runs from 1926 to 2016, as available for different predictors. The research decisions are sorted on their impact on the CW test statistics.

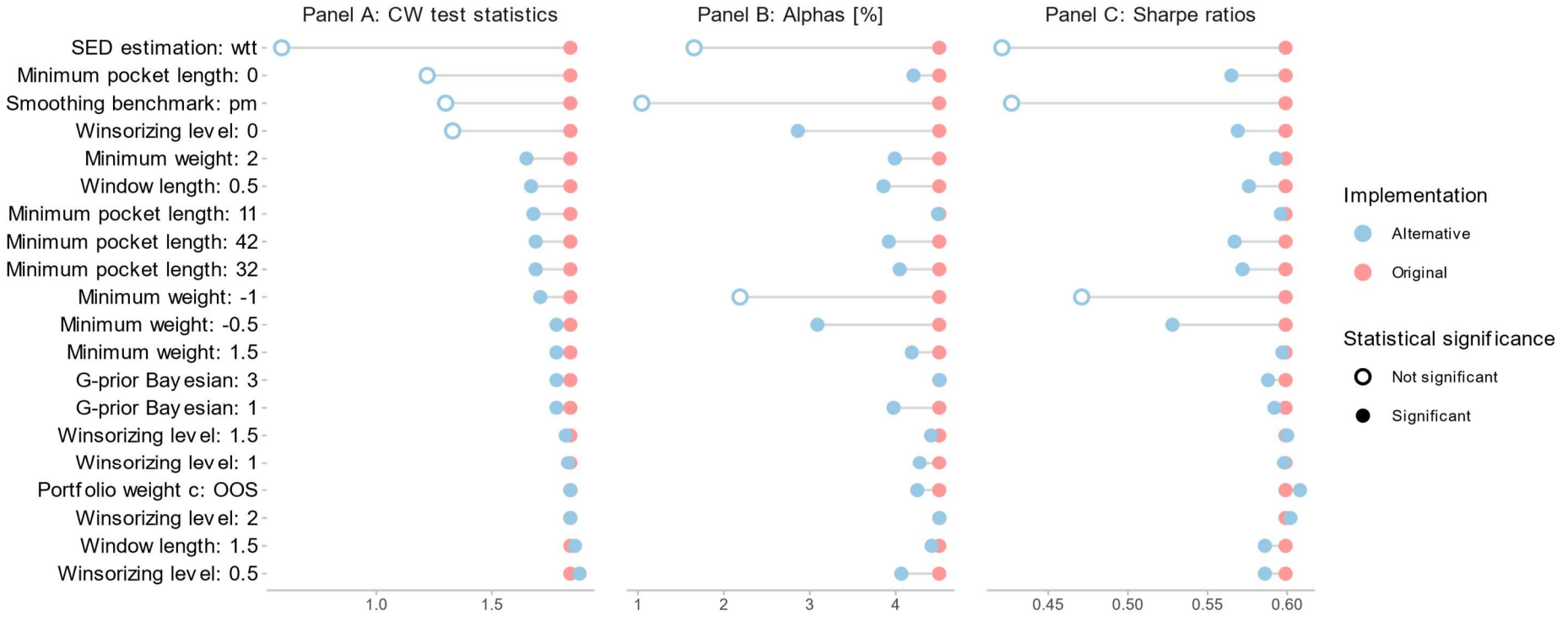
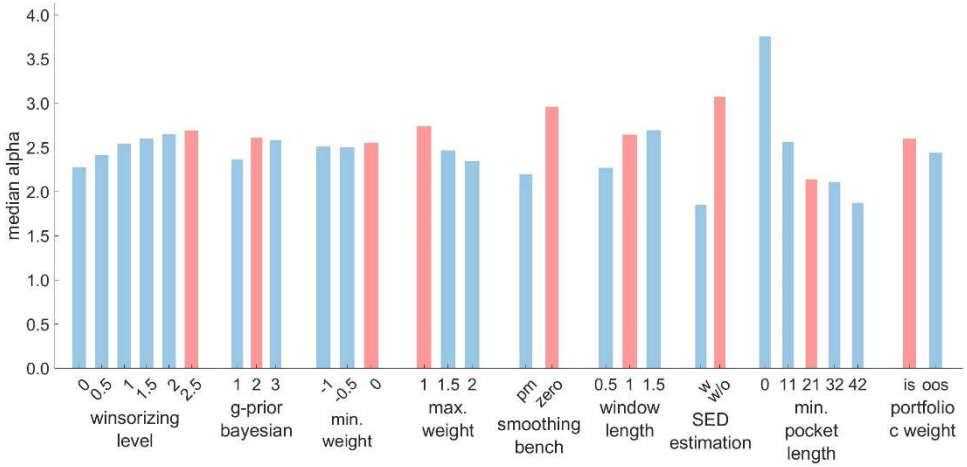


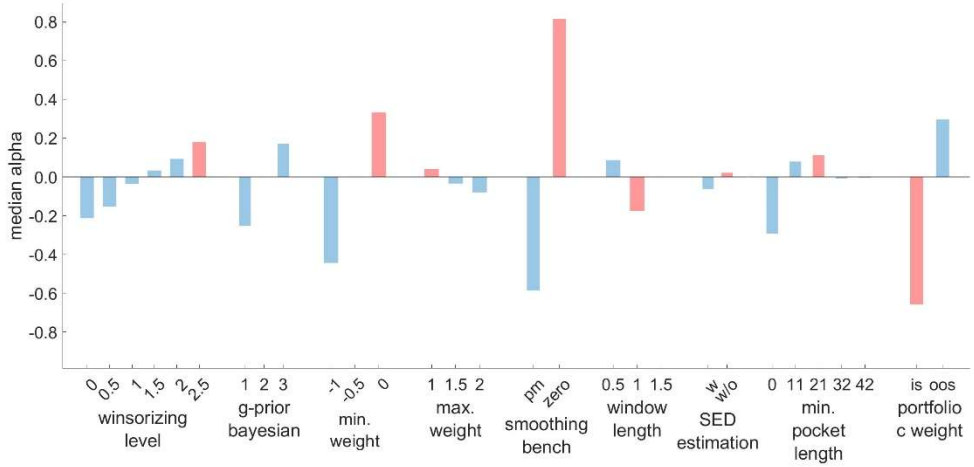
Figure 7. Performance in Subperiods

The figure reports the median alphas (Panels A) and the annualized Sharpe ratios (Panels B) across the distributions of 19,440 replications funneled over individual research design choices. The figures are based on a pooled set of forecasts from all prediction models from Farmer et al. (2023): dividend price ratio (*dp*), treasury bill rate (*tbl*), term spread (*tsp*), and realized variance (*rvar*), and five composite models: principal components (*pc*), four-variable multivariate (*mv*) forecast, and three variants of combination forecasts (*comb1*, *comb2*, *comb3*). The red bars indicate the original research design choice from Farmer et al. (2024). The calculations are based on the original data from Farmer et al. (2023). The full study period runs from 1926 to 2016, as available for different time series, and the figure also reports the results for the subperiods until 1989 and 1990 to 2016.

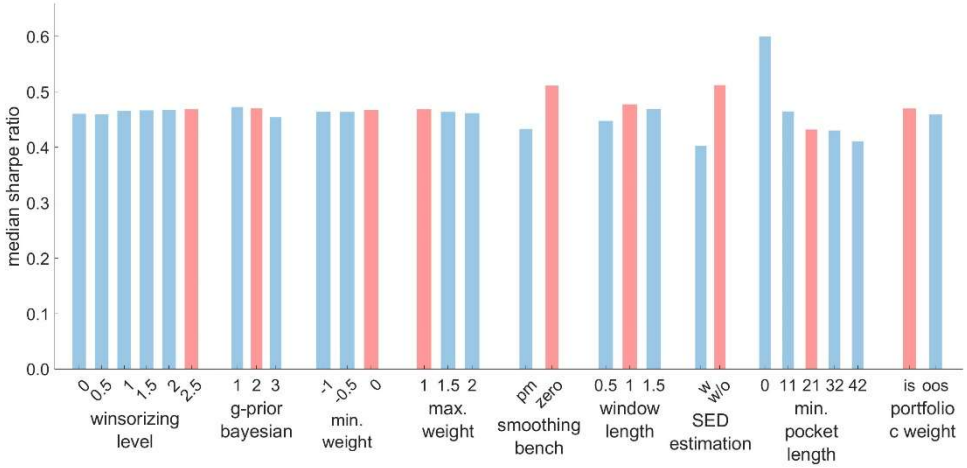
Panel A.1: Alphas pre-1989



Panel A.2: Alphas post-1989



Panel B.1: Sharpe ratios pre-1989



Panel B.2: Sharpe ratios post-1989

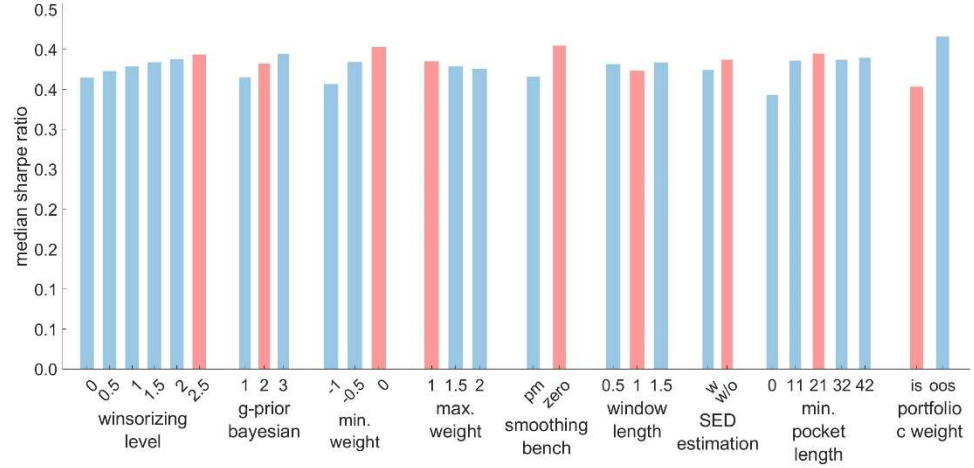


Table 1. The Role of Trading Costs

The table reports performance statistics for active asset allocation strategies based on time-varying market return predictions. The forecasts come from four individual predictors: dividend price ratio (*dp*), treasury bill rate (*tbl*), term spread (*tsp*), and realized variance (*rvar*), and five composite models: principal components (pc), four-variable multivariate (mv) forecast, and three variants of combination forecasts (*comb1*, *comb2*, *comb3*). Panel A reports the alphas (in %), and Panel B focuses on Sharpe ratios. The asterisks * and ** indicate values significantly higher than zero at the 5% and 1% levels, respectively, and the crosses † and †† denote the values lower than zero at the same significance levels. The significance in Panel A is calculated based on Newey and West (1987) adjusted t-statistics, and the Sharpe ratios in Panel B are examined using the tests of Ledoit and Wolf (2008). The performance is reported after adjusting with different levels of transaction costs, from 0 basis points (bps), i.e., effectively implying no costs, to 20 basis points. The full study period runs from 1926 to 2016, as available for different time series, and the table also reports the subperiods until 1989 and 1990 to 2016.

	Trading costs: 0 bps			Trading costs: 5 bps			Trading costs: 10 bps			Trading costs: 15 bps			Trading costs: 20 bps		
	Full	Pre-1989	Post-1989	Full	Pre-1989	Post-1989	Full	Pre-1989	Post-1989	Full	Pre-1989	Post-1989	Full	Pre-1989	Post-1989
Panel A: Alphas															
dp	2.74*	2.16	3.28	1.07	1.72	0.63	-0.59	1.28	-2.02	-2.26	0.83	-4.67†	-3.93††	0.39	-7.32††
tbl	5.48**	5.84**	5.63*	4.56**	4.60**	3.45	3.64**	3.36*	1.27	2.72*	2.13	-0.91	1.80	0.89	-3.09
tsp	4.42**	4.37**	5.05*	2.96*	3.59*	2.84	1.51	2.81	0.63	0.05	2.03	-1.58	-1.41	1.25	-3.79
tvar	3.41**	3.18	4.59*	1.88	2.29	2.37	0.34	1.40	0.16	-1.20	0.51	-2.06	-2.73†	-0.38	-4.27
mv	3.59**	3.54*	4.91*	2.73*	2.11	2.69	1.87	0.69	0.47	1.00	-0.74	-1.75	0.14	-2.16	-3.98
pc	4.57**	4.11*	5.14*	3.47**	3.46*	2.85	2.37*	2.82	0.57	1.27	2.17	-1.71	0.17	1.52	-3.99
comb1	4.51**	4.35**	4.85*	2.11	3.24*	1.84	-0.29	2.13	-1.16	-2.69†	1.02	-4.16	-5.09††	-0.09	-7.16††
comb2	4.47**	4.79**	4.47*	2.58*	3.66*	1.63	0.69	2.53	-1.20	-1.20	1.40	-4.03	-3.09†	0.27	-6.87††
comb3	2.26*	4.85**	0.29	0.30	3.41*	-2.51	-1.65	1.97	-5.30†	-3.61††	0.53	-8.10††	-5.56††	-0.91	-10.90††
Panel B: Sharpe ratios															
dp	0.48	0.42	0.49	0.36	0.38	0.35	0.25	0.33	0.22	0.14†	0.29	0.08††	0.02††	0.25	-0.07††
tbl	0.68**	0.71**	0.62**	0.62*	0.61*	0.50	0.55*	0.50	0.38	0.48	0.39	0.26	0.42	0.29	0.13†
tsp	0.60*	0.59*	0.59*	0.50	0.52	0.47	0.40	0.46	0.35	0.30	0.39	0.23	0.20†	0.33	0.10†
tvar	0.52	0.47	0.57	0.42	0.40	0.45	0.31	0.33	0.33	0.21†	0.26	0.21	0.11††	0.18	0.08††
mv	0.55*	0.54	0.58	0.49	0.41	0.46	0.43	0.28	0.34	0.37	0.15	0.22	0.30	0.02	0.10††
pc	0.62**	0.57*	0.60*	0.54*	0.52	0.48	0.47	0.46	0.35	0.39	0.41	0.23	0.31	0.35	0.10††
comb1	0.60*	0.58*	0.58*	0.44	0.49	0.42	0.28	0.40	0.26	0.12††	0.30	0.10††	-0.04††	0.21	-0.07††
comb2	0.58*	0.60*	0.54*	0.45	0.51	0.39	0.32	0.42	0.24	0.19†	0.32	0.09†	0.06††	0.23	-0.07††
comb3	0.43	0.71*	0.27*	0.23	0.55	0.08†	0.04††	0.38	-0.12††	-0.16††	0.20	-0.31††	-0.36††	0.03	-0.51††