HEC MONTRÉAL

Can Top ESG Stocks Outperform the Investment Universe? Evidence From Mean-Variance Efficiency and Spanning Tests

by

Fatma Ammar

Under the supervision of

David Ardia

Department of Decision Sciences
(Financial Engineering)

In Partial fulfillment of the requirements for
the Degree of Master of Science (M.Sc.)

December 2024

Abstract

We apply a refined framework of mean-variance efficiency (MVE) and mean-variance spanning (MVS) following Gungor and Luger (2016) to assess the efficiency and risk-return properties of top ESG-rated stocks without imposing restrictive parametric assumptions on return distributions. Using daily U.S. equity returns, covering the period from January 2014 to December 2022 from CRSP, we test whether ESG-focused groups optimize the risk-return tradeoff and span the broader stock universe. Our results show that while most ESG-heavy groups do not reject the MVE hypothesis, only a carefully selected subset of ESG stocks fails to reject the MVS hypothesis. Additionally, we find that ESG scores outperform the Media Climate Change Concerns Index (MCCC) in terms of stability and long-term efficiency.

JEL Classification: C12, C14, C53, G11, G23.

Keywords: ESG, Socially responsible investing, Mean-variance spanning, Mean-variance efficiency.

Acknowledgments

At the end of my master's thesis, I would like to take this opportunity to express my appreciation and gratitude to all the people who contributed to the success of this work. I thank my research director, Professor David Ardia, whose invaluable support, guidance, and expertise have profoundly influenced this work. I am genuinely grateful for the opportunity to work under his mentorship and for all the support throughout my studies in the Financial Engineering program. I extend my appreciation to Rosnel Sessinou for his significant input on the methodology of this work. I thank Sermin Gungor and Richard Luger for providing the R code for their tests. I am profoundly thankful to my parents and siblings for their love, encouragement, and sacrifices throughout my academic journey. To my husband, Firas, who has been my biggest support system, thank you for your unwavering love, patience, and belief in me. To my friends who stood by me, your presence has made all the difference.

1 Introduction

In recent years, socially responsible investing (SRI) —defined as an investment approach that integrates Environmental, Social, and Governance (ESG) factors—has been growing exponentially in the asset management industry. By integrating ESG factors, SRI promises to align financial performance with broader societal goals. A pivotal moment in this evolution came when 181 CEOs from major U.S corporations endorsed the Business Roundtable's 2019 statement, redefining corporate purpose to prioritize the welfare of stakeholders, including communities, the environment, and investors.[1]

Global interest in ESG investing is further underscored by the urgent need to address climate change, one of humanity's most pressing challenges. With countries like Canada pledging to reduce its greenhouse gas emissions under the Paris Agreement.[2], understanding how investments impact the environment and society has become critical to achieving a sustainable, carbon-neutral economy. Reflecting this urgency, global ESG assets under management (AUM) surpassed $30 trillion in 2022 and are projected to hit $40 trillion by 2030, making up over a quarter of predicted total AUM.[3] These trends reflect a growing consensus that ESG investing supports sustainable goals and is a pivotal part of long-term financial performance and risk management.

Despite the momentum, key questions remain unanswered: Do top-rated ESG stocks offer an optimal risk-return tradeoff? Are these stocks sufficient to achieve mean-variance efficiency, or does including medium- and low-rated ESG stocks enhance the efficiency of the investment universe? Understanding this is crucial as investors seek to balance sustainability goals with financial performance.

This study seeks to address these gaps by employing mean-variance efficiency (MVE) and mean-variance spanning (MVS) frameworks to evaluate the performance of asset groups based

[1] Business Roundtable Redefines the Purpose of a Corporation to Promote 'An Economy That Serves All Americans' *Business Roundtable*, August 19, 2019.

[2] Adoption of the Paris Agreement *UNFCCC*, December 12, 2015.

[3] Global ESG assets predicted to hit $40 trillion by 2030, despite challenging environment, forecasts Bloomberg Intelligence. *Bloomberg*, February 8, 2024.

on ESG ratings and Media Coverage on Climate Change (MCCC) exposure. Using U.S. stock returns from January 2014 to December 2022, we construct K benchmark and N test universes of stocks, ranked according to ESG scores and MCCC exposure. Building on the methodology of Gungor and Luger (2016), this research examines whether top-rated ESG stocks or those ranked by MCCC exposure can span the broader investment universe or if including lower-rated ESG stocks enhances the mean-variance frontier. The methodology overcomes traditional limitations, such as the reliance on i.i.d. disturbances and multivariate normality, by applying equation-by-equation F-statistics in the multivariate linear model (MLR), thus eliminating disturbance covariance. Additionally, we employ non-parametric bounds tests with Monte Carlo resampling to account for nuisance parameters. This approach makes the method well-suited for large asset universes where the number of stocks exceeds the available periods. The findings contribute to academic research and investment strategies by offering empirical evidence on the efficiency and spanning capabilities of ESG- and MCCC-based groups. Top ESG-rated stocks demonstrate mean-variance efficiency while expanding the universe to include lower-rated stocks diminishes this efficiency. By contrast, MCCC-based asset groups show less consistent results, with frequent rejections of efficiency and spanning hypotheses, reflecting sensitivity to short-term sentiment on climate change.

This paper makes three key contributions. First, it provides empirical insights into the mean-variance efficiency of ESG and MCCC-based stock groups. Second, it evaluates their ability to span larger investment universes. Third, it highlights the trade-offs between relying on ESG ratings for long-term stability and using the MCCC index as a supplementary tool for monitoring climate-related sentiment shifts.

The remainder of this thesis is organized as follows. Section 2 underlines the relevant literature review. Section 3 details the adopted methodologies. Section 4 discusses the data used. Results and analysis are displayed in Section 5, and finally, Section 6 concludes.

2 Literature Review

This section motivates the main empirical questions explored in this research. We begin with an overview of the growing relevance of ESG investing, followed by a review of the mixed findings on the performance of ESG-focused stocks. We then explore the two methodologies for evaluating the performance of stock universes: mean-variance efficiency and mean-variance spanning tests.

2.1 Overview

The assets under management committed to ESG investment strategies have been surging, rising from less than $10 trillion in 2006 to over $120 trillion by 2021.[4] This massive growth follows the creation of the term "ESG," which was first created by major financial institutions in 2004 as a response to a request from Kofi Annan, the UN Secretary-General. Since its introduction, "ESG" has become the standard term for environmental, social, and governance practices.

ESG issues are difficult to measure financially, yet they can decisively affect investments' risk-return profiles. Environmental issues, for instance, directly impact a company's financial performance. Concerns about fossil fuel assets and climate change express themselves in shareholder resolutions at the annual meetings of large oil corporations, such as Shell. The Gulf of Mexico oil spill in 2010 represents the most significant environmental disaster in U.S. history (Nima, 2011), exemplifies this impact, costing BP $23 billion due to the careless cost-cutting corporate culture and excessive risk-taking that caused the spill (Griggs, 2011). Regarding social risks, the Google sexual harassment scandal sparked outrage among workers and led over 20,000 employees globally to leave Google offices in 2019. This incident also prompted shareholders to file a lawsuit against the corporation over its treatment of allegations of executives' sexual misconduct, which resulted in a $310 million settlement (Brown and Peterson, 2022). Highlighting the importance of effective governance, Volkswagen's admission of the emissions cheating scan-

[4] UNPRI's Anuual Report 2021 "Enhance our global footprint" *UNPRI*, 2021.

dal in 2015 resulted in equity market value losses of more than $20 billion within five trading days and abnormal losses for its suppliers due to spillover effects (Barth et al., 2022). This scandal damaged consumer confidence in diesel vehicles and had a lasting impact on the brand's reputation. In recognition of the importance of these issues and the worldwide government support for ESG principles, the United Nations adopts the Sustainable Development Goals.[5] In addition to these extreme events, other ongoing megatrends, such as natural resource scarcity and changing demographics, influence investment strategies. Consequently, detecting and assessing ESG risks have become an integral part of investment decision-making.

The term "ESG investing" is used almost interchangeably with relatively traditional "socially responsible investing," "impact investing," or the most recent "responsible investing" and "sustainable investing." Socially Responsible Investing (SRI) identifies investment risks and opportunities based on ESG metrics (Widyawati, 2020). Over recent years, the market for ESG investing has grown exponentially. The share of global asset owners applying ESG criteria to at least 25% of their total investments increases from 48% in 2017 to 75% in 2019.[6] A report by CNBC in early 2020 also indicates unprecedented inflows into sustainability-focused funds following the outbreak of the COVID-19 pandemic. Therefore, the need to address ESG criteria becomes clear (Krueger et al., 2020); however, what is less clear is the evidence that ESG factors relate, in a causal sense, to higher returns.

2.2 A Review of ESG Stocks and Returns

The consideration of ESG issues in investing for economic gain is a phenomenon that has been around for a while. Research on governance (G) is well-established and shows that better governance increases firm value (Gompers et al., 2003). Research on the effect of the environmental (E) dimension on stock returns is likewise developing and ongoing (Bolton and Kacperczyk, 2021, 2023; Pastor et al., 2021, 2022). The social dimension (S) is still a relatively new area of research (Briscoe-Tran et al., 2024). While sustainable investing undergoes extensive scholarship,

[5] Transforming our world: the 2030 Agenda for Sustainable Development *UN*, September 25, 2021.

[6] ESG Global Survey 2019: investing with Purpose for Performance *BNP Paribas*, May 20, 2019.

a consensus does not yet exist on the performance of ESG-based investments.

Alexander and Osthoff (2007) demonstrate a simple trading strategy using socially responsible ratings from KLD Research, showing that portfolios formed by purchasing high-rated stocks and selling low-rated stocks yield significant abnormal returns. In a broader analysis, Gunnar et al. (2015) conduct a meta-analysis of the existing literature and confirm the positive relationship between ESG factors and corporate financial performance. Pedersen et al. (2020) find that only portfolios based on governance aspects yield significant abnormal returns. At the same time, integrating the environmental and social criteria or overall ESG scores does not improve portfolio performance. Conversely, Harisson and Kacperczyk (2009) show that so-called sin stocks (i.e., companies operating in industries viewed as unethical such as alcohol, tobacco, gambling, and firearms) have higher expected returns due to their neglect by constrained investors. Also, Chava (2010) finds no significant relationship between expected returns and a firm's environmental factors. Rob et al. (2007) establish that the performance gap between ethical and conventional mutual funds remains statistically insignificant, challenging the belief that ESG investing inherently guarantees superior returns. Statman (2006) also reports no statistically significant distinctions between the returns of conventional indexes and social responsibility stock indices.

Other scholars bring different insights in their search for ESG investing trends. Madhavan et al. (2021) highlight the significance of factor exposures in ESG funds as they reveal that high ESG-rated portfolios exhibit distinct factor profiles that lead to superior risk-adjusted returns and that the systematic effects of factors indicate a correlation between ESG metrics and fund performance characteristics. Additionally, the study employs traditional asset pricing models such as Fama and French (2015) and M.Cahart (1997) models and proves that quality factors further delineate the performance expectations based on profitability and investment growth. Pedersen et al. (2020) further propose a model that combines traditional risk-return optimization with ESG considerations. They introduce the "ESG-efficient frontier" to illustrate how different investor types prioritize ESG factors in their portfolio choices. Their findings support the notion that we need a more sophisticated understanding of the efficiency of stock universes when integrating

ESG factors into investment strategies. Additionally, Pastor et al. (2021) predict that green firms outperform brown firms when climate change concerns increase unexpectedly. Ardia et al. (2023) test this prediction and develop a daily Media Climate Change Concerns (MCCC) index from ten leading newspapers and newswires. Unlike the ESG ratings that are provided annually or semi-annually by different rating agencies that could diverge (Dimson et al., 2020) and be relatively noisy (Berg et al., 2022), the MCCC index is built on daily and gives a more detailed view of the market reactions to climate change news.

Given these mixed findings, more research is necessary to fully comprehend the function of the sustainability matrices in optimizing asset universes, especially in relation to well-established financial theories such as mean-variance efficiency and spanning.

2.3 Mean-Variance Efficiency

Central to portfolio theory is the notion of mean-variance efficiency (MVE) introduced by Markowitz (1952), which forms the basis of Modern Portfolio Theory (MPT). It states that a portfolio is mean-variance efficient when no other portfolio offers the same expected return with a lower level of risk. Building on this framework, The Capital Asset Pricing Model (CAPM), developed by Sharpe (1964), Treynor (1999), Lintner (1965) and Mossin (1966), extends this concept, suggesting that the market portfolio should be mean-variance efficient in an efficient market. Extensive research has since focused on testing portfolio mean-variance efficiency, particularly with the development of multivariate tests.

One of the central challenges in testing mean-variance efficiency has been dealing with cross-sectional dependence of the errors when running individual tests. Roll (1977) highlights that grouping individual stocks into portfolios can result in a loss of information about the cross-sectional behavior of individual stocks, as deviations from expected returns may cancel out, reducing the power of the test. To address this, Gibbons et al. (1989) demonstrate the importance of multivariate tests in mean-variance efficiency studies and propose the GRS test, a multivariate version of the t-statistics designed to account for cross-sectional dependence of the errors and

improve the power test, leading to more robust asset pricing models. However, the GRS test has several limitations: it first assumes independent and identically distributed (i.i.d) and normally distributed. Second, it requires the number of assets (N) to be smaller than the time-series observations (T) to avoid singularity in the covariance matrix. Finally, it focuses on unconditional efficiency, and neglects conditional efficiency under different market dynamics.

Several studies have attempted to address the limitations of the conventional mean-variance efficiency tests. Beaulieu et al. (2007), for instance, use simulations to relax the normality assumptions required by GRS by proposing a likelihood ratio test that accommodates non-Gaussian disturbances. The BDK test improves finite-sample methods and leads to fewer rejections of mean-variance efficiency compared to Gaussian-based tests. However, t requires the error distribution to be specified with a finite set of nuisance parameters, which can be restrictive in practice. In contrast, Gungor and Luger (2009) introduce two distribution-free non-parametric sign tests for single-factor models that allow non-normal error distributions but necessitate it to be cross-sectionally independent and conditionally symmetrically distributed around zero. In their later work, Gungor and Luger (2013) broaden this approach to accommodate multiple-factor models.In another approach, Pesaran and Yamagata (2012) develop a new multivariate asymptotic test (i.e., a test that becomes more accurate and reliable as the sample size increases) that outperforms the GRS test for large test sets where $N > T$. The PY test aggregates t-statistics for individual assets and deploys a threshold estimator to account for cross-sectional correlations in the disturbances. However, it assumes weakly and sparsely correlated disturbances, making it less effective as correlation increase

2.4 Mean-Variance Spanning

Mean-variance spanning (MVS) extends the MVE framework by testing whether adding new assets to a benchmark group improves the minimum-variance frontier. Introduced by Huberman and Kandel (1987), this approach evaluates whether the minimum-variance frontier of a given set of benchmark assets changes when new test assets are added, meaning if the new assets are redun-

dant in improving the investment opportunity set. The HK method assesses whether expanding the asset group enhances the risk-return tradeoff by testing whether the frontier with additional assets matches the original one. Transitioning from MVE to MVS thus helps investigate the efficiency of the benchmark group and whether including additional assets can meaningfully alter the efficiency frontier. Furthermore, the HK framework validates linear factor models like the CAPM and the Asset Pricing Theory (APT) to ensure that the integration of test assets aligns with these models' assumptions. The HK test uses a likelihood ratio to examine the mean-variance spanning hypothesis. The findings show that the spanning hypothesis holds over short intervals but is not supported over long periods due to instability in the coefficients.

Building on Huberman and Kandel (1987)'s foundational work, Kan and Zhou (2012) introduce two new mean-variance spanning tests based on the Wald and Lagrange multiplier principles to address the limitations in the original likelihood ratio test, particularly with respect to normality assumptions. They also propose a step-down test and a generalized method of moments (GMM) test for cases where normality does not hold. Gungor and Luger (2016) further contribute to this discourse by developing a finite-sample procedure for testing both the mean-variance efficiency and spanning hypotheses without requiring parametric assumptions about the distribution of disturbances. This methodology employs an equation-by-equation approach to derive exact distribution-free tests for MLR models, accommodating non-normality and time-varying covariances. The procedure improves test power as the time and cross-sectional dimensions increase. Compared to traditional tests (GRS, HK), which struggle with large asset numbers, the Gungor and Luger (2016) framework employs Monte Carlo simulations to allow for computationally inexpensive testing regardless of sample size even when N exceeds T. Their method allows for analyzing financial models that exhibit time-varying conditional covariance structures, such as Multivariate Generalized Autoregressive Conditional Heteroscedasticity (GARCH).

Despite the growing body of literature on ESG investing, little research has focused on whether universes of stocks selected using ESG criteria are mean-variance efficient or if they span the broader investment universe. This study aims to fill this gap by applying the methodolo-

gies developed by Gungor and Luger (2016) to ESG universes.

3 Methodology

This section outlines the methodology from Gungor and Luger (2016), adopting their notations and procedures as the foundation. We first introduce the mean-variance efficiency (MVE) and mean-variance spanning (MVS) hypotheses testing, along with the traditional GRS and HK parametric tests. Next, we present the GL test procedure, an extension of conventional tests within the general MLR framework. We later apply this non-parametric test to test MVE and MVS hypotheses on ESG and MCCC-based stock groups. We incorporate the Fama and French (2015) factors and the MCCC index in regression models to further refine our analysis. The exposure to media sentiment on climate change allows us to construct MCCC-based stock groups and compare them to the ESG stock groups.

3.1 Parametric Tests

We introduce the mean-variance efficiency and spanning hypotheses along with the exact GRS and HK tests, as they provide the foundational framework for the analysis and are essential for understanding the subsequent non-parametric approach.

3.1.1 Mean-Variance Efficiency

Let $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ be a probability filtered space endowed with the filtration $\mathbb{F} = \{t \in T : \mathcal{F}_t\}$ represent the evolution of information available up to time t. We are particularly interested in understanding how the information flow encapsulated by the filtration $\mathbb{F}$ affects the return dynamics of an investment universe that includes a risk-free asset, a set of K benchmark risky assets, and an additional set of N test assets. We would like to see whether there is a relationship between the minimum-variance frontier spanned by the K benchmark assets and the frontier of the combined $N + K$ assets. We use their excess returns to evaluate the relationship between the test and the benchmark assets. At time t, the risk-free return is given by r_{ft}, the benchmark asset returns are

given by r_{Kt}, and the test asset returns are given by r_t. Accordingly, the excess returns on benchmark assets are $z_{Kt} = r_{Kt} - r_{ft}$, while the excess returns on test assets are given by $z_t = r_t - r_{ft}$. We assume the following linear relationship:

$$z_t = a + \beta z_{Kt} + \varepsilon_t, \tag{1}$$

where a is an N-vector of intercepts, β is an $N \times K$ matrix of sensitivities to the benchmark assets, and ε_t is an N-vector of disturbances with $E[\varepsilon_t \mid \mathcal{F}_t] = 0$ and $E[\varepsilon_t \varepsilon_t' \mid \mathcal{F}_t] = \Sigma$. If a group of K benchmark assets is mean-variance efficient, then $E[z_t] = \beta E[z_{Kt}]$. The usual expected return-beta representation's N conditions can be evaluated by testing the null hypothesis:

$$H_E : a = 0, \tag{2}$$

This implies that all pricing errors are zero, meaning the test assets offer no additional explanatory power beyond the benchmark assets.

A multivariate F test of H_E is proposed by the classic mean-variance efficiency test, GRS test, introduced by Gibbons et al. (1989), which states that all the pricing errors comprising the vector a are jointly equal to zero. Conditional on the $T \times K$ collection of components $Z_K = [z_{K1}, ..., z_{KT}]'$, their test implies that the vectors of disturbance terms ε_t, $t = 1, ..., T$, are independent and normally distributed around zero with a cross-sectional covariance matrix that is time-invariant; i.e., $\varepsilon_t \mid Z_K \sim i.i.dN(0, \Sigma)$.

MVE Unconstrained Model

Here, the intercept term a is estimated freely without any restrictions. If $a \neq 0$, then the benchmark assets K are not mean-variance efficient. Under normality, the methods of maximum likelihood and ordinary least squares (OLS) yield the same unconstrained estimates for a and β:

$$\hat{a} = \bar{z} - \hat{\beta}\bar{z}_K,$$

$$\hat{\beta} = \left[\sum_{t=1}^{T}(z_t - \bar{z})(z_{Kt} - \bar{z}_K)'\right]\left[\sum_{t=1}^{T}(z_{Kt} - \bar{z}_K)(z_{Kt} - \bar{z}_K)'\right]^{-1},$$

where $\bar{z} = \frac{1}{T}\sum_{t=1}^{T} z_t$ and $\bar{z}_K = \frac{1}{T}\sum_{t=1}^{T} z_{Kt}$. The unconstrained estimate of the disturbance covariance matrix is:

$$\hat{\Sigma} = \frac{1}{T}\sum_{t=1}^{T}(z_t - \hat{a} - \hat{\beta} z_{Kt})(z_t - \hat{a} - \hat{\beta} z_{Kt})'. \tag{3}$$

MVE Constrained Model

For the constrained model, the intercept term $a = 0$, means that we assume that the benchmark assets are sufficient to describe the return dynamics of the test assets. The estimates are:

$$\hat{\beta}_0 = \left[\sum_{t=1}^{T} z_{Kt} z_{Kt}'\right]^{-1} \left[\sum_{t=1}^{T} z_{Kt} z_t'\right],$$

$$\hat{\Sigma}_0 = \frac{1}{T}\sum_{t=1}^{T}(z_t - \hat{\beta}_0 z_{Kt})(z_t - \hat{\beta}_0 z_{Kt})'. \tag{4}$$

The GRS test statistic for H_E is

$$J_{E,1} = \frac{(T-N-K)}{N}\left[1 + \bar{z}_K' \hat{\Omega}^{-1} \bar{z}_K\right]^{-1} \hat{a}' \hat{\Sigma}^{-1} \hat{a}, \tag{5}$$

where $J_{E,1}$ tests whether the vector of alphas ($\hat{a}$) is significantly different from zero, indicating inefficiencies. The term $\hat{\Omega} = \frac{1}{T}\sum_{t=1}^{T}(z_{Kt} - \bar{z}_K)(z_{Kt} - \bar{z}_K)'$ is the estimated covariance matrix of the benchmark returns. The GRS test statistic can similarly be expressed as:

$$J_{E,1} = \frac{(T-N-K)}{N}\left[\frac{|\hat{\Sigma}_0|}{|\hat{\Sigma}|} - 1\right], \tag{6}$$

which shows that $J_{E,1}$ can be interpreted as a Likelihood Ratio test. Under the null hypothesis H_E, the statistic $J_{E,1}$ follows a central F distribution with N degrees of freedom in the numerator and $(T - N - K)$ degrees of freedom in the denominator (Billou, 2004). We can use it to test whether the test assets provide additional explanatory power beyond the benchmark assets.

3.1.2 Mean-Variance Spanning

Mean-variance spanning occurs when the minimum-variance frontier formed by a set of K benchmark assets (with $K \geq 2$) remains unchanged after adding N test assets. This implies that the K benchmark assets fully represent the diversification opportunities, and the N test assets do not enhance the risk-return tradeoff.

To develop the spanning hypothesis, consider the following statistical model of returns:

$$r_t = a + \beta r_{Kt} + \varepsilon_t, \tag{7}$$

where the disturbance vector ε_t now satisfies $E[\varepsilon_t \mid \mathcal{F}_t] = 0$ and $E[\varepsilon_t \varepsilon_t' \mid \mathcal{F}_t] = \Sigma$. Here, r_t and r_{Kt} are specified in terms of returns, not excess returns. Following Huberman and Kandel (1987), the mean-variance spanning hypothesis imposes the $2N$ restrictions:

$$H_S : a = 0, \delta = 0, \tag{8}$$

where $\delta = \iota_N - \beta \iota_K$ and ι_i is an i-vector of ones. The first condition ensures that the benchmark assets can span returns on the test assets up to a zero-mean, orthogonal factor. The second condition implies that, for each test asset, a combination of the benchmark assets with the same mean return as the test asset but with no additional diversification benefit exists. So the null hypothesis H_S holds: for every test asset, there exists an asset universe of the K benchmark assets with the same mean return as the test asset (since $a = 0$ and $\beta \iota_K = \iota_N$), but with a lower variance (as $\text{Cov}(r_{Kt}, \varepsilon_t) = 0$ and Σ is positive definite). In such a case, the test assets do not enhance the mean-variance frontier spanned by the benchmark assets (Kan and Zhou, 2012).

Huberman and Kandel (1987) propose a procedure similar to the GRS test to test this hypothesis. Given the $T \times K$ matrix of benchmark returns $R_K = [r_{K1}, \ldots, r_{KT}]'$, the HK test requires that $\varepsilon_t \mid R_K \sim$ i.i.d.$N(0, \Sigma)$.

MVS Unconstrained Model

In the unconstrained version of the model, the OLS estimates of the parameters are analogous to those in the GRS test for mean-variance efficiency, given by:

$$\hat{a} = \bar{r} - \hat{\beta} \bar{r}_{Kt},$$
$$\hat{\beta} = \left[\sum_{t=1}^{T} (r_t - \bar{r})(r_{Kt} - \bar{r}_{Kt})' \right] \left[\sum_{t=1}^{T} (r_{Kt} - \bar{r}_{Kt})(r_{Kt} - \bar{r}_{Kt})' \right]^{-1},$$

where $\bar{r} = \frac{1}{T} \sum_{t=1}^{T} r_t$ and $\bar{r}_{Kt} = \frac{1}{T} \sum_{t=1}^{T} r_{Kt}$. The unconstrained estimate of the disturbance covariance matrix is found as

$$\hat{\Sigma} = \frac{1}{T} \sum_{t=1}^{T} (r_t - \hat{a} - \hat{\beta} r_{Kt})(r_t - \hat{a} - \hat{\beta} r_{Kt})'. \tag{9}$$

MVS Constrained Model

We can apply the restrictions in the null hypothesis by splitting the matrix β into two [b_1, C]. Here, b_1 represents an $N \times 1$ vector, while C corresponds to an $N \times (K-1)$ matrix. Partitioning the vector r_{Kt} into the first row r_{1t} and the remaining $K-1$ rows $r_{(K-1)t}$,

$$r_t = a + b_1 r_{1t} + C r_{(K-1)t} + \varepsilon_t,$$

subject to the constraint $\beta l_N = l_N$, which implies that $b_1 + C l_{K-1} = l_N$. Substituting the restrictions $a = 0$ and $b_1 = l_N - C l_{K-1}$, the model becomes:

$$r_t - l_N r_{1t} = C(r_{(K-1)t} - l_{K-1} r_{1t}) + \varepsilon_t. \tag{10}$$

The constrained estimates are given by:

$$\hat{C}_0 = \left[\sum_{t=1}^{T}(r_t - l_N r_{1t})(r_{(K-1)t} - l_{K-1} r_{1t})'\right]\left[\sum_{t=1}^{T}(r_{(K-1)t} - l_{K-1} r_{1t})(r_{(K-1)t} - l_{K-1} r_{1t})'\right]^{-1},$$

$$\hat{b}_{1,0} = l_N - \hat{C}_0 l_{K-1},$$

$$\hat{\Sigma}_0 = \frac{1}{T}\sum_{t=1}^{T}(r_t - \hat{\beta}_0 r_{Kt})(r_t - \hat{\beta}_0 r_{Kt})', \tag{11}$$

where $\hat{\beta}_0 = [\hat{b}_{1,0}, \hat{C}_0]$. Following the LR form, the HK test can be written as:

$$J_S = \frac{(T-N-K)}{N}\left[\sqrt{\frac{|\hat{\Sigma}_0|}{|\hat{\Sigma}|}} - 1\right], \tag{12}$$

and under the null hypothesis H_S, this statistic follows a central F distribution with $2N$ degrees of freedom in the numerator and $2(T-N-K)$ degrees of freedom in the denominator.

3.2 Non-Parametric Tests

We present the non-parametric bounds test of efficiency and spanning introduced by Gungor and Luger (2016), which relaxes the four assumptions of the exact J_E and J_S tests outlined in the previous section. Specifically, these tests address the limitations of (i) independence of disturbances, (ii) identically distributed disturbances, (iii) normally distributed disturbances, and (iv) the condition $N \leq T - K - 1$.

3.2.1 MLR Framework

The analysis is conducted within a MLR framework, represented by:

$$Y = XB + \varepsilon, \tag{13}$$

where Y is a $T \times N$ matrix of dependent variables, X is a $T \times (K+1)$ matrix of regressors, and $\varepsilon = [\epsilon_1, \ldots, \epsilon_T]'$ is the $T \times N$ matrix of residuals. The parameter matrix $B = [a, \beta]'$, has dimensions $(K+1) \times N$. For model (1), $Y = [z_1, \ldots, z_T]'$ and $X = [\iota_T, Z_k]$. In contrast, for model (7), $Y = [r_1, \ldots, r_T]'$ and $X = [\iota_T, R_k]$.

The parameter matrix B is estimated subject to linear constraints to test the null hypothesis:

$$H_0 : HB = D, \tag{14}$$

where H is an $h \times (K+1)$ matrix of constants of rank h, and D is an $h \times N$ matrix of constants. Specifically, the efficiency hypothesis in (2) is obtained by setting $H = [1, 0, \ldots, 0]$ and $D = [0, \ldots, 0]$. For the spanning hypothesis in equation (8):

$$H = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 0 & \ldots & 0 \\ 1 & \ldots & 1 \end{bmatrix}.$$

The same hypothesis is tested across all equations in the MLR system.

MLR Unrestricted Model

With the MLR model, we can derive the unrestricted OLS estimates and residuals as follows:

$$\hat{B}(Y) = (X'X)^{-1}X'Y, \tag{15}$$

$$\hat{\varepsilon}(Y) = Y - X\hat{B}(Y) = MY = M\varepsilon,$$

where $M = I - X(X'X)^{-1}X'$. Here the $i$the column of $\hat{B}(Y) = [\hat{B}_1(Y), \ldots, \hat{B}_N(Y)]$ minimizes the $i$the diagonal element of the sum-of-squares and cross-products matrix

$$\mathcal{E} = (Y - XB)'(Y - XB),$$

and its estimate is

$$\hat{\mathcal{E}}(Y) = \hat{\varepsilon}'(Y)\hat{\varepsilon}(Y). \tag{16}$$

MLR Restricted Model

We minimize the residual sum-of-squares in $\mathcal{E}$ subject to the restrictions in the null hypothesis (14), resulting in the constrained estimates and residuals:

$$\hat{B}_0(Y) = \hat{B}(Y) - (X'X)^{-1}H'[H(X'X)^{-1}H']^{-1}[H\hat{B}(Y) - D], \tag{17}$$

$$\hat{\varepsilon}_0(Y) = Y - X\hat{B}_0(Y) = M_0 Y = M_0\varepsilon,$$

with $M_0 = M + X(X'X)^{-1}H'[H(X'X)^{-1}H']^{-1}H(X'X)^{-1}X'$. The corresponding restricted residual sum-of-squares and cross-products matrix is

$$\hat{\mathcal{E}}_0(Y) = \hat{\varepsilon}_0'(Y)\hat{\varepsilon}_0(Y). \tag{18}$$

We want to allow for time-varying conditional covariance structures of unknown form while maintaining flexibility in the distribution of disturbances. We would also like to avoid the singularity problem for the matrices $\hat{\mathcal{E}}(Y)$ and $\hat{\mathcal{E}}_0(Y)$ when N > T to compute the usual statistics.

Gungor and Luger (2016) provide a test procedure that is also derived from (16)) and (18), but avoids the singularity problem by not requiring the determinants $|\hat{\Sigma}_t|$ and $|\hat{\Sigma}_0|$ shown in (6) and (12) for the GRS and HK tests. The distributional theory underlying their approach rests on a multivariate Assumption 1, which includes the normal distribution assumed by GRS and HK as a special case.

Assumption 1 (Reflective Symmetry)**.** *The cross-sectional disturbance vectors ε_t, $t = 1, \ldots, T$, are jointly continuous and reflectively symmetric so that*

$$(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_T \mid X) \stackrel{d}{=} (\pm\varepsilon_1, \pm\varepsilon_2, \ldots, \pm\varepsilon_T \mid X),$$

where each ε_t is assigned either a positive or negative sign with equal probability and the symbol "$\stackrel{d}{=}$" stands for the equality in distribution.

The assumption states that conditional on X, ε_t has the same distribution as $-\varepsilon_t$. This allows for a broad class of distributions, including elliptically symmetric distributions, which are compatible with expected utility maximization regardless of investor preferences. Several time-varying covariance models, such as multivariate GARCH or stochastic volatility models, also satisfy the assumption.

3.2.2 Test Procedure

The test procedure described here uses equation-by-equation F-statistics, which can be computed from the abovementioned unrestricted and restricted OLS estimates. These F-statistics evaluate whether imposing restrictions on the model (i.e., MVE hypothesis and MVS hypothesis) results in a loss of explanatory power. The $N \times 1$ vector of F statistics is given by:

$$F(Y) = \frac{\left(diag\left\{\hat{\mathcal{E}}_0(Y)\right\} - diag\left\{\hat{\mathcal{E}}(Y)\right\}\right)/h}{diag\left\{\hat{\mathcal{E}}(Y)\right\}/(T-K-1)}, \tag{19}$$

where diag$\{\cdot\}$ extracts the diagonal elements of a square matrix, which means we are considering only the variances of the residuals for each equation; h is the number of rows of the restriction matrix H in the MLR null hypothesis, and the division in the numerator and denominator is carried out element by element. The single-equation F statistic is represented by the i-the element of the N-vector $F(Y) = [F_1(Y), \ldots, F_N(Y)]'$ which takes the usual form of:

$$F_i(Y) = \frac{\left(\text{RSS}_{0,i}(Y) - \text{RSS}_i(Y)\right)/h}{\text{RSS}_i(Y)/(T-K-1)},$$

where the terms $\text{RSS}_i(Y)$ and $\text{RSS}_{0,i}(Y)$ represent the residual sum-of-squares of each model. Here, the degrees-of-freedom term $(T-K-1)/h$ could be omitted from (19) since it is just a constant under the proposed permutation approach. We consider the maximum F statistics:

$$F_{\max}(Y) = \max\{F_1(Y), \ldots, F_N(Y)\}, \tag{20}$$

which selects the individual F statistic that indicates the greatest violation from the null hypothesis. The $F_{\max}(Y)$ statistic highlights the equation where the restriction has the most significant impact, which helps to pinpoint potential inefficiencies. It is important to mention that $F_{\max}(Y)$

can be computed even when $N > T$, as the $F_i(Y)$ statistics can be derived from individual equations. As the sample size increases, the precision of the estimates a_i improves.

3.2.3 Building blocks

The $F_{\max}(Y)$ statistic under H_0 is influenced by the nuisance parameters B that are not specified by the null hypothesis, which can affect the distribution of the test statistic, making it difficult to derive accurate critical values for hypothesis testing. To enable hypothesis testing in a distribution-free way, we establish exact bounds for the distribution of the test statistics under H_0. These bounds address the unrestricted elements of B using a point null hypothesis:

$$H_0^* : H_0, B = B^*, \tag{21}$$

where B^* represents specified values chosen to satisfy the null hypothesis, ensuring $H_0^* \subseteq H_0$. Let $\varepsilon^* = Y - XB^*$, which, under H_0^*, aligns with ε, the true model residuals. Notably, the structure of H_0^* depends on the choice of B^*, which in turn affects the H_0-restricted residuals ε^*.

Bootstrap Sample Construction

We employ the bootstrap approach to generate samples. We introduce $\tilde{s} = [\tilde{s}_1, \ldots, \tilde{s}_T]'$ a T-vector consisting of independent Bernoulli random variables with $\Pr[\tilde{s}_t = 1] = \Pr[\tilde{s}_t = -1] = \frac{1}{2}$ for all t. A bootstrap sample of dependent variables is defined as:

$$\tilde{Y} = XB^* + \tilde{s} \cdot \varepsilon^*, \tag{22}$$

where $\tilde{s} \cdot \varepsilon^*$ represents the modified residuals. So by applying the random signs $\tilde{s}$, we are generating a new dataset $\tilde{Y}$ that is a random perturbation of the original model but still retains the same underlying covariance structure. This construction preserves the contemporaneous covariance structure among the elements of ε^*. Under H_0^* in (21) and conditional on X, we have $Y \overset{d}{=} \tilde{Y}$ (i.e, Y and $\tilde{Y}$ have the same distribution). From Theorem 1.3.7 in Randles and Wolfe (1979), if $Y \overset{d}{=} \tilde{Y}$ and $\mathcal{F}(\cdot)$ is a measurable function defined on the common support of Y and $\tilde{Y}$, then $\mathcal{F}(Y) \overset{d}{=} \mathcal{F}(\tilde{Y})$. This means that if we compute a test statistic $F(\tilde{Y})$, it will have the same dis-

tribution as $\mathcal{F}(Y)$, allowing us to use $\tilde{Y}$ to approximate the behavior of the test statistic under H_0.

Proposition 1 (Equally Likely Property)**.** *Suppose the MLR model in (13) holds with the Reflective Summary Assumption. Let $\tilde{Y}$ be a bootstrap sample generated according to Equation (21) for a specific realization of $\tilde{s}$, and consider the statistic $F(\tilde{Y})$ computed from this bootstrap sample. Then, under H_0^* in (21) and given X, the 2^T values of $\mathcal{F}(\tilde{Y})$ obtained from all possible realizations of $\tilde{s}$ are equally likely values for $\mathcal{F}(Y)$.*

The Proposition shows that $\mathcal{F}(Y)$ is pivotal under H_0, meaning its bootstrap distribution does not depend on any unknown nuisance parameters. Critical values can, in principle, be derived from the conditional distribution of $\mathcal{F}(Y)$ based on the 2^T equally likely possibilities represented by $\mathcal{F}(\tilde{Y})$. However, obtaining this distribution by enumerating all realizations of $\tilde{s}$ is impractical.

Monte Carlo Test Procedure

We employ the Monte Carlo (MC) test technique (Barnard, 1963; Birnbaum, 1974; Dwass, 1957) to approximate the distribution under the null hypothesis H_0. Instead of enumerating all possible bootstrap samples, we randomly generate $M-1$ bootstrap samples $\tilde{Y}_1, \ldots, \tilde{Y}_{M-1}$. For each sample, we compute the $\mathcal{F}(\cdot)$ statistic to yield $\mathcal{F}(\tilde{Y}_m)$ for $m = 1, \ldots, M-1$. When using the MC test procedure, we generate several bootstrap samples and compute a test statistic for each sample. However, since the statistic $\mathcal{F}(\cdot)$ is calculated from a finite set of values, some of them may be the same across different bootstrap samples, which leads to ties. These ties complicate ranking the observed $\mathcal{F}(Y)$ among bootstrap values, affecting the accuracy of our p-value calculation. To manage this, we adopt a tie-breaking rule (Dufour, 2006). We draw M i.i.d. variables U_m from a continuous uniform distribution on $[0, 1]$, independent of the $\mathcal{F}(\cdot)$ statistics, and pair the U and $\mathcal{F}(\cdot)$ statistics. We compute the lexicographic rank of $(\mathcal{F}(Y), U_M)$ as follows:

$$\tilde{R}_M(\mathcal{F}(Y)) = 1 + \sum_{m=1}^{M-1} I[\mathcal{F}(Y) > \mathcal{F}(\tilde{Y}_m)] + \sum_{m=1}^{M-1} I[\mathcal{F}(Y) = F(\tilde{Y}_m)] I[U_M > U_m], \tag{23}$$

where $I[A]$ denotes the indicator function of event A. Recognizing that the pairs

$$(\mathcal{F}(\tilde{Y}_1), U_1), \ldots, (\mathcal{F}(\tilde{Y}_M), U_{M-1}), (\mathcal{F}(Y), U_M)$$

are exchangeable under H_0, we can derive from Lemma 2.3 in (Dufour, 2006) that the lexicographic ranks are uniformly distributed across the integers $1, \ldots, M$, specifically:

$$\Pr[\tilde{R}_M[\mathcal{F}(Y)] = m] = \frac{1}{M}, \quad \text{for} m = 1, \ldots, M.$$

Thus, the MC p-value can be expressed as

$$\tilde{p}_M[\mathcal{F}(Y)] = \frac{M - \tilde{R}_M(\mathcal{F}(Y)) + 1}{M}. \tag{24}$$

where $\tilde{R}_M[\mathcal{F}(Y)]$ is the rank of $(\mathcal{F}(Y), U_M)$ given by Equation (23). This p-value allows us to determine whether the observed statistic $\mathcal{F}(Y)$ is extreme relative to the bootstrap distribution. If αM is an integer then

$$\Pr[\tilde{p}_M[\mathcal{F}(Y)] \leq \alpha \mid H_0] = \alpha.$$

The MC test for H_0^* enables the formulation of our proposed bounds tests for H_0, the hypothesis of interest. The main idea is to derive both a liberal and a conservative test, each with a nominal level α. The null hypothesis H_0 will not be rejected when the liberal test does not reject it, and it will be rejected when the conservative test yields a significant result.

3.2.4 Liberal and Conservative Tests

The null hypothesis assumes that $B^* = \hat{B}_0$, which represents the OLS estimate of B under H_0. By construction, we have $H\hat{B}_0 = D$, meaning that H_0^* is compatible with the original null hypothesis H_0. The residuals, denoted by ε^*, are equivalent to those obtained under H_0, which simplifies to $\varepsilon^* = \hat{\varepsilon}_0$.

For the liberal test, denote by $\tilde{P}_M^L(\mathcal{F}(Y))$ the associated MC p-value computed according to (24). This liberal p-value satisfies $\Pr\left[\tilde{P}_M^C(\mathcal{F}(Y)) > \alpha \mid X\right] \leq 1 - \alpha$ under H_0, implying that the decision rule to do not reject H_0 when $\tilde{P}_M^L(\mathcal{F}(Y)) > \alpha$ is valid. This decision rule is based on the observation that $H_0^* \subseteq H_0$; thus, if H_0^* is not rejected, neither is H_0.

For the conservative test, we introduce a test statistic specific to this point null hypothesis (21). Let the residual sum-of-squares and cross-products matrix to H_0^* be expressed as $\hat{\mathcal{E}}^* = \hat{\varepsilon}^{*'}\hat{\varepsilon}^*$. We

then consider the $N \times 1$ vector of test statistics:

$$\mathrm{F}^{C}(\mathrm{Y}) = \frac{\left(\mathrm{diag}\left\{\hat{\varepsilon}^{*}\right\} - \mathrm{diag}\left\{\hat{\varepsilon}(\mathrm{Y})\right\}\right)/h}{\mathrm{diag}\left\{\hat{\varepsilon}(\mathrm{Y})\right\}/(T-K-1)}, \tag{25}$$

where the superscript C denotes that this is a conservative test statistic. When computed with the original sample Y, we have $F^{C}(Y) = F(Y)$ as we set $B^{*} = \hat{B}_{0}$.For any bootstrap sample $\tilde{Y}$, generated according to (25), the following inequalities hold:

$$\mathrm{diag}\left\{\varepsilon^{*}\right\} \geq \mathrm{diag}\left\{\hat{\varepsilon}_{0}(\tilde{Y})\right\} \geq \mathrm{diag}\left\{\hat{\varepsilon}(\tilde{Y})\right\}, \tag{26}$$

where these comparisons are element-wise. This means that an OLS residual sum of squares calculated with restrictions cannot be smaller than one calculated with fewer restrictions (Davidson and MacKinnon, 2004). From these inequalities, it follows that:

$$F(\tilde{Y}) \leq \mathrm{F}^{C}(\mathrm{Y}). \tag{27}$$

As with the liberal statistics, the conservative statistics $F^{C}(\cdot)$ can be aggregated using a p-norm. Denote this aggregation as $F_{C}(\cdot)$, which could either represent $F_{p}^{C}(\cdot)$ or $F_{\max}^{C}(\cdot)$. The relationships we described imply that $\Pr[\mathcal{F}(\cdot) > \theta] \leq \Pr[\mathcal{F}^{C}(\cdot) > \theta]$, for any threshold value $\theta \in \mathbb{R}$. Define θ_{α} as a critical value such that $\Pr[\mathcal{F}(Y) > \theta_{\alpha} \mid X] = \alpha$ whenH_{0}^{*} holds, and similarly define θ_{α}^{C} via: $\Pr[\mathcal{F}^{C}(\mathrm{Y}) > \theta_{\alpha}^{C} \mid X] = \alpha$ underH_{0}^{*}. It follows that: $\theta_{\alpha} \leq \theta_{\alpha}^{C}$, implying that:

$$\Pr[\mathcal{F}(Y) > \theta_{\alpha}^{C} \mid X] \leq \alpha \quad \text{when} \mathcal{F}(Y) \text{ follows its} H_{0}\text{-distribution.} \tag{28}$$

Thus, if the joint F bounds test based on θ_{α}^{C} is significant, then the exact joint F test based on θ_{α} is also significant at level α. We apply the MC test technique to implement the bounds test.

Proposition 2 (Bounds MC p-values)**.** *Suppose the MLR model in* (13) *with Assumption 1 holds. Further, consider a statistic* $\mathcal{F}(Y)$ *for testing* H_{0} *and the corresponding conservative test statistic* $\mathcal{F}^{C}(Y)$*. Define the liberal and conservative MC p-values as*

$$\tilde{p}_{M}^{L}[\mathcal{F}(Y)] = \frac{M - \tilde{R}_{M}[\mathcal{F}(Y)] + 1}{M} \quad \textit{and} \quad \tilde{p}_{M}^{C}[\mathcal{F}(Y)] = \frac{M - \tilde{R}_{M}^{C}[\mathcal{F}(Y)] + 1}{M},$$

where $\tilde{R}_M[\mathcal{F}(Y)]$ *and* $\tilde{R}^C_M[\mathcal{F}(Y)]$ *are the lexicographic ranks of* $\mathcal{F}(Y)$ *among* $\mathcal{F}(\tilde{Y}_m)$ *and* $\mathcal{F}^C(\tilde{Y}_m)$*, respectively, for* $m = 1, \ldots, M-1$*. The bootstrap samples* $\tilde{Y}_m$ *are generated according to the model (22), and the lexicographic ranks are computed as*

$$\tilde{R}_M[\mathcal{F}(Y)] = 1 + \sum_{m=1}^{M-1} \mathbb{I}\left[\mathcal{F}(Y) > \mathcal{F}(\tilde{Y}_m)\right] + \sum_{m=1}^{M-1} \mathbb{I}\left[\mathcal{F}(Y) = \mathcal{F}(\tilde{Y}_m)\right] \times \mathbb{I}[U_M > U_m],$$

$$\tilde{R}^C_M[\mathcal{F}(Y)] = 1 + \sum_{m=1}^{M-1} \mathbb{I}\left[\mathcal{F}(Y) > \mathcal{F}^C(\tilde{Y}_m)\right] + \sum_{m=1}^{M-1} \mathbb{I}\left[\mathcal{F}(Y) = \mathcal{F}^C(\tilde{Y}_m)\right] \times \mathbb{I}[U_M > U_m],$$

where U_m*,* $m = 1, \ldots, M$*, are i.i.d. uniform variates on* $[0, 1]$*, independently of the* $\mathcal{F}$ *statistics. If* αM *is an integer, then* $\Pr\left[\tilde{p}^L_M(\mathcal{F}(Y)) > \alpha \mid X\right] \leq 1 - \alpha$ *and* $\Pr\left[\tilde{p}^C_M(\mathcal{F}(Y)) \leq \alpha \mid X\right] \leq \alpha$*, under* H_0 *in the null hypothesis (14).*

This conclusion is derived from Proposition 2.4 in Dufour (2006), which addresses the validity of MC tests for general statistics. Notably, in Proposition 2, it is essential that the same bootstrap sample $\tilde{Y}_m$ is used to compute both $\mathcal{F}(\tilde{Y}_m)$ and $\mathcal{F}^C(\tilde{Y}_m)$. Additionally, the same set of uniform random variables $U_1, \ldots, U_M$ should be applied when calculating both $\tilde{R}_M[\mathcal{F}(Y)]$ and $\tilde{R}^C_M[\mathcal{F}(Y)]$. These conditions are necessary to ensure consistency and prevent any contradictory results between the liberal and conservative MC p-values.

Bounds MC Test Decision Rule

The decision rule for the MC bounds test of $H_0 : HB = D$ at level α is as follows:

- Reject H_0 if $\tilde{P}^C_M(\mathcal{F}(Y)) \leq 5\%$.
- Do not reject H_0 if $\tilde{P}^L_M(\mathcal{F}(Y)) > 5\%$.
- Inconclusive if neither condition is met.

The logic behind this rule is similar to the well-known bounds test of Durbin and Watson (1950, 1951) for detecting autocorrelation in regression models.

To sum up, the non-parametric test presents the foundational framework of our study evaluating the efficiency and spanning properties of top ESG-rated and MCCC-based stocks. The following section builds on this framework by introducing the methodology for extracting stocks' exposure to climate sensitivity.

3.3 Fama French Model With Media Climate Change Concerns Sensitivity

The Fama and French (2015) model takes the following form:

$$R_{it} - R_{Ft} = a_i + b_i(R_{Mt} - R_{Ft}) + s_i\, SMB_t + h_i\, HML_t + r_i\, RMW_t + c_i\, CMA_t + \varepsilon_{it} \quad (29)$$

where R_{it} is the return of the stock at time t, R_{Ft} is the risk-free rate, and R_{Mt} is the market return. SMB_t is the return spread of small-cap stocks minus large-cap stocks; HML_t is the return spread of high book-to-market (B/M) ratio stocks minus low (B/M) ratio stocks; RMW_t is the return spread of stocks with robust profitability minus those with weak profitability, and CMA_t is the return spread of stocks with conservative investment policies minus those with aggressive investment policies. The error term ε_{it} follows a normal distribution, $\varepsilon_{it} \sim \mathcal{N}(0, \sigma^2)$. If the exposures to the five factors b_i, s_i, h_i, r_i, and c_i capture all variation in expected returns, the intercept a_i is zero for all securities and portfolios i. We extend the model by incorporating a new variable, m_i, which represents the exposure of each stock to the Market Climate Change Component (MCCC), which captures market reactions to climate-related events and policies. The extended model is represented as follows:

$$R_{it} - R_{Ft} = a_i + b_i(R_{Mt} - R_{Ft}) + s_i\, SMB_t + h_i\, HML_t + r_i\, RMW_t + c_i\, CMA_t + m_i\, MCCC_t + \varepsilon_{it} \quad (30)$$

The Fama-French factors account for broad systematic risks, including market, size, value, profitability, and investment factors. This isolates climate sensitivity from broader market risks and allows the MCCC variable to capture the specific exposure of stocks to climate-related events and policies. This ensures that the impact of climate change on stock returns is captured independently of other risk factors.

4 Data

This section provides a detailed overview of our analysis's data sources and types. We first describe the stock returns data, followed by ESG ratings, and then the finalized datasets for ESG

hypotheses testing. The primary data source is Wharton Research Data Services (WRDS). In addition to the ESG data, we incorporate explanatory variables that capture systematic risk factors –Fama and French (2015) Factors and MCCC index, where the latter serves as an alternative dynamic measure to ESG to assess mean-variance efficiency and mean-variance spanning.

4.1 Returns Data

We source the stock returns data from the Center for Research in Security Prices (CRSP), which provides comprehensive security price, return, and volume data for the NYSE, AMEX, and NASDAQ stock markets. The database is free from survivorship bias and accounts for organizational events, including name changes, mergers, and liquidations. We use US-listed firms' daily returns from January 2, 2014, to December 31, 2022.

We retrieve daily returns data and the CRSP unique permanent security level identifiers (*PERMNO*) from the daily stock returns database to obtain the dataset. We do not consider ESG factors at this stage, so the initial dataset compromises both ESG-rated and ESG-unrated stocks. We also retrieve stock tickers from the stock names database and merge them with the data. We consider only the latest available ticker for each company to avoid any potential biases caused by historical tickers. Furthermore, we aggregate the returns by date and ticker to construct a time series dataset in an extended format. Finally, we transform the data into a panel data format, with each column representing the return of an individual stock, to facilitate efficient analysis. Table 1 summarizes the number of stocks with available daily returns from 2014 to 2022. The number of stocks included in the dataset increased steadily over the sample period.

Table 1: Evolution of the Stock Return Database
This table provides an overview of the number of individual stocks with available daily return data for each year from 2014 to 2022.

Year	2014	2015	2016	2017	2018	2019	2020	2021	2022
Stocks with Returns	1881	1997	2076	2124	2209	2298	2352	2401	2469

4.2 ESG Data

We use Trucost ESG Disclosure Scores, also known as S&P Global ESG Scores, to obtain ESG scores of the U.S.-listed companies. Unlike other ESG datasets that rely solely on publicly available information, S&P Global ESG Scores are generated from a combination of verified company disclosures, media, stakeholder analysis, and in-depth company engagement through the S&P Global Corporate Sustainability Assessment (CSA), providing unparalleled access to ESG insights before they reach others.

We collect ESG scores for U.S-listed companies classified as "Operating" from the Trucost database that matched the tickers obtained from our initial returns dataset. The time series data is monthly. The ESG data is monthly, and to ensure consistency, we generate all possible year-month combinations from January 2014 to December 2022 and fill any gaps using a forward-fill method. Subsequently, we remove unrated stocks to obtain monthly ESG scores. We repeat this process for each ESG dimension—Environmental, Social, and Governance— by adjusting the criteria in the extraction query. Table 2 illustrates the evolution of the number of ESG-rated stocks in our dataset from 2014 to 2022. The size of the datasets significantly increased after 2020, showcasing the increasing corporate transparency and reporting on ESG factors.

Table 2: Evolution of ESG-Rated Stocks
This table presents the number of U.S.-listed stocks with available ESG ratings, segmented by Environmental, Social, and Governance factors, from 2014 to 2022. The dataset is sourced from Trucost (S&P Global ESG Scores).

Year	2014	2015	2016	2017	2018	2019	2020	2021	2022
Size of ESG data	389	455	493	542	584	705	1058	1895	1579
Size of Environmental data	389	455	493	542	584	705	1058	1895	1579
Size of Social data	389	455	493	542	584	705	1058	1895	1579
Size of Governance data	389	455	493	542	584	705	1058	1895	1579

The S&P Global ESG Scores range from 0 to 100. Table 3 summarizes the descriptive statistics of ESG metrics and their dimensions—Environmental, Social, and Governance. Over the years, the mean ratings have declined, notably in 2020. This decline coincides with a sharp increase in ESG-rated stocks (Table 2), many of which received lower ESG scores. Among the ESG dimensions, governance maintains higher statistics than environmental and social scores. Gov-

ernance practices are well-established due to longstanding regulatory and stakeholder demands. In contrast, Environmental and Social metrics are newer focus areas, with many companies still developing tracking and disclosure frameworks.

Table 3: Descriptive Statistics of ESG, Environmental, Social, and Governance Ratings
The table presents descriptive statistics for ESG metrics over the years 2014 to 2022, including minimum (Min), first quartile (Q1), median, mean, third quartile (Q3), and maximum (Max) values.

Year	ESG					
	Min	Q1	Median	Mean	Q3	Max
2014	0	32	41	43.8	53	86
2015	0	32	40	43.6	52	88
2016	17	30	38	41.3	49	89
2017	15	29	37	40.9	49.3	89
2018	7	26	35	38.0	46	87
2019	3	18	29	31.8	42	90
2020	3	11	21.7	24.1	31	91
2021	0	12	23	24.7	31	91
2022	0	15	24	26.7	34	89

Year	Environmental					
	Min	Q1	Median	Mean	Q3	Max
2014	0	16	30	34.0	48	94
2015	0	19.9	32	35.8	50	94
2016	2	19	31	35.2	48	93
2017	0	20	31	36.0	50.7	96
2018	0	14.2	27	32.3	45	98
2019	0	10	23.7	28.6	43	98
2020	0	1	18	20.8	28	98
2021	0	1	18	20.9	28	98
2022	0	5	18	22.2	31	98

Year	Social					
	Min	Q1	Median	Mean	Q3	Max
2014	10	26	35	37.8	46	86
2015	9	24	32	36.1	44	91
2016	6	22	29	34.3	42	92
2017	0	20	28	33.1	43	92
2018	0	18.5	28	32.0	41.8	91
2019	0	10	22	25.1	36	90
2020	0	4	16	17.8	25	92
2021	0	5	16	18.1	25	92
2022	0	9	18.9	21.1	29	87

Year	Governance					
	Min	Q1	Median	Mean	Q3	Max
2014	0	46	54	55.4	63	90
2015	0	46	54	55.1	63	88
2016	26	42	49.4	51.4	58.4	91
2017	0	40	48	50.1	56	91
2018	0	35	43	45.1	52	87
2019	5	27	37	38.8	48	88
2020	5	20	27.6	30.1	36	88
2021	6	21	29	31.2	37	88
2022	6	23	31	33.5	40	89

4.3 Filtered and Aligned Data

To conduct the mean-variance efficiency and mean-variance spanning tests, we align the returns data from CRSP and the ESG data from Trucost into matrices with identical $T \times N$ dimensions, where T = 2266 daily observations from 2 January September 2014 to 31 December 2022, and N denotes the number of stocks. We include only stocks with available ESG ratings from Trucost and return data from CRSP. Table 4 provides an overview of the selected stocks.

Table 4: Evolution of ESG-Rated Stocks After Data Filtration
This table presents the number of stocks with both daily returns and ESG ratings available after filtering for missing or incomplete values in the Trucost ESG database and CRSP return dataset from 2014 to 2022.

Year	2014	2015	2016	2017	2018	2019	2020	2021	2022
Size of ESG data	389	453	493	538	584	705	1055	1892	1575
Size of Environmental data	389	453	493	538	584	705	1055	1892	1575
Size of Social data	389	453	493	538	584	705	1055	1892	1575
Size of Governance data	389	453	493	538	584	705	1055	1892	1575

4.4 Systematic Risk Factors and Climate Concerns

For our research, we incorporate exposure to the Media Climate Change Concerns Index (MCCC) to perform mean-variance spanning (MVS) and mean-variance efficiency (MVE) test developed by Ardia et al. (2023) sourced from the Sentometrics Research website[7] which captures media attention on climate change and its potential influence on market dynamics. We regress stocks return on the MCCC and the Fama and French (2015) five-factor mimicking portfolios, available on Kenneth French's website.[8] These factors—market excess return, size (SMB), value (HML), profitability (RMW), and investment (CMA)— account for traditional systematic risks. Afterward, we utilize the MCCC index to rank stocks based on their exposure to this to this climate-related risk measure. Figure 1 shows the index's daily evolution from 2014 to 2022.

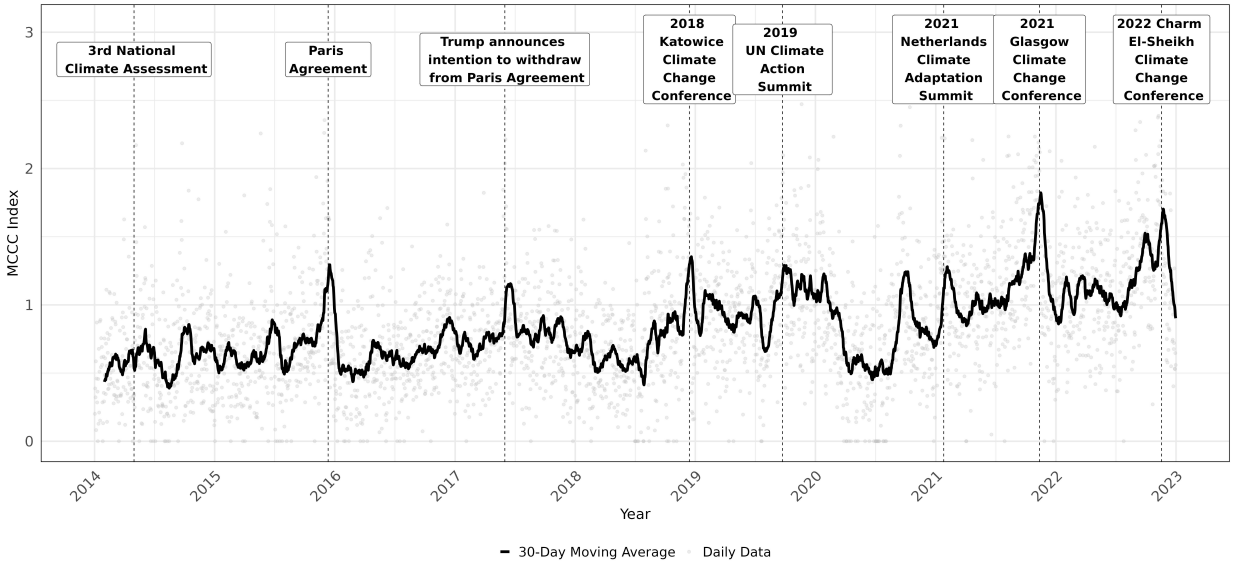

Figure 1: Trends in Media Climate Change Concerns (MCCC) Index With Key Climate Events

This figure shows the daily Media Climate Change Concerns (MCCC) index (gray points) with its 30-day moving average (black line) from January 2014 to December 2022. Key climate events, including notable conferences and policy announcements, are highlighted (in boxes).

[7] Sentometrics Research: MCCC Data, November 14, 2020.

[8] Kenneth R. French: Fama/French 5 Factors (2x3)

5 Empirical Application

This section is divided into two main parts. In one part, we conduct mean-variance efficiency (MVE) and mean-variance spanning (MVS) tests using ESG scores. We also examine the efficiency and spanning capacity of ESG as well as Environmental (E), Social (S), and Governance (G) stock groups across different sizes. In the other part, we test the same hypotheses using the exposure to the Media Climate Change Concerns (MCCC) index as an alternative dynamic measure to evaluate sustainability-driven stock selection.

5.1 Hypothesis Testing Using ESG Score

We construct our ESG universes of stocks from a forward-looking perspective. At the start of each year, we rank stocks based on their ESG (or E, S, and G) scores and form two main groups: the benchmark universe, which consists of the top K ESG-rated stocks, and the test universe, which include the following top N stocks. We hold the groups for one year, then re-form them based on updated ESG rankings. This strategy incorporates all available information at a given time and avoids future data or forecasts.

Using the framework of Gungor and Luger (2016), we perform mean-variance efficiency (MVE) and spanning (MVS) tests to evaluate the performance of ESG stock groups. MVE tests assess whether the top K ESG stocks offer optimal risk-adjusted returns, while MVS, examines whether adding the next group of stocks, the top N ESG stocks, improves the risk-return trade-off of the benchmark universe. We examine various stock group configurations for K = 10, 20, 30, and N = 10, 20, 30 to evaluate the ability of top ESG universes to "span" the subsequent top N stocks. The actual number of stocks in the benchmark (K) and test (N) groups may exceed the target values due to tied ESG scores at selection thresholds. In such cases, all stocks with the same score are included, resulting in larger group sizes. Table 5 captures this consideration, with each group's actual number of stocks adjusted to account for tied ESG scores in the top K and top N selections. The adopted methodology accommodates hypothesis testing even when

the number of stocks exceeds the number of days T – cases where $K + N > 252$.

Table 5: The Size of the Groups Ranked Based on ESG Scores
This table presents the composition of a benchmark universe (K) that consists of the top ESG stocks and test universe (N) that includes the following top ESG stocks in the context of mean-variance spanning.

Year	K	N			K	N			K	N		
	10	10	20	30	20	10	20	30	30	10	20	30
2014	33	46	95	177	79	49	131	216	128	82	167	254
2015	33	42	89	181	75	47	139	270	122	92	223	320
2016	22	30	82	156	52	52	126	264	104	74	212	340
2017	15	35	77	132	52	42	97	187	92	55	145	236
2018	20	42	87	138	62	45	96	200	107	51	155	307
2019	22	38	90	178	60	52	140	299	112	88	247	425
2020	16	25	113	224	41	88	199	338	129	111	250	375
2021	20	24	64	116	44	40	92	186	84	52	146	244
2022	19	33	76	164	52	43	131	226	95	88	183	310

The sizes of the benchmark and the test group also differ across the Environmental (E), Social (S), and Governance(G) dimensions. Tables A.1, A.2, and A.3 in the Appendix show these sizes.

5.1.1 ESG Mean-Variance Efficiency Test

We apply the non-parametric test developed by Gungor and Luger (2016) as outlined in Section 3.2 for mean-variance efficiency with $F_{\max}$ statistics described in (20) using $M = 500$ Monte Carlo simulations, so the smallest possible MC p-value is 0.2%. We perform the MC test at the nominal $\alpha = 5\%$ significance level.

Table 6 summarizes the results of the mean-variance efficiency tests. Most results across all panels (K = 10, 20, and 30) do not reject the MVE hypothesis and indicate that the top ESG-rated benchmark universes are mean-variance efficient across various test sets (N = 10, 20, and 30) for most of the years from 2014 to 2022. A few inconclusive outcomes in 2017 may reflect the effect of the Paris Agreement in 2016. The absence of rejections, combined with the Liberal Monte-Carlo (LMC) p-values remaining above the significance threshold for most cases, suggest that the top ESG-rated stock groups optimally balance risk and return. These universes of stocks cover the efficient universe space, with no evidence of superior combinations among the test sets. Therefore, the mean-variance efficiency test results validate the hypothesis that

ESG-focused groups formed using top-rated stocks are efficient in the risk-return tradeoff. This finding supports the idea that integrating ESG criteria does not compromise performance.

Table 6: ESG Mean-Variance Efficiency Test Results: Gungor and Luger (2016) Test
This table presents results for three benchmark stock sets (K = 10, 20, 30) and test stocks (N = 10, 20, 30) from 2014 to 2022. Key metrics include the F_{max} statistic, BMC p-value, LMC p-value, and final decision on mean-variance efficiency. With α = 5%, the conservative MC p-value is reported if $\tilde{P}_M^C(F_{max}(Y)) \leq 5\%$, the liberal MC p-value if $\tilde{P}_M^L(F_{max}(Y)) > 5\%$, and both if inconclusive. The symbol "-" is used whenever the p-values are not reported. Decisions are denoted by (✓) for "Do not reject," (✗) for "Reject," and (**?**) for "Inconclusive," based on p-values.

Year	K = 10											
	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.013	-	0.970	✓	0.036	-	0.322	✓	0.045	-	0.250	✓
2015	0.030	-	0.310	✓	0.031	-	0.502	✓	0.031	-	0.750	✓
2016	0.009	-	0.992	✓	0.040	-	0.156	✓	0.040	-	0.306	✓
2017	0.017	-	0.754	✓	0.022	-	0.766	✓	0.056	1.000	0.036	**?**
2018	0.017	-	0.862	✓	0.043	-	0.142	✓	0.043	-	0.200	✓
2019	0.019	-	0.730	✓	0.019	-	0.960	✓	0.019	-	0.730	✓
2020	0.010	-	0.960	✓	0.010	-	1.000	✓	0.010	-	1.000	✓
2021	0.012	-	0.922	✓	0.026	-	0.608	✓	0.026	-	0.608	✓
2022	0.014	-	0.886	✓	0.014	-	0.992	✓	0.014	-	0.992	✓
Year	K = 20											
	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.028	-	0.754	✓	0.054	-	0.284	✓	0.054	-	0.410	✓
2015	0.026	-	0.784	✓	0.026	-	0.984	✓	0.034	-	0.974	✓
2016	0.039	-	0.244	✓	0.039	-	0.504	✓	0.039	-	0.722	✓
2017	0.029	-	0.572	✓	0.029	-	0.796	✓	0.054	-	0.222	✓
2018	0.043	-	0.184	✓	0.043	-	0.338	✓	0.043	-	0.618	✓
2019	0.013	-	0.998	✓	0.019	-	1.000	✓	0.021	-	1.000	✓
2020	0.016	-	0.990	✓	0.016	-	1.000	✓	0.026	-	0.994	✓
2021	0.019	-	0.830	✓	0.022	-	0.944	✓	0.025	-	0.980	✓
2022	0.014	-	0.980	✓	0.014	-	1.000	✓	0.019	-	1.000	✓
Year	K = 30											
	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.034	-	0.980	✓	0.034	-	1.000	✓	0.068	-	0.654	✓
2015	0.036	-	0.958	✓	0.050	-	0.872	✓	0.052	-	0.942	✓
2016	0.030	-	0.940	✓	0.047	-	0.884	✓	0.047	-	0.950	✓
2017	0.016	-	0.998	✓	0.094	1.000	0.032	**?**	0.094	1.000	0.044	**?**
2018	0.017	-	0.998	✓	0.030	-	1.000	✓	0.031	-	1.000	✓
2019	0.023	-	0.996	✓	0.049	-	0.882	✓	0.049	-	0.960	✓
2020	0.038	-	0.938	✓	0.038	-	0.994	✓	0.055	-	0.916	✓
2021	0.020	-	0.966	✓	0.035	-	0.870	✓	0.035	-	0.956	✓
2022	0.025	-	0.992	✓	0.025	-	1.000	✓	0.029	-	1.000	✓

5.1.2 ESG Mean-Variance Spanning Test

We test the mean-variance spanning (MVS) hypothesis using the non-parametric F_{max} test by Gungor and Luger (2016) to assess whether benchmark stocks span the test universe for different

values of K and N. The goal is to determine whether certain benchmark universes efficiently represent the risk-return tradeoff of a broader set of risky stocks.

Table 7 displays the results of the MVS test. The test rejects the spanning hypothesis in approximately 70% of the cases. For the smallest benchmark universe (K = 10), spanning outcomes are mostly inconclusive when tested against different N test universes, with only one non-rejecting decision per N. Expanding the benchmark universe to K = 20 increases the frequency of non-rejection spanning, particularly for $N = 10$ and N = 20. Notably, we observe fewer inconclusive results during 2016–2018 (post-Paris Agreement) and 2020 (COVID-19 crisis). When the benchmark size increases to K = 30, spanning outcomes improve further, with non-rejections of the MVS hypothesis observed more consistently and fewer inconclusive outcomes. Notably, for N = 10, the hypothesis is not rejected across all observed years, whereas for larger N, it is not the case. While expanding to a larger group of stocks enhances spanning initially, excessively including lower-ranked ESG stocks dilutes efficiency and does not improve the universe's performance.

5.1.3 Tests for Environmental, Social, and Governance Ratings

We examine whether focusing on top-rated Environmental (E), Social (S), or Governance (G) scores brings different results compared to ESG for mean-variance efficiency (MVE) and mean-variance spanning (MVS). Tables A.4 and A.5 in Appendix I present the Environmental (E) results. The MVE tests for E scores consistently show non-rejection, similar to the outcomes for overall ESG scores. However, the MVS test results in rejections approximately two-thirds of the time, slightly less frequent than for comprehensive ESG scores. Notably, when K = 30, we observe more non-rejection decisions, particularly for cases involving N = 10 and also N = 20. In 2017, the MVS hypothesis is rejected for K = 10 with N = 10 and N = 20. This decision is likely influenced by the 2016 Paris Agreement on climate change, as it logically influences Environmental scores. Tables A.6 and A.7 summarize the hypothesis testing results for the Social (S) dimension. The MVE and MVS results for Social ratings are similar to those for overall ESG, with mostly inconclusive outcomes and some non-rejection of the spanning hypothesis for larger

N and K values. Tables A.8 and A.9 present the results for the Governance (G) dimension. This dimension also displays higher non-rejection rates, similar to the Environmental (E) scores, with no noteworthy deviations from the ESG outcomes.

Table 7: ESG Mean-Variance Spanning Test Results: Gungor and Luger (2016) Test
This table presents results for three benchmark stock sets (K = 10, 20, 30) and test stocks (N = 10, 20, 30) from 2014 to 2022. Key metrics include the F_{max} statistic, BMC p-value, LMC p-value, and final decision on mean-variance efficiency. With α = 5%, the conservative MC p-value is reported if $\tilde{P}^{C}_{M}(F_{max}(Y)) \leq 5\%$, the liberal MC p-value if $\tilde{P}^{L}_{M}(F_{max}(Y)) > 5\%$, and both if inconclusive. The symbol "-" is used whenever the p-values are not reported. Decisions are denoted by (✓) for "Do not reject," (✗) for "Reject," and (**?**) for "Inconclusive," based on p-values.

	K = 10											
Year	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.070	1.000	0.028	?	0.070	-	0.082	✓	0.070	-	0.138	✓
2015	0.118	1.000	0.002	?	0.118	1.000	0.002	?	0.198	1.000	0.002	?
2016	0.084	1.000	0.012	?	0.086	1.000	0.018	?	0.103	1.000	0.002	?
2017	0.192	0.180	0.002	?	0.192	0.180	0.002	?	0.212	0.134	0.002	?
2018	0.110	1.000	0.002	?	0.110	1.000	0.002	?	0.110	1.000	0.002	?
2019	0.081	1.000	0.004	?	0.081	1.000	0.014	?	0.081	1.000	0.003	?
2020	0.050	-	0.352	✓	0.121	1.000	0.034	?	0.138	1.000	0.036	?
2021	0.060	1.000	0.042	?	0.091	1.000	0.002	?	0.091	1.000	0.002	?
2022	0.067	1.000	0.018	?	0.091	1.000	0.006	?	0.161	0.998	0.002	?
	K = 20											
Year	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.035	-	0.944	✓	0.056	-	0.728	✓	0.121	1.000	0.012	?
2015	0.081	-	0.054	✓	0.096	0.126	0.048	?	0.126	1.000	0.018	?
2016	0.117	1.000	0.002	?	0.126	1.000	0.002	?	0.126	1.000	0.002	?
2017	0.099	1.000	0.004	?	0.099	1.000	0.010	?	0.133	1.000	0.004	?
2018	0.117	1.000	0.006	?	0.117	1.000	0.012	?	0.126	1.000	0.126	?
2019	0.039	-	0.798	✓	0.039	-	0.982	✓	0.169	1.000	0.002	?
2020	0.104	1.000	0.026	?	0.119	1.000	0.034	?	0.3	1.000	0.003	?
2021	0.078	-	0.102	✓	0.078	-	0.056	✓	0.078	-	0.102	✓
2022	0.039	-	0.620	✓	0.124	1.000	0.002	?	0.83	0.004	-	✗
	K = 30											
Year	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.077	-	0.512	✓	0.154	1.000	0.026	?	0.154	1.000	0.044	?
2015	0.120	-	0.060	✓	0.132	1.000	0.046	?	0.132	-	0.066	✓
2016	0.073	-	0.358	✓	0.124	1.000	0.040	?	0.124	-	0.052	✓
2017	0.043	-	0.852	✓	0.157	1.000	0.004	?	0.157	1.000	0.004	?
2018	0.067	-	0.354	✓	0.126	1.000	0.040	?	0.126	-	0.072	✓
2019	0.098	-	0.144	✓	0.209	1.000	0.006	?	0.209	1.000	0.006	?
2020	0.086	-	0.490	✓	0.154	1.000	0.024	?	0.154	1.000	0.042	?
2021	0.057	-	0.370	✓	0.111	1.000	0.028	?	0.111	1.000	0.042	?
2022	0.101	-	0.056	✓	0.771	1.000	0.002	?	0.771	1.000	0.002	?

In summary, MVE evaluates the standalone performance of the top K ESG-rated stocks and consistently shows non-rejection across most years, indicating that these stocks independently achieve optimal risk-return tradeoffs. In contrast, MVS examines whether adding test stocks enhances the risk-return frontier. Frequent rejections suggest insufficient diversification for smaller K values (K = 10 or 20). However, as K increases to 30, non-rejection rates improve significantly, demonstrating that larger benchmark groups better span the investment universe's efficient frontier. Adding lower-ranked ESG stocks beyond the top 30 can dilute efficiency and yield less conclusive results. Overall, focusing on the top-rated ESG stocks, particularly the top 30, achieves a better balance of efficiency and market representation.

5.2 Hypothesis Testing Using MCCC Exposure

We run time series regressions for all stocks each year using the five factors from Fama and French (2015) along with the MCCC index as explanatory variables, as outlined in Section 3.3. We extract and store the coefficients associated with the MCCC index to measure the stocks' sensitivity to media sentiment related to climate change. Using this sensitivity, we construct the benchmark and test universes of stocks following the same methodology applied to ESG rankings. At the start of each year, we rank stocks based on their MCCC exposures to form benchmark universes of the top K stocks and test universes comprising the next top N. Unlike the ESG-based rankings, the MCCC approach precisely matches the K and N specifications (e.g., K = 10 results in precisely ten stocks) due to the variability in stock exposures to the MCCC index.

5.3 MCCC Mean-Variance Efficiency Test

Unlike the consistent non-rejection observed in the MVE hypothesis testing for ESG (or E, S, and G) scores, MCCC results showcase inconclusive decisions in more than half of the cases. We observe consistency in the values and the decisions within each benchmark size K regardless of the sizes of the test stocks N. Rejections are present for 2017, 2018, and 2019, while 2016 saw

full non-rejection. We observe partial non-rejection in 2022 for K = 20 and K = 30. The MCCC index, susceptible to short-term market sentiment, proves less effective than ESG in providing a robust risk-return tradeoff.

Table 8: MCCC Mean-Variance Efficiency Test Results: Gungor and Luger (2016) Test
This table presents results for three benchmark stock sets (K = 10, 20, 30) and test stocks (N = 10, 20, 30) from 2014 to 2022. Key metrics include the F_{max} statistic, BMC p-value, LMC p-value, and final decision on mean-variance efficiency. With α = 5%, the conservative MC p-value is reported if $\tilde{P}_M^C(F_{max}(Y)) \leq 5\%$, the liberal MC p-value if $\tilde{P}_M^L(F_{max}(Y)) > 5\%$, and both if inconclusive. The symbol "-" is used whenever the p-values are not reported. Decisions are denoted by (✓) for "Do not reject," (✗) for "Reject," and (**?**) for "Inconclusive," based on p-values.

	K = 10											
Year	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.039	0.042	0.150	?	0.039	0.042	0.150	?	0.039	0.042	0.150	?
2015	0.120	0.002	0.052	?	0.120	0.002	0.052	?	0.120	0.002	0.052	?
2016	0.021	-	0.422	✓	0.021	-	0.422	✓	0.021	-	0.422	✓
2017	0.091	0.010	-	✗	0.091	0.010	-	✗	0.091	0.010	-	✗
2018	0.008	-	0.882	✓	0.008	-	0.882	✓	0.008	-	0.882	✓
2019	0.010	0.010	0.678	?	0.010	0.010	0.678	?	0.010	0.010	0.678	?
2020	0.023	0.016	0.33	?	0.023	0.016	0.33	?	0.023	0.016	0.33	?
2021	0.009	0.002	0.728	?	0.009	0.002	0.728	?	0.009	0.002	0.728	?
2022	0.016	0.002	0.282	?	0.0158	0.002	0.282	?	0.0158	0.002	0.282	?
	K = 20											
Year	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.011	0.004	0.936	?	0.011	0.004	0.936	?	0.011	0.004	0.936	?
2015	0.041	-	0.402	✓	0.041	-	0.402	✓	0.041	-	0.402	✓
2016	0.015	-	0.896	✓	0.015	-	0.896	✓	0.015	-	0.896	✓
2017	0.111	0.008	-	✗	0.111	0.008	-	✗	0.111	0.008	-	✗
2018	0.016	0.004	0.834	?	0.016	0.004	0.834	?	0.016	0.004	0.834	?
2019	0.199	0.004	-	✗	0.038	0.004	0.304	✗	0.038	0.004	0.304	✗
2020	0.032	0.002	0.268	?	0.032	0.002	0.268	?	0.032	0.002	0.268	?
2021	0.011	0.006	0.914	?	0.011	0.006	0.914	?	0.011	0.006	0.914	?
2022	0.028	-	0.31	✓	0.028	-	0.31	✓	0.028	-	0.31	✓
	K = 30											
Year	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.015	0.004	0.926	?	0.015	0.004	0.926	?	0.015	0.004	0.926	?
2015	0.044	0.568	0.008	?	0.044	0.568	0.008	?	0.044	0.568	0.008	?
2016	0.035	-	0.684	✓	0.035	-	0.684	✓	0.035	-	0.684	✓
2017	0.031	0.040	0.706	?	0.031	0.040	0.706	?	0.031	0.040	0.706	?
2018	0.089	0.004	-	✗	0.089	0.004	-	✗	0.089	0.004	-	✗
2019	0.038	0.004	0.304	?	0.038	0.004	0.304	?	0.038	0.004	0.304	?
2020	0.019	-	0.808	✓	0.019	-	0.808	✓	0.019	-	0.808	✓
2021	0.019	0.002	0.844	?	0.019	0.002	0.844	?	0.019	0.002	0.844	?
2022	0.032	-	0.386	✓	0.032	-	0.386	✓	0.032	-	0.386	✓

5.4 MCCC Mean-Variance Spanning Test

The MVS test results for MCCC exposure highlight consistent decisions across different N values within each benchmark size K. Overall, MCCC results show a higher frequency of rejections compared to ESG or individual E, S, and G scores. The MCCC-based groups of stocks often fail to span the test stocks. We highlight a notable non-rejection of the spanning hypothesis 2016 for K = 10 and K = 20, a finding that aligns with the mean-variance efficiency hypothesis. In 2020, non-rejection decisions were present for K = 20 and K = 30. However, the spanning test outcomes for 2014, 2015 and 2017 remain inconclusive.

In summary, the MCCC hypothesis testing results are less consistent than ESG, with more frequent rejections and inconclusive outcomes in MVE and MVS tests. MVE results show frequent standalone inefficiencies for benchmark stocks, except in 2016 and 2022, where larger benchmark sizes achieved partial non-rejection of the hypothesis. MVS results, however, exhibit even higher rejection rates, particularly for smaller K, indicating that MCCC benchmarks fail to span the test stocks effectively. Overall, sustainable investors should prioritize ESG ratings as the primary sustainable criterion for stock selection when making yearly investment decisions. We suggest taking the MCCC index as a supplementary tool to monitor short-term sentiment shifts and assess climate-related risks. Still, it should not replace ESG scores in strategic investment decisions.

Table 9: MCCC Mean-Variance Spanning Test Results: Gungor and Luger (2016) Test
This table presents results for three benchmark stock sets (K = 10, 20, 30) and test stocks (N = 10, 20, 30) from 2014 to 2022. Key metrics include the F_{max} statistic, BMC p-value, LMC p-value, and final decision on mean-variance efficiency. With α = 5%, the conservative MC p-value is reported if $\tilde{P}_M^C(F_{max}(Y)) \leq 5\%$, the liberal MC p-value if $\tilde{P}_M^L(F_{max}(Y)) > 5\%$, and both if inconclusive. The symbol "-" is used whenever the p-values are not reported. Decisions are denoted by (✓) for "Do not reject," (✗) for "Reject," and (**?**) for "Inconclusive," based on p-values.

Year	K = 10											
	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.048	0.028	0.186	**?**	0.048	0.028	0.186	**?**	0.048	0.028	0.186	**?**
2015	0.103	0.002	0.604	**?**	0.103	0.002	0.604	**?**	0.103	0.002	0.604	**?**
2016	0.069	-	0.452	✓	0.069	-	0.452	✓	0.069	-	0.452	✓
2017	0.091	0.010	0.616	**?**	0.091	0.010	0.616	**?**	0.091	0.010	0.616	**?**
2018	0.048	-	0.716	✓	0.048	-	0.716	✓	0.048	-	0.716	✓
2019	0.068	0.004	0.600	**?**	0.068	0.004	0.600	**?**	0.068	0.004	0.600	**?**
2020	0.112	0.002	0.072	**?**	0.112	0.002	0.072	**?**	0.112	0.002	0.072	**?**
2021	0.075	0.002	0.126	**?**	0.075	0.002	0.126	**?**	0.075	0.002	0.126	**?**
2022	0.089	0.002	0.282	**?**	0.089	0.002	0.282	**?**	0.089	0.002	0.282	**?**
Year	K = 20											
	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.088	0.002	0.380	**?**	0.088	0.002	0.380	**?**	0.088	0.002	0.380	**?**
2015	0.137	0.014	0.064	**?**	0.137	0.014	0.064	**?**	0.137	0.014	0.064	**?**
2016	0.069	0.660	-	✓	0.069	0.660	-	✓	0.069	0.660	-	✓
2017	0.111	0.004	0.190	**?**	0.111	0.004	0.190	**?**	0.111	0.004	0.190	**?**
2018	0.092	0.002	0.528	**?**	0.092	0.002	0.528	**?**	0.092	0.002	0.528	**?**
2019	0.303	0.002	-	✗	0.303	0.002	-	✗	0.303	0.002	-	✗
2020	0.077	0.002	0.536	✓	0.077	0.002	0.536	✓	0.077	0.002	0.536	✓
2021	0.075	0.002	0.290	**?**	0.075	0.002	0.290	**?**	0.075	0.002	0.290	**?**
2022	0.171	0.028	-	✗	0.171	0.028	-	✗	0.171	0.028	-	✗
Year	K = 30											
	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.074	0.002	0.454	**?**	0.074	0.002	0.454	**?**	0.074	0.002	0.454	**?**
2015	0.098	0.002	0.356	**?**	0.098	0.002	0.356	**?**	0.098	0.002	0.356	**?**
2016	0.081	0.044	0.280	**?**	0.081	0.044	0.280	**?**	0.081	0.044	0.280	**?**
2017	0.094	0.016	0.416	**?**	0.094	0.016	0.416	**?**	0.094	0.016	0.416	**?**
2018	0.064	0.010	0.786	**?**	0.064	0.010	0.786	**?**	0.064	0.010	0.786	**?**
2019	0.053	0.020	0.816	**?**	0.053	0.020	0.816	**?**	0.053	0.020	0.816	**?**
2020	0.029	-	0.900	✓	0.029	-	0.900	✓	0.029	-	0.900	✓
2021	0.136	0.002	-	✗	0.136	0.002	-	✗	0.136	0.002	-	✗
2022	0.058	-	0.420	✓	0.058	-	0.420	✓	0.058	-	0.420	✓

6 Conclusion

Traditional methods for analyzing mean-variance efficiency and spanning are constrained by assumptions of independent and identically distributed disturbances, multivariate normality, and limited scalability to large stock universes. This research adopts the advanced framework of

Gungor and Luger (2016), which eliminates restrictions on the MLR residual distributions and operates equation by equation, so it remains applicable to any stock size. We evaluate whether socially responsible stocks optimize the risk-return tradeoff and span the broader investment universe. We apply this approach to different groups of stocks, using the comprehensive ESG ratings, the E, S, and G dimensions, and the exposure to the MCCC index as an alternative measure.

We find that the top 30 ESG-rated stocks are mean-variance efficient and adequately span the investment universe. However, expanding these universes to include lower-rated ESG stocks diminishes efficiency, so diversification beyond top-rated stocks does not improve risk-return tradeoffs. Although the MCCC index helps capture short-term market sentiment, it is less useful for long-term investment strategies due to its higher volatility and less reliable efficiency results than ESG ratings. ESG ratings, with their comprehensive yearly assessments, provide a more reliable foundation for long-term sustainable stock selection strategies.

This research adds to the expanding literature on sustainable investing and practical investment strategies by offering empirical evidence on the efficiency and spanning capabilities of ESG-focused stock groups. A limitation is that it depends solely on annual ESG ratings, which may overlook the dynamic changes in businesses' sustainability practices throughout the year. Future research could extend the U.S. analysis to global markets to further explore sustainable investment opportunities.

References

Alexander, K., Osthoff, P., 2007. The effect of socially responsible investing on portfolio performance. European Financial Management 13, 908–922.

Ardia, D., Bluteau, K., Boudt, K., Inghelbrecht, K., 2023. Climate change concerns and the performance of green vs. brown stocks. Management Science 69, 7607–7632.

Barnard, G., 1963. Comment on 'the spectral analysis of point processes' by m.s. bartlett. Journal of the Royal Statistical Society (Series B) 25, 294.

Barth, F., Eckert, C., Gatzert, N., Scholz, H., 2022. Spillover effects from the Volkswagen emissions scandal: An analysis of stock, corporate bond, and credit default swap markets. Schmalenbach Journal of Business Research 74, 37–76.

Beaulieu, M.C., Dufour, J.M., Khalaf, L., 2007. Multivariate tests of mean-variance efficiency with possibly non-Gaussian errors: An exact simulation-based approach. Journal of Business & Economic Statistics 25, 398–410.

Berg, F., Kölbel, J.F., Rigobon, R., 2022. Aggregate confusion: The divergence of ESG ratings. Review of Finance 26, 1315–1344.

Billou, N., 2004. Tests of the CAPM and Fama and French three-factor model.

Birnbaum, Z., 1974. Computers and unconventional test statistics, in: Proschan, F., Serfling, R. (Eds.), Reliability and Biometry. SIAM, Philadelphia, pp. 441–458.

Bolton, P., Kacperczyk, M., 2021. Do investors care about carbon risk? Journal of Financial Economics 142, 517–549.

Bolton, P., Kacperczyk, M., 2023. Global pricing of carbon-transition risk. Journal of Finance 78, 3677–3754.

Briscoe-Tran, H., Meier, I., Elabd, R., Sokolovski, V., 2024. Social premiums. Working paper.

Brown, G., Peterson, R.S., 2022. The Imbalanced board: Google. Springer International Publishing, Cham.

Chava, S., 2010. Socially responsible investing and expected stock returns. Working paper.

Dimson, E., Marsh, P., Staunton, M., 2020. Divergent ESG ratings. Journal of Portfolio Management 47, 75–87.

Dufour, J.M., 2006. Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics. Journal of Econometrics 133, 443–477.

Durbin, J., Watson, G.S., 1950. Testing for serial correlation in least squares regression: I. Biometrika 37, 409–428.

Durbin, J., Watson, G.S., 1951. Testing for serial correlation in least squares regression: II. Biometrika 38, 159–178.

Dwass, M., 1957. Modified randomization tests for nonparametric hypotheses. Annals of Mathematical Statistics 28, 181–187.

Fama, E.F., French, K.R., 2015. A five-factor asset pricing model. Journal of Financial Economics 116, 1–22.

Gibbons, M.R., Ross, S., Shanken, J., 1989. A test of the efficiency of a given portfolio. Econometrica 57, 1121–52.

Gompers, P.A., Ishii, J.L., Metrick, A., 2003. Corporate governance and equity prices. Quarterly Journal of Economics 118, 107–155.

Griggs, J.W., 2011. BP gulf of Mexico oil spill. Energy Law Journal 32, 14–57.

Gungor, S., Luger, R., 2009. Exact distribution-free tests of mean-variance efficiency. Journal of Empirical Finance 16, 816–829.

Gungor, S., Luger, R., 2013. Testing linear factor pricing models with large cross-sections: a distribution-free approach. Journal of Business and Economic Statistics , 66–77.

Gungor, S., Luger, R., 2016. Multivariate tests of mean-variance efficiency and spanning with a large number of assets and time-varying covariances. Journal of Business & Economic Statistics 34, 161–175.

Gunnar, F., Busch, T., Bassen, A., 2015. ESG and financial performance: aggregated evidence from more than 2000 empirical studies. Journal of Sustainable Finance & Investment 5, 210–233.

Harisson, H., Kacperczyk, M., 2009. The price of sin: The effect of social norms on markets. Journal of Financial Economics 93, 15–36.

Huberman, G., Kandel, S., 1987. Mean-variance spanning. Journal of Finance 42, 873–88.

Kan, R., Zhou, G., 2012. Tests of mean-variance spanning. Annals of Economics and Finance 13, 139–187.

Krueger, P., Sautner, Z., Starks, L., 2020. The importance of climate risks for institutional investors. Review of Financial Studies 33, 1067–1111.

Lintner, J., 1965. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. Review of Economics and Statistics 47, 13–37.

Madhavan, A., Sobczyk, A., Ang, A., 2021. Toward ESG alpha: Analyzing ESG exposures through a factor lens. Financial Analysts Journal 77, 69–88.

Markowitz, H., 1952. Portfolio selection. Journal of Finance 7, 77–91.

M.Cahart, M., 1997. On persistence in mutual fund performance. Journal of Finance 52, 57–82.

Mossin, J., 1966. Equilibrium in a capital asset market. Econometrica 34, 768–783.

Nima, B., 2011. After the spill is gone: The gulf of Mexico, environmental crime, and criminal law. Michigan Law Review 109, 1413–1461.

Pastor, L., Stambaugh, R., Taylor, L.A., 2022. Dissecting green returns. Journal of Financial Economics 146, 403–424.

Pastor, L., Stambaugh, R.F., Taylor, L.A., 2021. Sustainable investing in equilibrium. Journal of Financial Economics 142, 550–571.

Pedersen, L.H., Fitzgibbons, S., Pomorski, L., 2020. Responsible investing: The ESG-efficient frontier. Journal of Financial Economics 142, 572–597.

Pesaran, M.H., Yamagata, T., 2012. Testing CAPM with a large number of assets. AFA 2013 San Diego Meetings Paper.

Randles, R., Wolfe, D., 1979. Introduction to the Theory of Nonparametric Statistics. John Wiley & Sons, New York.

Rob, B., Jeroen, D., Rogér, O., 2007. The ethical mutual fund performance debate: new evidence from Canada. j. bus. ethics. Journal of Business Ethics 70, 111–124.

Roll, R., 1977. A critique of the asset pricing theory's tests part I: On past and potential testability of the theory. Journal of Financial Economics 4, 129–176.

Sharpe, W.F., 1964. Capital asset prices: A theory of market equilibrium under conditions of risk. Journal of Finance 19, 425–442.

Statman, M., 2006. Socially responsible indexes. Journal of Portfolio Management 32, 100–109.

Treynor, J.L., 1999. Toward a theory of market value of risky assets, in: Korajczyk, R.A. (Ed.), Asset Pricing and Portfolio Performance. Risk Books, London, pp. 15–22. Unpublished Manuscript, Final Version.

Widyawati, L., 2020. A systematic literature review of socially responsible investment and environmental social governance metrics. Business Strategy and the Environment 29, 619–637.

Appendix

I Environmental, Social, and Governance Tests

Table A.1: The Size of the Groups Ranked Based on Environmental Scores
This table presents the composition of benchmark stocks (K) that consist of the top Environmental stocks and test stocks (N) that include the following top Environmental stocks in the context of mean-variance spanning.

Year	K	N			K	N			K	N		
	10	10	20	30	20	10	20	30	30	10	20	30
2014	19	35	74	114	54	39	79	130	123	93	91	157
2015	28	28	73	113	56	45	85	132	103	101	87	194
2016	28	32	77	131	60	45	99	173	96	105	128	222
2017	17	34	75	130	51	41	96	147	92	55	106	170
2018	20	44	82	128	64	38	84	156	102	46	118	194
2019	19	29	73	122	48	44	93	154	92	49	110	198
2020	11	17	41	81	28	24	64	148	52	40	124	327
2021	20	45	78	130	65	33	85	146	98	52	113	201
2022	24	32	77	132	56	45	100	157	101	55	112	212

Table A.2: The Size of the Groups Ranked Based on Social Scores
This table presents the composition of benchmark stocks (K) that consist of the top Social stocks and test stocks (N) that include the following top Social stocks in the context of mean-variance spanning.

Year	K	N			K	N			K	N		
	10	10	20	30	20	10	20	30	30	10	20	30
2014	22	36	68	141	58	32	105	185	90	73	153	236
2015	22	31	72	119	53	41	88	208	94	47	167	295
2016	16	28	59	108	44	31	80	149	75	49	118	236
2017	18	28	64	108	46	36	80	146	82	44	110	157
2018	17	32	78	133	49	46	101	178	95	55	132	241
2019	16	29	69	145	45	40	116	209	85	76	169	330
2020	18	36	127	222	54	91	186	365	145	95	274	330
2021	14	24	59	91	38	35	67	146	73	32	84	160
2022	16	36	75	134	52	39	98	157	91	59	185	297

Table A.3: The Size of the Groups Ranked Based on Governmental Scores

This table presents the composition of benchmark stocks (K) that consist of the top Governmental stocks and test stocks (N) that include the following top Governmental stocks in the context of mean-variance spanning.

Year	K	N			K	N			K	N		
	10	10	20	30	20	10	20	30	30	10	20	30
2014	38	52	112	239	90	60	79	130	123	93	91	157
2015	39	53	156	296	92	103	85	132	103	101	87	194
2016	25	45	139	305	70	94	99	173	96	105	128	222
2017	32	53	102	195	85	49	96	147	92	55	106	170
2018	26	51	116	225	77	65	84	156	102	46	118	194
2019	24	44	124	287	68	80	93	154	92	49	110	198
2020	18	26	109	242	44	83	64	148	52	40	124	327
2021	24	32	69	136	65	46	37	146	98	52	113	201
2022	15	41	76	168	56	56	35	157	101	55	112	212

Table A.4: Environmental Mean-Variance Efficiency Test Results: Gungor and Luger (2016) Test
This table presents results for three benchmark stock sets (K = 10, 20, 30) and test stocks (N = 10, 20, 30) from 2014 to 2022. Key metrics include the F_{max} statistic, BMC p-value, LMC p-value, and final decision on mean-variance efficiency. With α = 5%, the conservative MC p-value is reported if $\tilde{P}_M^C(F_{max}(Y)) \leq 5\%$, the liberal MC p-value if $\tilde{P}_M^L(F_{max}(Y)) > 5\%$, and both if inconclusive. The symbol "-" is used whenever the p-values are not reported. Decisions are denoted by (✓) for "Do not reject," (✗) for "Reject," and (**?**) for "Inconclusive," based on p-values.

Year	K = 10											
	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.021	-	0.612	✓	0.020	-	0.862	✓	0.045	-	0.118	✓
2015	0.009	-	0.986	✓	0.032	-	0.408	✓	0.032	-	0.572	✓
2016	0.017	-	0.806	✓	0.032	-	0.448	✓	0.032	-	0.612	✓
2017	0.015	-	0.852	✓	0.0178	-	0.942	✓	0.019	-	0.970	✓
2018	0.015	-	0.956	✓	0.0401	-	0.178	✓	0.040	-	0.248	✓
2019	0.015	-	0.830	✓	0.023	-	0.760	✓	0.023	-	0.894	✓
2020	0.006	-	0.976	✓	0.007	-	1.000	✓	0.025	-	0.678	✓
2021	0.019	-	0.774	✓	0.019	-	0.928	✓	0.019	-	0.980	✓
2022	0.018	-	0.754	✓	0.019	-	0.946	✓	0.021	-	0.966	✓
Year	K = 20											
	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.015	-	0.948	✓	0.028	-	0.788	✓	0.043	-	0.378	✓
2015	0.038	-	0.212	✓	0.038	-	0.380	✓	0.038	-	0.564	✓
2016	0.031	-	0.476	✓	0.031	-	0.780	✓	0.031	-	0.926	✓
2017	0.026	-	0.638	✓	0.034	-	0.586	✓	0.039	-	0.556	✓
2018	0.045	-	0.136	✓	0.045	-	0.246	✓	0.045	-	0.434	✓
2019	0.0154	-	0.966	✓	0.015	-	0.998	✓	0.019	-	1.000	✓
2020	0.008	-	0.992	✓	0.023	-	0.800	✓	0.023	-	0.952	✓
2021	0.027	-	0.566	✓	0.027	-	0.872	✓	0.027	-	0.972	✓
2022	0.012	-	0.998	✓	0.012	-	1.000	✓	0.015	-	1.000	✓
Year	K = 30											
	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.021	-	0.936	✓	0.049	-	0.416	✓	0.049	-	0.580	✓
2015	0.036	-	0.564	✓	0.037	-	0.826	✓	0.043	-	0.892	✓
2016	0.030	-	0.864	✓	0.035	-	0.968	✓	0.037	-	0.978	✓
2017	0.021	-	0.954	✓	0.049	-	0.480	✓	0.054	-	0.468	**?**
2018	0.026	-	0.872	✓	0.044	-	0.670	✓	0.044	-	0.842	✓
2019	0.024	-	0.910	✓	0.032	-	0.928	✓	0.032	-	0.988	✓
2020	0.023	-	0.806	✓	0.023	-	0.982	✓	0.023	-	1.00	✓
2021	0.022	-	0.968	✓	0.025	-	0.992	✓	0.046	-	0.810	✓
2022	0.023	-	0.962	✓	0.028	-	0.970	✓	0.028	-	1.00	✓

Table A.5: Environmental Mean-Variance Spanning Test Results: Gungor and Luger (2016) Test
This table presents results for three benchmark stock sets (K = 10, 20, 30) and test stocks (N = 10, 20, 30) from 2014 to 2022. Key metrics include the F_{max} statistic, BMC p-value, LMC p-value, and final decision on mean-variance efficiency. With α = 5%, the conservative MC p-value is reported if $\tilde{P}_M^C(F_{max}(Y)) \leq 5\%$, the liberal MC p-value if $\tilde{P}_M^L(F_{max}(Y)) > 5\%$, and both if inconclusive. The symbol "-" is used whenever the p-values are not reported. Decisions are denoted by (✓) for "Do not reject," (✗) for "Reject," and (**?**) for "Inconclusive," based on p-values.

Year	K = 10											
	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.090	1.000	0.008	**?**	0.129	1.000	0.004	**?**	0.129	1.000	0.004	**?**
2015	0.095	1.000	0.002	**?**	0.095	1.000	0.006	**?**	0.105	1.000	0.008	**?**
2016	0.105	1.000	0.002	**?**	0.105	1.000	0.002	**?**	0.105	1.000	0.004	**?**
2017	0.179	0.408	0.002	**?**	0.331	0.004	-	✗	0.331	0.004	-	✗
2018	0.240	0.322	0.002	**?**	0.24	0.452	0.002	**?**	0.240	0.512	0.002	**?**
2019	0.116	0.996	0.004	**?**	0.229	0.112	0.002	**?**	0.229	0.170	0.002	**?**
2020	0.106	0.998	0.016	**?**	0.192	0.784	0.004	**?**	0.192	0.916	0.006	**?**
2021	0.036	-	0.554	✓	0.078	1.000	0.022	**?**	0.102	1.000	0.002	**?**
2022	0.195	0.784	0.002	**?**	0.185	0.972	0.002	**?**	0.195	0.996	0.002	**?**
Year	K = 20											
	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.069	-	0.062	✓	0.070	-	0.118	✓	0.070	-	0.208	✓
2015	0.079	1.000	0.026	**?**	0.107	1.000	0.012	**?**	0.107	1.000	0.022	**?**
2016	0.064	-	0.150	✓	0.064	-	0.284	✓	0.091	-	0.052	✓
2017	0.115	1.000	0.004	**?**	0.115	1.000	0.004	**?**	0.119	1.000	0.008	**?**
2018	0.138	1.000	0.002	**?**	0.138	1.000	0.002	**?**	0.138	1.000	0.002	**?**
2019	0.086	1.000	0.008	**?**	0.086	1.000	0.016	**?**	0.086	1.000	0.050	**?**
2020	0.121	1.000	0.004	**?**	0.167	1.000	0.002	**?**	0.167	1.000	0.006	**?**
2021	0.1	-	0.002	**?**	0.100	-	0.012	✓	0.100	1.000	0.028	**?**
2022	0.039	-	0.620	✓	0.065	-	0.200	✓	0.065	-	0.274	✓
Year	K = 30											
	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.045	-	0.734	✓	0.059	-	0.642	✓	0.125	1.000	0.020	**?**
2015	0.066	-	0.254	✓	0.066	-	0.510	✓	0.123	1.000	0.028	**?**
2016	0.069	-	0.338	✓	0.069	-	0.618	✓	0.124	-	0.052	✓
2017	0.084	-	0.098	✓	0.084	-	0.188	✓	0.069	-	0.796	**?**
2018	0.060	-	0.416	✓	0.154	1.000	0.008	**?**	0.154	1.000	0.012	**?**
2019	0.052	-	0.532	✓	0.054	-	0.780	✓	0.098	-	0.106	✓
2020	0.122	1.000	0.010	**?**	0.122	1.000	0.030	**?**	0.098	0.106	0.002	**?**
2021	0.112	1.000	0.020	**?**	0.112	-	0.064	✓	0.218	1.000	0.002	**?**
2022	0.046	-	0.802	✓	0.760	-	0.374	✓	0.670	1.000	0.002	**?**

Table A.6: Social Mean-Variance Efficiency Test Results: Gungor and Luger (2016) Test
This table presents results for three benchmark stock sets (K = 10, 20, 30) and test stocks (N = 10, 20, 30) from 2014 to 2022. Key metrics include the F_{max} statistic, BMC p-value, LMC p-value, and final decision on mean-variance efficiency. With α = 5%, the conservative MC p-value is reported if $\tilde{P}_M^C(F_{max}(Y)) \leq 5\%$, the liberal MC p-value if $\tilde{P}_M^L(F_{max}(Y)) > 5\%$, and both if inconclusive. The symbol "-" is used whenever the p-values are not reported. Decisions are denoted by (✓) for "Do not reject," (✗) for "Reject," and (**?**) for "Inconclusive," based on p-values.

	K = 10											
Year	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.019	-	0.756	✓	0.019	-	0.910	✓	0.029	-	0.666	✓
2015	0.028	-	0.300	✓	0.028	-	0.504	✓	0.034	-	0.398	✓
2016	0.038	-	0.074	✓	0.038	-	0.156	✓	0.037	-	0.252	✓
2017	0.020	-	0.580	✓	0.021	-	0.794	✓	0.021	-	0.920	**?**
2018	0.017	-	0.764	✓	0.036	-	0.200	✓	0.036	-	0.314	✓
2019	0.024	-	0.374	✓	0.024	-	0.664	✓	0.024	-	0.888	✓
2020	0.007	-	1.000	✓	0.007	-	1.000	✓	0.019	-	1.000	✓
2021	0.009	-	0.954	✓	0.023	-	0.600	✓	0.024	-	0.736	✓
2022	0.016	-	0.860	✓	0.016	-	0.952	✓	0.021	-	0.944	✓
	K = 20											
Year	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.021	-	0.754	✓	0.047	-	0.244	✓	0.047	-	0.364	✓
2015	0.018	-	0.924	✓	0.021	-	0.964	✓	0.031	-	0.942	✓
2016	0.013	-	0.960	✓	0.015	-	1.000	✓	0.015	-	1.000	✓
2017	0.028	-	0.526	✓	0.028	-	0.782	✓	0.070	1.000	0.030	**?**
2018	0.052	1.000	0.046	**?**	0.052	-	0.098	✓	0.052	-	0.194	✓
2019	0.019	-	0.838	✓	0.019	-	0.994	✓	0.019	-	1.000	✓
2020	0.011	-	1.000	✓	0.033	-	0.864	✓	0.033	-	0.964	✓
2021	0.025	-	0.482	✓	0.025	-	0.720	✓	0.025	-	0.898	✓
2022	0.013	-	0.986	✓	0.013	-	1.000	✓	0.019	-	1.000	✓
	K = 30											
Year	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.038	-	0.594	✓	0.039	-	0.834	✓	0.046	-	0.758	✓
2015	0.021	-	0.964	✓	0.032	-	0.980	✓	0.031	-	1.000	✓
2016	0.018	-	0.964	✓	0.022	-	0.994	✓	0.035	-	0.936	✓
2017	0.017	-	0.986	✓	0.123	1.000	0.004	**?**	0.123	1.000	0.004	**?**
2018	0.019	-	0.986	✓	0.025	-	0.996	✓	0.028	-	1.000	✓
2019	0.026	-	0.928	✓	0.028	-	0.992	✓	0.035	-	0.992	✓
2020	0.041	-	0.946	✓	0.042	-	0.996	✓	0.054	-	0.974	✓
2021	0.016	-	0.950	✓	0.016	-	1.000	✓	0.016	-	1.000	✓
2022	0.022	-	0.964	✓	0.034	-	0.970	✓	0.034	-	0.994	✓

Table A.7: Social Mean-Variance Spanning Test Results: Gungor and Luger (2016) Test
This table presents results for three benchmark stock sets (K = 10, 20, 30) and test stocks (N = 10, 20, 30) from 2014 to 2022. Key metrics include the F_{max} statistic, BMC p-value, LMC p-value, and final decision on mean-variance efficiency. With α = 5%, the conservative MC p-value is reported if $\tilde{P}_M^C(F_{max}(Y)) \leq 5\%$, the liberal MC p-value if $\tilde{P}_M^L(F_{max}(Y)) > 5\%$, and both if inconclusive. The symbol "-" is used whenever the p-values are not reported. Decisions are denoted by (✓) for "Do not reject," (✗) for "Reject," and (**?**) for "Inconclusive," based on p-values.

Year	K = 10											
	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.078	1.000	0.010	?	0.080	1.000	0.014	?	0.088	1.000	0.024	?
2015	0.156	0.988	0.002	?	0.156	1.000	0.002	?	0.262	0.324	0.002	?
2016	0.180	0.362	0.002	?	0.180	0.596	0.002	?	0.207	0.468	0.002	?
2017	0.109	1.000	0.002	?	0.109	1.000	0.002	?	0.131	1.000	0.002	?
2018	0.114	0.994	0.002	?	0.114	1.000	0.002	?	0.125	1.000	0.002	?
2019	0.129	0.830	0.002	?	0.150	0.790	0.002	?	0.150	0.974	0.002	?
2020	0.255	0.858	0.002	?	0.288	0.942	0.002	?	0.606	0.006	-	✗
2021	0.043	-	0.156	✓	0.060	-	0.052	✓	0.086	1.000	0.006	?
2022	0.093	1.000	0.006	?	0.129	0.982	0.002	?	0.223	0.224	0.002	?
Year	K = 20											
	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.062	-	0.114	✓	0.062	-	0.328	✓	0.126	-	0.002	✓
2015	0.083	1.000	0.026	?	0.153	1.000	0.002	?	0.153	1.000	0.002	?
2016	0.117	1.000	0.002	?	0.117	1.000	0.002	?	0.117	-	0.002	✓
2017	0.050	-	0.288	✓	0.070	-	0.092	✓	0.156	1.000	0.002	?
2018	0.079	1.000	0.014	?	0.082	1.000	0.026	?	0.092	1.000	0.002	?
2019	0.079	1.000	0.020	?	0.079	-	0.058	✓	0.090	1.000	0.024	?
2020	0.117	1.000	0.016	?	0.370	1.000	0.002	?	0.370	1.000	0.002	?
2021	0.065	-	0.066	✓	0.069	-	0.072	✓	0.069	-	0.120	?
2022	0.048	-	0.368	✓	0.109	1.000	0.004	?	0.120	1.000	0.008	✓
Year	K = 30											
	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.060	-	0.478	✓	0.154	1.000	0.002	?	0.154	1.000	0.002	?
2015	0.127	1.000	0.006	?	0.127	1.000	0.008	?	0.127	1.000	0.016	?
2016	0.104	1.000	0.008	✓	0.104	1.000	0.022	?	0.161	1.000	0.002	?
2017	0.093	1.000	0.030	?	0.231	1.000	0.002	?	0.231	1.000	0.002	?
2018	0.046	-	0.816	✓	0.076	-	0.358	✓	0.107	-	0.096	✓
2019	0.046	-	0.860	✓	0.076	-	0.322	✓	0.090	-	0.244	✓
2020	0.224	-	0.004	?	0.224	1.000	0.006	?	0.224	0.100	0.006	?
2021	0.069	-	0.092	?	0.080	-	0.100	✓	0.112	1.000	0.010	?
2022	0.112	1.000	0.014	✓	0.112	1.000	0.044	?	0.112	-	0.066	✓

Table A.8: Governance Mean-Variance Efficiency Test Results: Gungor and Luger (2016) Test
This table presents results for three benchmark stock sets (K = 10, 20, 30) and test stocks (N = 10, 20, 30) from 2014 to 2022. Key metrics include the F_{max} statistic, BMC p-value, LMC p-value, and final decision on mean-variance efficiency. With α = 5%, the conservative MC p-value is reported if $\tilde{P}_M^C(F_{max}(Y)) \leq 5\%$, the liberal MC p-value if $\tilde{P}_M^L(F_{max}(Y)) > 5\%$, and both if inconclusive. The symbol "-" is used whenever the p-values are not reported. Decisions are denoted by (✓) for "Do not reject," (✗) for "Reject," and (**?**) for "Inconclusive," based on p-values.

	K = 10											
Year	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.021	-	0.840	✓	0.024	-	0.918	✓	0.034	-	0.828	✓
2015	0.025	-	0.668	✓	0.030	-	0.830	✓	0.031	-	0.894	✓
2016	0.033	-	0.254	✓	0.032	-	0.518	✓	0.032	-	0.894	✓
2017	0.033	-	0.290	✓	0.033	-	0.488	✓	0.038	-	0.486	**?**
2018	0.022	-	0.290	✓	0.037	-	0.348	✓	0.54	-	0.372	✓
2019	0.011	-	0.992	✓	0.012	-	1.000	✓	0.019	-	1.00	✓
2020	0.007	-	0.996	✓	0.016	-	0.998	✓	0.022	-	0.986	✓
2021	0.013	-	0.938	✓	0.017	-	0.908	✓	0.019	-	0.982	✓
2022	0.015	-	0.898	✓	0.028	-	0.524	✓	0.030	-	0.636	✓
	K = 20											
Year	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.036	-	0.584	✓	0.057	-	0.396	✓	0.057	-	0.484	✓
2015	0.0240	-	0.732	✓	0.040	-	0.944	✓	0.040	-	0.976	✓
2016	0.017	-	1.000	✓	0.019	-	1.000	✓	0.042	-	0.890	✓
2017	0.018	-	0.978	✓	0.022	-	1.000	✓	0.033	-	0.990	✓
2018	0.044	-	0.316	✓	0.044	-	0.586	✓	0.044	-	0.818	✓
2019	0.017	-	0.992	✓	0.042	-	0.730	✓	0.042	-	0.876	✓
2020	0.014	-	0.998	✓	0.022	-	1.000	✓	0.023	-	1.000	✓
2021	0.043	-	0.098	✓	0.043	-	0.278	✓	0.043	-	0.464	✓
2022	0.017	-	0.910	✓	0.017	-	1.000	✓	0.023	-	1.000	✓
	K = 30											
Year	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.059	-	0.862	✓	0.059	-	0.946	✓	0.059	-	0.958	✓
2015	0.114	-	0.730	✓	0.138	-	0.668	✓	0.138	-	0.708	✓
2016	0.043	-	0.996	✓	0.057	-	0.994	✓	0.057	-	0.996	✓
2017	0.058	-	0.572	✓	0.061	-	0.780	✓	0.068	-	0.732	✓
2018	0.035	-	0.992	✓	0.051	-	0.990	✓	0.055	-	0.998	✓
2019	0.050	-	0.976	✓	0.053	-	0.994	✓	0.060	-	0.996	✓
2020	0.026	-	1.000	✓	0.076	-	0.408	✓	0.076	-	0.514	✓
2021	0.015	-	0.998	✓	0.037	-	0.862	✓	0.042	-	0.922	✓
2022	0.023	-	0.998	✓	0.036	-	0.980	✓	0.036	-	0.998	✓

Table A.9: Governance Mean-Variance Spanning Test Results: Gungor and Luger (2016) Test
This table presents results for three benchmark stock sets (K = 10, 20, 30) and test stocks (N = 10, 20, 30) from 2014 to 2022. Key metrics include the F_{max} statistic, BMC p-value, LMC p-value, and final decision on mean-variance efficiency. With α = 5%, the conservative MC p-value is reported if $\tilde{P}_M^C(F_{max}(Y)) \leq 5\%$, the liberal MC p-value if $\tilde{P}_M^L(F_{max}(Y)) > 5\%$, and both if inconclusive. The symbol "-" is used whenever the p-values are not reported. Decisions are denoted by (✓) for "Do not reject," (✗) for "Reject," and (**?**) for "Inconclusive," based on p-values.

	K = 10											
Year	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.074	1.000	0.038	**?**	0.074	-	0.080	✓	0.114	1.000	0.008	**?**
2015	0.131	1.000	0.002	**?**	0.142	1.000	0.002	**?**	0.233	1.000	0.002	**?**
2016	0.110	1.000	0.002	**?**	0.117	1.000	0.002	**?**	0.117	1.000	0.002	**?**
2017	0.095	1.000	0.002	**?**	0.097	1.000	0.004	**?**	0.097	1.000	0.012	**?**
2018	0.083	1.000	0.083	**?**	0.101	1.000	0.008	**?**	0.018	1.000	0.018	**?**
2019	0.075	1.000	0.075	**?**	0.075	1.000	0.050	**?**	0.075	-	0.074	**?**
2020	0.092	1.000	0.034	**?**	0.132	1.000	0.014	**?**	0.192	0.916	0.006	**?**
2021	0.085	1.000	0.002	**?**	0.085	-	0.004	**?**	0.106	1.000	0.002	**?**
2022	0.084	1.000	0.010	**?**	0.101	-	0.002	**?**	0.229	0.046	-	✗
	K = 20											
Year	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.041	-	0.934	✓	0.085	-	0.258	✓	0.085	-	0.340	✓
2015	0.076	0.252	0.026	**?**	0.188	0.002	0.048	**?**	0.188	1.000	0.002	**?**
2016	0.121	1.000	0.121	**?**	0.121	1.000	0.018	**?**	0.121	1.000	0.002	**?**
2017	0.039	-	0.844	✓	0.067	-	0.506	✓	0.067	-	0.716	✓
2018	0.094	1.000	0.050	**?**	0.094	-	0.104	✓	0.201	1.000	0.002	**?**
2019	0.055	-	0.055	✓	0.133	1.000	0.002	**?**	0.198	1.000	0.004	**?**
2020	0.097	0.038	0.004	**?**	0.285	1.000	0.002	**?**	0.285	1.000	0.002	**?**
2021	0.074	1.000	0.012	**?**	0.074	-	0.052	✓	0.083	-	0.060	✓
2022	0.038	-	0.636	✓	0.145	1.000	0.002	**?**	0.145	1.000	0.002	**?**
	K = 30											
Year	N = 10				N = 20				N = 30			
	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision	F_{max}	BMC	LMC	Decision
2014	0.111	-	0.338	✓	0.111	-	0.562	✓	0.111	-	0.622	✓
2015	0.211	-	0.382	✓	0.211	-	0.554	✓	0.211	-	0.590	✓
2016	0.109	-	0.920	✓	0.109	-	0.890	✓	0.109	-	0.920	✓
2017	0.091	-	0.414	✓	0.107	-	0.414	✓	0.127	-	0.222	✓
2018	0.101	-	0.442	✓	0.142	-	0.180	✓	0.210	1.000	0.01	**?**
2019	0.189	1.000	0.020	**?**	0.189	1.000	0.040	**?**	0.189	-	0.068	✓
2020	0.143	1.000	0.024	**?**	0.368	1.000	0.002	**?**	0.368	0.106	0.002	**?**
2021	0.092	1.000	0.036	**?**	0.099	-	0.066	✓	0.942	0.218	0.002	**?**
2022	0.138	-	0.008	✓	0.138	1.000	0.014	**?**	0.138	1.000	0.018	**?**