Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

The M6 forecasting competition: Bridging the gap between forecasting and investment decisions

Spyros Makridakis [a], Evangelos Spiliotis [b], Ross Hollyman [c,d], Fotios Petropoulos [c,a,*], Norman Swanson [e], Anil Gaba [f]

[a] *Makridakis Open Forecasting Center, Institute For the Future, University of Nicosia, Cyprus*
[b] *Forecasting and Strategy Unit, School of Electrical and Computer Engineering, National Technical University of Athens, Greece*
[c] *School of Management, University of Bath, UK*
[d] *Business School, University of Exeter, Exeter, UK*
[e] *Department of Economics, Rutgers University, NJ, USA*
[f] *INSEAD, Singapore*

ARTICLE INFO

Keywords:
Forecasting competitions
M competitions
Forecast accuracy
Investment decisions
Assets

ABSTRACT

The M6 forecasting competition, the sixth in the Makridakis competition sequence, focused on financial forecasting. A key objective of the M6 competition was to contribute to the debate surrounding the Efficient Market Hypothesis by examining how and why market participants make investment decisions. To address these objectives, the M6 competition investigated forecasting accuracy and investment performance in a universe of 100 publicly traded assets. The competition employed live evaluation on real data across multiple periods, a cross-sectional setting where participants predicted asset performance relative to that of other assets, and a direct evaluation of the utility of forecasts. In this way, we were able to measure the benefits of accurate forecasting and assess the importance of forecasting when making investment decisions. Our findings highlight the challenges that participants faced when attempting to accurately forecast the relative performance of assets, the great difficulty associated with trying to consistently outperform the market, the limited connection between submitted forecasts and investment decisions, the value added by information exchange and the "wisdom of crowds", and the value of utilizing risk models when attempting to connect prediction and investing decisions.

© 2024 The Author(s). Published by Elsevier B.V. on behalf of International Institute of Forecasters. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Investing involves allocating money or resources with the expectation of future profit. People invest for various reasons—saving for retirement, building wealth, funding education, or achieving various financial goals—with different types of investments offering varying degrees of expected return and risk. The stock market is the asset class of choice for most investors and a focal point of interest for academics and practitioners. The M6 forecasting competition requested participants to invest in 50 US stocks and 50 international exchange-traded funds (ETFs) with the winners sharing $300,000 in prizes based on their forecasting and investment performance.

The objective of the M6 competition was to contribute to the debate around the Efficient Market Hypothesis (EMH), advanced by Fama (1969) and popularized a few years later by Malkiel (1973). The EMH simply states that asset prices fully reflect all available information,

[*] Corresponding author at: School of Management, University of Bath, UK.
E-mail address: f.petropoulos@bath.ac.uk (F. Petropoulos).

https://doi.org/10.1016/j.ijforecast.2024.11.002
0169-2070/© 2024 The Author(s). Published by Elsevier B.V. on behalf of International Institute of Forecasters. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Please cite this article as: S. Makridakis, E. Spiliotis, R. Hollyman et al., The M6 forecasting competition: Bridging the gap between forecasting and investment decisions. International Journal of Forecasting (2024), https://doi.org/10.1016/j.ijforecast.2024.11.002.

making it impossible to consistently outperform the market through stock picking or market timing. Empirical evidence on the validity of the EMH, comparing the performance of all types of active funds versus corresponding passively managed benchmarks, is provided yearly by Armour (2023). The data suggest that the success of active funds diminishes considerably in comparison to corresponding market averages as the time horizon of evaluation increases.

Yet there are a few investors and firms that seem to consistently outperform the market benchmarks, generating positive alphas, by stock picking or by exploiting a number of market inefficiencies (Pedersen, 2015). These investors include the legendary Warren Buffett (who according to the 2022 letter to Berkshire shareholders (Buffett, 2023) has generated compounded annual gains of about 10% higher than the corresponding returns on the S&P 500 (see Fig. 1, top panel)), Peter Lynch, George Soros, Carl Icahn, and the firms Bridgewater Associates and Renaissance Technologies, among others. Are these investors defying the EMH? And if they are, for how long do they manage to do so and how is it done, beyond pure luck? These are interesting questions. In the case of Berkshire, higher performance can be attributed to the application of Benjamin Graham's value investing (Graham, 1949) by Buffett, who was his student at Columbia University and applies a modified version of Graham's value investing to stock selection. According to the 2022 letter to Berkshire shareholders, the firm produced a phenomenal overall compounded gain between 1964 and 2022 of 3,787,464% versus 24,708% for the S&P 500. Is Buffett a great stock picker, beating the EMH over such a long period of 58 years?

The performance of Berkshire versus that of S&P 500 is shown in the top panel of Fig. 1. Evidently, different results arise if the period is separated into two sub-periods, one between 1965 and 2002 (Fig. 1, middle panel) and the other between 2003 and 2022 (Fig. 1, bottom panel). In the middle panel of Fig. 1, Berkshire's gains are considerably higher than the S&P 500 ones, reaching a high of close to 20% in 1989 and settling to 15.64% higher at the end of 2002. Thus, Berkshire outperformed the S&P 500 quite effectively in the 1965 to 2002 period, when comparing compounded annual gains.

The bottom panel of Fig. 1 covers the 20-year period between 2003 and 2022 and displays a completely different picture from that of the middle panel. In this panel, we observe no significant differences in the annual compounded gains between Berkshire and the S&P 500. There may be many reasons for this fundamental change in performance, with a major one being increased competition, as Buffett's success and investment principles became widely known and more investors and funds began imitating his approaches. Other reasons may center around the significant increase in the size of his investment portfolio, and changes in the economic environment, including low interest rates, quantitative easing, and technological advancements in stock selection methodologies used by competing investors. Additionally, Buffett's conservative investment philosophy, focusing on long-term value investing, may not have kept pace with the more aggressive, high-growth investment strategies being followed by others since the beginning of the 21st century.

Fig. 1 provides three different ways of presenting the same data, raising questions about the fairest way of doing so. It is clear, however, that whilst Buffett consistently created significant positive alpha (α) for 38 years (Fig. 1, middle panel) that trend did not persist post-2002. The top panel of Fig. 1 shows consistent compounded gains of Berkshire over S&P 500 for the entire 1965–2022 period being greatly influenced by its considerable pre-2003 gains. Put differently, the bottom panel of Fig. 1, exhibits no difference in compounded gains between Berkshire and the S&P 500 when starting comparisons after 2002, unless pre-2003 performance is added to the mix. Needless to say, and as illustrated in this figure, there is no guarantee that past out-performance extends into the future. This does not necessarily mean that Buffett has stopped being a great investor. Perhaps markets have simply become more efficient, as information is now disseminated faster because it is made instantaneously available to all players in the markets. What will happen in the future, given that Buffett's stock selection has shifted to a current concentration of 78% of the entire portfolio in just five stocks, with Apple being 50% of the total (Best, 2023)? Perhaps this approach will result in Berkshire's returns and positive alpha increasing. Still, what will happen if Apple, for example, suffers from reduced future revenue and stock price growth? Is there too much reliance on accurately forecasting future relative returns and not enough reliance on diversifying risk? For an excellent and comprehensive analysis of Buffett's alpha, the interested reader is referred to Frazzini et al. (2018).

One of the primary objectives of the M6 competition was to investigate the value of accurate forecasting for investment decision making. In our data, we found no clear connection between the two. What was even more interesting was the fact that on average, the top-performing teams in the forecasting task constructed relatively inefficient portfolios, while the top-performing teams in the investment task submitted less accurate forecasts. At the same time, top teams still performed relatively well, with 23% beating the forecasting benchmark and achieving an R-squared of 0.099 between forecasting and the investment decision ranks. Based in part on these findings, we believe that even small improvements in forecasting accuracy can provide significant improvements in investment decisions, and we plan to explore this hypothesis and a number of related hypotheses in the remainder of this paper. We begin in Section 2 by discussing the design and execution of the M6 competition. We then provide an overview of the team submissions comprising our experimental data in Section 3. Ten different hypotheses are then postulated and tested in Section 4. In Section 5, we continue our analysis by discussing our winning submissions and the methods utilized by the winning teams. In Section 6, we present an investment risk model that allows us to compare the investment performance of M6 teams with a standardized benchmark. Finally, our key findings and insights are collected in Section 7, and our concluding remarks and directions for future research are gathered in Section 8. An appendix contains technical details of the investment risk model discussed in Section 6.

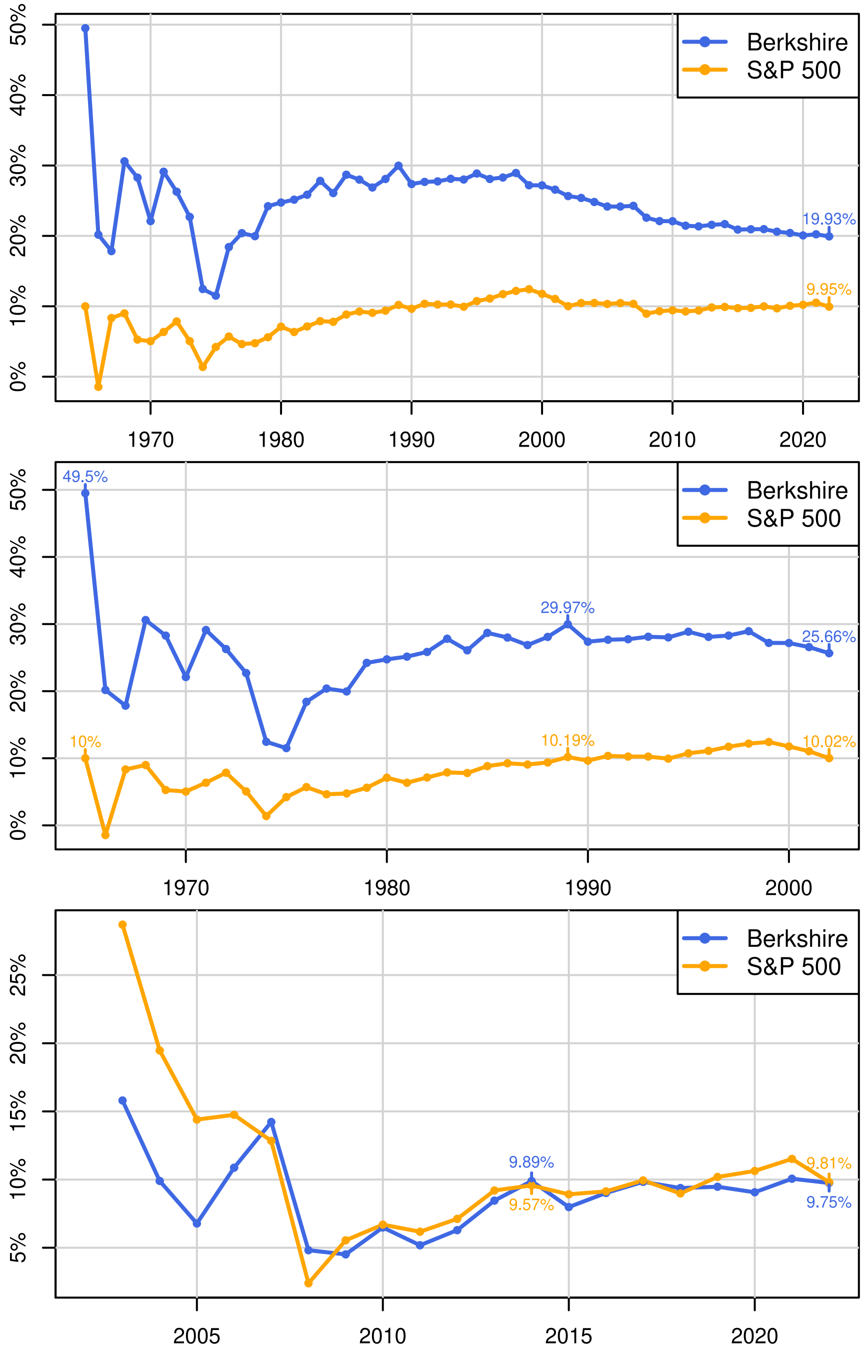


Fig. 1. Performance of Berkshire versus that of S&P 500. Top panel: 1965–2022. Middle panel: 1965–2002. Bottom panel: 2003–2022.

2. M6 competition design and execution

The M6 forecasting and investment competition was made up of participating teams[1] who were asked to submit forecasts of the relative performance of a number of tradable assets, as well as the associated investment positions for these assets. Their objective was to achieve the best performance in two challenges, forecasting and investment, and to achieve the best overall performance after combining both challenges. In this section, we discuss the design of the M6 competition in detail, with a focus on the assets modeled by competition participants, the process of the competition (i.e. submission points and submission requirements), and the measures used to evaluate performance.

[1] Each team could consist of up to five participants: a team leader who was responsible for making the submissions, and four members. None of the team members were allowed to be part of another team.

2.1. Design innovations

The M6 forecasting competition offered three major design innovations compared to the previous M competitions (Makridakis et al., 2021). These design innovations are as follows:

Live evaluation on real data and over multiple periods

Four of the previous five M forecasting competitions (M1 (Makridakis et al., 1982), M3 (Makridakis & Hibon, 2000), M4 (Makridakis et al., 2020b), and M5 (Makridakis et al., 2022b)) focused on measuring the performance of forecasts submitted on a single evaluation window; a technique also known as fixed-origin evaluation, Tashman (see also 2000). The organizers of these competitions split the available data for each time series into in-sample and out-of-sample periods. The in-sample data were disclosed to the participants, without any other information that would reasonably allow them to recover the nature or identity of the target time series and, thus, the values of the out-of-sample observations. To avoid efforts by participants to identify the target time series by value-matching, the organizers of the M4 competition (a competition that made use of predominantly publicly available time series) scaled the data using a random multiplier. The participants were asked to submit forecasts for the out-of-sample period, and their forecasts were then compared to the actual out-of-sample data. This design is relatively straightforward and allows the organizers to readily analyze forecast performance.

The only exception to the above approach was that taken in the M2 competition (Makridakis et al., 1993), where a three-phase approach was adopted. The organizers provided participants with a first batch of data, which the participants used to produce their first set of forecasts. The organizers then sent participants an updated version of the data, and participants submitted updated forecasts. Finally, the organizers offered not only an updated version of the data but also an analysis of the participants' forecasts, and a final set of forecasts were produced by the participants. This iterative process allowed participants to improve their forecasts via the provision of feedback on their previous forecasts. Data in this competition were proprietary and were masked via the use of multipliers.

In contrast, the M6 competition was built on real, live data through a series of 12 rolling submission points that covered almost an entire calendar year. More importantly, the primary data used in the M6 competition were fully identified, open, and publicly available. In fact, the M6 forecasting competition was the first competition where data were not provided directly to participants. Instead, a set of 100 publicly tradable assets (50 stocks and 50 ETFs) were specified, and their unique identifiers were shared with the participants. The organizing team did offer an interface to collect data associated with the adjusted closing values of the selected assets, but participants had the option to use their own data sources (either from freely available sources such as Yahoo Finance, Google Finance, etc., or from subscription-based services) to collect historical prices of the assets, and to decide on the amount and frequency of the historical data that they wished to utilize in their analyses. As the names and identifiers of the assets were fully transparent, participants also had the opportunity to collect additional supporting data (again from freely available or subscription-based sources) for use in their models. Such supporting data could include contextual information from news and (social) media, fundamentals, accounting data, or economic data, for example.

For each of the 12 submission points, participating teams were asked to submit their forecasts (and investment positions) after incorporating new information that had accumulated up until the submission point. The competition featured a live evaluation system that was updated every day, and the participants' performance during the current month and all previous months and quarters was public information. This system offered participants unique real-time feedback on their submissions, as it allowed them to monitor and compare their performance relative to that of other teams, and to make adjustments to their modeling approaches over time. Of note is that participants were only given real-time information related to other participants' scores and positions on the leaderboard. They did not have knowledge of the forecasts and investment decisions of their competitors.

Nature of the forecasting task

As mentioned above, previous M forecasting competitions were primarily time series forecasting competitions. Participants were given a set of historical values for the target series (sometimes along with other exogenous variables) and the objective was to construct point predictions of future values of the target variable. In the last two competitions, M4 and M5, one or many pairs of prediction intervals (for predefined nominal coverage rates) around these point forecasts were also submitted by participants. To address this challenge, some participants in the M4 and M5 forecasting competitions employed approaches that relied on cross-learning. While this was straightforward and, to a degree, implicitly suggested from the design of M5 and the use of hierarchical structured data, this was not the case for M4 where the observations across the series were not even aligned in terms of time. However, this did not discourage the winning submissions of the M4 competition (Montero-Manso et al., 2020; Smyl, 2020) from successfully applying global models.

The M6 forecasting competition was not a pure time series forecasting competition, given the importance of cross-sectional forecasting when making investment decisions. Participating teams were not asked to submit forecasts of the future values of each of the assets nor associated prediction intervals. Instead, we asked participants to estimate the probability that each of the assets would be ranked within the first, second, third, fourth, or fifth quintile, when comparing relative returns across all 100 assets. In other words, forecasts in the M6 competition were not values associated with a single asset but relative predictions that took into account the predicted performance of one asset relative to that of all other assets.

Evaluating forecast utility

The previous five M competitions focused on the accuracy of the point forecasts, while the last two additionally focused on the evaluation of forecast uncertainty via the evaluation of prediction intervals. Additionally, the latter competitions offered insights into the tradeoffs between forecasting performance and computational complexity. However, none of the previous forecasting competitions explicitly focused on how forecasting performance (or forecast quality) is implicitly or explicitly linked to the forecast utility (or forecast value), i.e. how better forecasts might lead to better decisions and measurable (utility) benefits based on said decisions. For example, in the context of the M5 forecasting competition and retail forecasting, forecast utility could have been measured by assessing the connection between improvements in forecast accuracy and inventory-based performance associated with holding and backlog costs.

In the M6 forecasting competition, we explicitly evaluated utility (and provided participants with an incentive to maximize a specific definition thereof). In particular, participants were asked to submit not only forecasts associated with 100 assets, but also investment positions for each of these assets. Our objective was to assess whether or not the participants utilized their forecasts to make informed investment decisions, and to assess whether such decisions translated into risk-adjusted return gains.

2.2. Data

The investment universe of the M6 competition consisted of two classes of assets:

- 50 stocks from the Standard and Poor's (S&P) 500 index, and
- 50 international exchange-traded funds (ETFs).

The 50 stocks and 50 ETFs were selected to be broadly representative of the market. Additionally, stocks were selected from various sectors (communication services, energy, financial, health care, materials, etc.). We computed (in November 2021) the following features for each of the stocks:

- Average stock price (over the last 250 days)
- Coefficient of variation for the stock price (over the last 250 days)
- Coefficient of variation for the stock price (since the beginning of 2018)
- Average daily returns (since the beginning of 2018)
- Standard deviation of daily returns (since the beginning of 2018)
- Total returns (over the last 250 days)
- Total returns (since the beginning of 2018)
- Average volume (over the last 250 days)
- Coefficient of variation for the volume (over the last 250 days)

Given these features, we performed K-means clustering for each sector to gain insight into the range of stocks picked within each sector. Then, we randomly sampled the desired number of stocks from each sector (so that the total stocks selected from each sector reflected the size of the sector), making sure that the population of each cluster was taken into consideration (so larger clusters contributed more series, but ensuring that some series were sampled from the remaining clusters). For more details about the process used for constructing the M6 universe of assets, refer to the supplementary material.

Note that the organizing team provided an easy way for the participants to access the asset price data, which we updated daily, via a customized submission website dedicated to the M6 forecasting competition (https://m6competition.com/). However, as discussed above, this was not the sole available source of historical data. Given the open and public nature of the data and the assets, participants were able to collect historical asset prices (and other data) from alternative sources/providers, and they were able to choose the data frequency and sample periods used in their analyses. Organizers also provided a forum (https://mofc.unic.ac.cy/forum/list/) where participants could discuss topics they considered important, ask questions, clarify and exchange ideas, and disseminate code and other information.[2] The forum was also used by the M6 organizers for making announcements.

2.3. Process and timeline

The M6 competition was a live forecasting competition, lasting for 12 months, starting in February 2022, and ending a year later in 2023. It consisted of a single month trial run and 12 rolling origins where participants were asked to provide their submissions and were evaluated, once actual data became available. For each submission point, participants were asked to provide their forecasts and investment decisions over the next four weeks (usually the next 20 trading days). The submission deadline for each point was 18:00 GMT the Sunday before the start of the corresponding investment period. All submissions were made through the customized submission website dedicated to the competition.

The deadline for submissions for the trial practice run, which was not taken into account when compiling team performance measures, was February 6, 2023. Four weeks after the deadline for submissions for the trial run, the first actual submission point took place, followed by another 11 submission points. The interval between two consecutive submission points was four weeks (i.e. equal to the forecasting horizon). In other words, the rolling origin evaluation process of the competition involved 12 non-overlapping four-week periods. The 12 submission points were divided into four quarters, as presented in Table 1.

2.4. Submission requirements

The teams submitted (i) forecasts and (ii) investment decisions for all 100 assets for the next four-week period, given each submission point. At each submission

[2] Exchanging information privately was not allowed, as this would be equivalent to participating with multiple teams. The organizers disqualified teams for which there was evidence of private information exchange, as well as teams whose members were on multiple teams.

Table 1
Timeline (submission points) of the M6 forecasting competition.

Quarter	Month 1	Month 2	Month 3
1	March 6, 2022	April 3, 2022	May 1, 2022
2	May 29, 2022	June 26, 2022	July 24, 2022
3	August 21, 2022	September 18, 2022	October 16, 2022
4	November 13, 2022	December 11, 2022	January 8, 2023

Table 2
Example of the submission format for the M6 forecasting competition.

ID	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Decision
ABBV	0	0.1	0.2	0.5	0.2	0
CNC	0	0	1	0	0	0
GOOG	0.1	0.1	0.1	0.1	0.6	0.5
EWG	0.5	0.4	0.05	0.05	0	0
BMY	0.2	0.2	0.2	0.2	0.2	0
OGN	0	0	0.1	0.4	0.5	0.3
DRE	0.7	0.3	0	0	0	−0.2
UNH	0	0	1	0	0	0
...	...	...	...	...	...	...

point, a participating team was asked to submit a single file consisting of 100 rows (one row for each asset) and seven values per row. Below, we summarize the directions given to participants for the seven values that were to be provided for each of the 100 assets.

- The first value of each row should indicate the asset to which the forecasts and the investment decisions of the respective row refer. The acronym of each asset served as an identifier (e.g. "GOOGL" or "ATVI").
- The second through sixth values should be values summing to unity that refer to the probabilities with which the asset will have percentage returns that are within the first, second, third, fourth, or fifth quintile, across all assets (stocks and ETFs). In other words, we asked participants to provide probabilistic forecasts for the performance (percentage returns) of each asset relative to the other assets. We automatically checked for invalid submissions (i.e. submissions where the sum of the five probabilities did not sum to one, or submissions with negative probability values) and returned such submissions to the participants.
- The seventh value for each asset should be a numerical value corresponding to the investment weight for that asset. The weights should be positive for long positions, negative for short positions, or zero for no position. For instance, if three assets were assigned weights 0.5, 0.3, and −0.2, respectively, and all other assets were assigned weights of 0, this would mean that the participant invested in only three assets with positions long, long, and short and with a budget allocation of 50%, 30%, and 20%, respectively. If the sum of the absolute weights exceeded 1 (or 100%), then the submission was automatically considered to be invalid and returned to the participants. If the sum of the absolute weights was less than 1 (less than 100%), then the remainder was assumed to be assigned to an asset with zero return and zero risk (i.e. no investment). However, if the sum of the absolute weights was below 0.25 (25%), then a warning message was given and the submission was considered to be invalid and returned to the participants. In other words, the competition required participants to make some investment decisions and to take some risk.

Table 2 presents the first eight rows of an example submission file. In this example, the participant decided to invest in three assets ("GOOG", "OGN", and "DRE") with weights of 50%, 30%, and 20% (or 0.5, 0.3, and 0.2) and positions long, long, and short, respectively. Additionally, the participant predicted that there was a probability of 0.1, 0.2, 0.5, and 0.2 that the first asset ("ABBV") would have a percentage return within the 2nd, 3rd, 4th, and 5th quintiles, respectively. Also, the participant was confident that the expected percentage returns for the second asset ("CNC") would be within the 3rd quintile across all assets.

Note that on the occasions when participants decided not to submit forecasts and investment decisions at particular submission points, we assumed that their previous (latest) submission carried over. In other words, their forecasts and investment decisions did not change. In this regard, if for instance a team made a single submission at the first month of the competition, this same submission would be used to evaluate its performance across all 12 submission points of M6.

Note also that although participating teams were allowed to change their submissions for each submission point up to five times each day until the submission deadline, the last submission made was the only one considered for evaluation.

2.5. Measuring performance

Measuring the performance of the forecasts

The forecasting performance for a particular submission point was measured by the ranked probability score (RPS; Epstein, 1969; Gneiting & Raftery, 2007). The realized percentage total returns of all assets (stocks and ETFs) over the period were divided into quintiles, ranking from 1 (worst performing) to 5 (best performing). Given 100 assets, 20 of these would receive a rank of 5, 20 a rank of 4, and so forth. In cases involving a tie on the margins of the classes, the tied assets were all assigned the respective average rank. For instance, if four assets were tied in 18th place, then they all received a rank of $(5 + 5 + 5 + 4)/4 = 4.75$, with the three "5"s in this expression being the rank of the three assets in the first quintile, and the "4" being the rank of the asset in the second quintile.

The actual return ranking of each asset, i, and each period, T, is described by a vector $q_{i,T}$ of order 5.

- In the case of no ties on the margins of the classes, the elements in this vector, $q_{i,T,k}$, with $k \in 1, \ldots, 5$, are set equal to one if the asset is ranked in quintile k, and zero otherwise. For instance, if asset i is

ranked in the third quintile at period T, then $q_{i,T}$ has values 0, 0, 1, 0, and 0.
- In the case of ties on the margins of the classes, the values assigned to the elements of the vector $q_{i,T}$ were calculated such that the tied classes were assigned non-zero weights, with the respective weighted average being equal to the actual rank. For instance, following the above example of a 4.75 rank, the values of $q_{i,T}$ would be 0, 0, 0, 0.25, and 0.75, such that $0 \times 1 + 0 \times 2 + 0 \times 3 + 0.25 \times 4 + 0.75 \times 5 = 4.75$.

Similarly, we constructed a vector denoting the probabilities of each rank for a particular asset, $f_{i,T}$, as submitted by each participating team at each submission point.

The RPS for asset i in period T was then calculated as

$$RPS_{i,T} = \frac{1}{5} \sum_{j=1}^{5} \left(\sum_{k=1}^{j} q_{i,T,k} - \sum_{k=1}^{j} f_{i,T,k} \right)^2 . \tag{1}$$

The RPS is constructed by summing over j in order to measure the accuracy of forecasted quintiles in cumulative fashion. In other words, the RPS is the sum of several Brier scores evaluated over different probability thresholds (in our case, quintiles). The RPS for a given participant for period T is a simple average of the RPS values across all assets:

$$RPS_{i,T} = \frac{1}{100} \sum_{i=1}^{100} RPS_{i,T} . \tag{2}$$

The overall RPS for multiple submission points T_1 to T_2 is

$$RPS_{T_1:T_2} = \frac{1}{100(T_2 - T_1 + 1)} \sum_{T=T_1}^{T_2} \sum_{i=1}^{100} RPS_{i,T} . \tag{3}$$

The RPS is zero for a perfect score, and positive otherwise. The RPS for a naive method for which the probabilities of each quintile being realized are equal to 0.2 for all assets (i.e. each asset is equally likely to have a performance in the 1st, 2nd, 3rd, 4th, or 5th quintile) is equal to 0.16. From now on, this naive method will be referred as the forecasting benchmark.

Measuring the performance of the investment decisions

The performance of the investment decisions is measured by means of a variant of the information ratio, IR, defined as the ratio of the portfolio return, *ret*, to the standard deviation of the portfolio return, *sdp*. Namely, risk-adjusted returns are defined as

$$IR = \frac{ret}{sdp}, \tag{4}$$

where *ret* denotes continuously compounded portfolio returns, and *sdp* denotes the standard deviation of these returns measured at a daily frequency.

Note that numerators and denominators in reported IR values are not annualized. Additionally, the IR is a variant of the typical information ratio, but with the benchmark return set equal to 0; and it is also a variant of the Sharpe ratio, but with the risk-free rate set to 0. All return calculations begin with the daily portfolio holding period return, calculated as

$$RET_d = \sum_{i=1}^{N} w_i \left(\frac{S_{i,d}}{S_{i,d-1}} - 1 \right), \tag{5}$$

where N denotes the number of assets, w_i is a portfolio weight, and $S_{i,d}$ denotes the price of asset i at the end of trading day d, with $d - 1$ referring to the previous trading day. In all return calculations, prices are adjusted closing prices. Continuously compounded portfolio returns are then calculated as $ret_d = \ln(1 + RET_d)$.

In the above expressions, RET_d is measured for a single day, d, and is the percentage return (gain/loss) associated with each asset selected for investment, averaged by the corresponding investment decision weight for each asset. Returns for a holding period longer than one day are calculated as the sum of daily returns. In particular, the return for the holding period from d_1 to d_2 is calculated as:

$$ret_{d_1:d_2} = \sum_{d=d_1}^{d_2} ret_d . \tag{6}$$

The standard deviation, $sdp_{d_1:d_2}$, is calculated using the same $d_2 - d_1 + 1$ values of ret_d as those used in the calculation of $ret_{d_1:d_2}$. In particular, $varp_{d_1:d_2} = \frac{1}{T-1} \sum_{d=d_1}^{d_2} \left(ret_d - T^{-1} ret_{d_1:d_2}\right)^2$ and $sdp_{d_1:d_2} = \sqrt{varp_{d_1:d_2}}$ with $T = d_2 - d_1 + 1$.

Higher IR values suggest better investment performance generated per unit of risk taken. In order to benchmark the performance of the participating teams, an approach where equal long positions are taken for all 100 assets (investment weights of 0.1) was assumed.

Measuring the combined performance of the forecasts and the investment decisions

The combined performance is measured by means of the arithmetic mean of the ranks of the ranked probability score, RPS, and performance of the investment decision, IR, which implies that equal importance is given to the two tasks/challenges. As such, the overall rank for a submission, OR, is calculated as:

$$OR = \frac{\text{rank}(RPS) + \text{rank}(IR)}{2}, \tag{7}$$

where rank(·) denotes the rank of a participant relative to all other participants for that measure (RPS or IR). To calculate the overall forecasting rank across all 12 submission points, we take the arithmetic mean of the RPS as calculated in each month.

Note that the M6 GitHub repository (https://github.com/Mcompetitions/M6-methods) provided sample code (R and Python) for evaluating submissions in terms of the RPS and IR, with the objective of clarifying the evaluation process and facilitating replication of the competition's results. A MS excel file with similar computations was also available for teams with limited programming background.

2.6. Prizes

In the M6 forecasting competition, we offered a total of $300,000 in prizes awarded based on the participants' performance on each of the competition's challenges (forecasting and investment) and based on their overall performance.

We offered $42,000 in performance prizes for each of the four quarters of the competition. The (nine) quarterly prizes were awarded to the participants with the first-, second-, and third-:

- best performance in terms of forecasting (evaluated by the ranked probability score),
- best performance in terms of investment decisions (evaluated by the information ratio), and
- best overall performance in both challenges (i.e. winning the quarterly duathlon prize)

In addition, we offered $124,000 in global prizes where we considered the performance across all four quarters. The (15) global prizes were awarded to the participants with the first- to fifth-:

- best performance in terms of forecasting,
- best performance in terms of investment decisions, and
- best overall performance in both challenges (i.e. winning the global duathlon prize)

Finally, we offered $8000 ($2000 per quarter) to the best-performing submissions by teams with students as members.[3]

Note that if a participant did not submit forecasts and investment decisions in the first month of a particular quarter, and if there were no submissions to be carried over from the previous quarter, then they were automatically ineligible for the prize for that particular quarter. Similarly, if a participant was ineligible for a prize for a single quarter, then that participant was automatically ineligible for the global prizes (awards based on performance across all 12 submission points). Namely, in order for a participating team to be eligible for a global prize, it had to submit forecasts and investment decisions from the very first month of the competition (after the trial run).

In order to be able to offer this significant prize pool, we relied on sponsorships from multiple organizations, including Google (Platinum Sponsor), Meta (Gold Sponsor), JP Morgan (Diamond Sponsor), SAS, the International Institute of Forecasters, Kinaxis, Intech, ForecastPro, causaLens, Rutgers University, Erasmus Business School, University of Nicosia, and the Makridakis Open Forecasting Center.

3. Participating teams and overview of submissions

The M6 competition involved 318 participants on 226 teams. About 80% of the teams consisted of a single participant, 10% consisted of two members, and the remaining 10% included more than two members. Almost half of the participants originated in the United States of America, India, France, China, Turkey, Germany, and Greece, while the remainder originated from 43 other countries. Although the affiliations and backgrounds of the participants were not always clear, we concluded based on questionnaires that around 60% of the participants were independent researchers, consultants, and data scientists, 25% worked in the industry, and 15% were academics. Moreover, 13 of the participants were students.

Of the 226 teams that entered the M6, 163 teams (about 72%) joined at the beginning of the competition, and were thus eligible for the global prizes. Most of the remaining 63 teams joined the competition at the beginning of the second, third, and fourth quarters, probably in order to be eligible for the corresponding quarterly prizes. Additionally, 59 teams participated in the trial run to familiarize themselves with the submission system and the evaluation setup of the competition. Table 3 summarizes the number of active teams and new submissions made per submission period.

As explained in Section 2.4, the participating teams were not obliged to update their submissions at every submission point. Instead, a team could retain a previous submission for one or multiple evaluation rounds. Teams that opted to stick with their previous forecasts and investment decisions did not have to re-submit the same submission file. This meant, for example, that teams that did not have the time to work on their submissions for a certain submission point were not disqualified. Table 3 summarizes the proportion of previous submissions used for evaluation purposes at each round of the competition. It is evident that as the competition proceeded, fewer and fewer teams regularly updated their submissions. In the second submission point, 36.4% of the teams were evaluated based on their initial submission, while in the third submission point, 28.1% and 15.7% of the teams were evaluated based on the submissions they made at the first and second submission points, respectively. Ultimately, at the last submission point of the competition, only 27.9% of the teams were evaluated based on a completely updated submission, with the remaining 17.6%, 12.0%, and 42.5% utilizing submissions respectively created more than one, three, and six months prior. In addition, of the 163 teams included in the global leaderboard, only 26 made an original submission at all 12 submission points, 64 made more than six submissions, and 39 made a single submission at the very beginning of the competition, as displayed in Fig. 2.

In contrast to real life where professional investors are often committed to being active, the observations above highlight the difficulties present when organizing live competitions that involve multiple evaluation rounds and cover a long period of time. Despite the incentives we tried to provide through the quarterly prizes, few new participants were attracted after the launch of the competition, perhaps due to their exclusion from the global prizes. Moreover, our findings demonstrate how challenging it is in practice for a team to remain dedicated to such a demanding duathlon. Fortunately, said dedication seemed to pay off in many cases, as all five winners in

[3] In order for a team to be eligible for the student prize, it had to consist solely of students, with the exception of one participant who could serve as supervisor.

Table 3
Number of active teams and new submissions made per submission period. For each period, the proportion of previous submissions used for evaluation purposes is also reported.

Period	Active	New	Submissions used for evaluation (%)											
	Teams	Submissions	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th	12th
1st	163	163	100.0											
2nd	176	112	36.4	63.6										
3rd	185	104	28.1	15.7	56.2									
4th	197	114	20.8	10.7	10.7	57.9								
5th	197	86	20.3	10.2	8.1	17.8	43.7							
6th	200	85	20.0	10.0	7.5	14.0	6.0	42.5						
7th	208	88	19.2	9.6	7.2	12.5	3.8	5.3	42.3					
8th	208	66	18.8	9.6	7.2	12.0	3.4	3.8	13.5	31.7				
9th	214	72	18.2	9.3	7.0	11.7	2.8	3.7	9.3	4.2	33.6			
10th	223	82	17.5	9.0	6.7	7.6	2.7	3.1	8.1	3.1	5.4	36.8		
11th	226	71	17.3	8.8	6.6	7.5	2.7	3.1	7.5	3.1	4.0	8.0	31.4	
12th	226	63	17.3	8.8	6.6	7.1	2.7	3.1	6.2	2.7	3.5	6.6	7.5	27.9

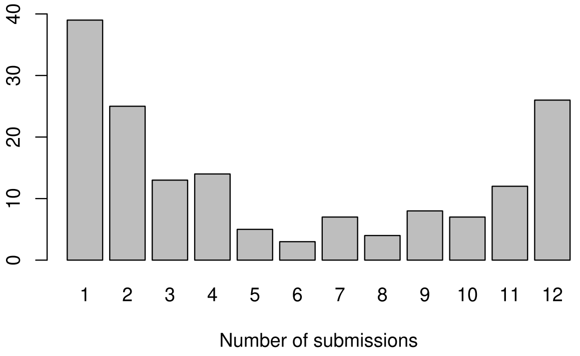


Fig. 2. Number of teams per submission count.

the forecasting track and four of the five winners in the investment track updated their submission every single round, while the same was true for three of the duathlon winners.[4] Nevertheless, the overall correlation between the number of submissions made and the performance of a team was small, suggesting that active participation is a necessary but not sufficient condition for winning a competition like the M6.

In terms of performance, of the 163 teams included in the global leaderboard, 38 (23.3%) managed to provide forecasts that were more accurate than the benchmark, 47 (28.8%) constructed better portfolios, and 11 (6.7%) achieved both a higher IR and higher RPS. It is also interesting that, as shown in Fig. 3, only three teams outperformed the benchmark's forecasts in all 12 months of the competition, and did so when comparing investment decisions. (One team did achieve a higher IR score than the benchmark in 11 different months, and three teams did so in nine different months.)

Fig. 4 visualizes the daily evolution of RPS and IR scores for the 163 teams included in the global leaderboard. When it comes to RPS, we observe that teams performed either similarly well to or significantly worse than the benchmark throughout the competition. On the contrary, there are a notable number of teams that performed either significantly better or significantly worse than the benchmark, with the majority of teams reporting lower IR scores in most of the periods. Drawing from the above discussion and by referring to Fig. 4, it is evident that beating the benchmark consistently was particularly difficult in practice, despite the simplistic forecasting and investing approaches employed by the benchmark.

4. Ten hypotheses and their evaluation

Similar to the M4 (Makridakis et al., 2020a) and M5 (Makridakis et al., 2022a) forecasting competitions, the organizers of the M6 formulated 10 hypotheses before its launch. The idea was to introduce some conjectures of both theoretical and practical interest prior to assessing any experimental evidence. In this section, we evaluate these 10 hypotheses.

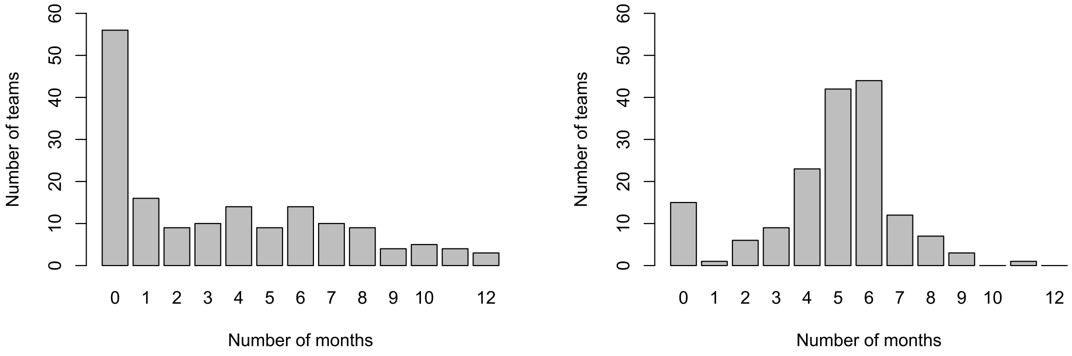
Hypothesis No. 1. *The EMH will hold for the great majority of teams, but the top-performing ones will manage to beat the market.*

Clearly we cannot (and do not) claim to accept or reject the EMH based on our data. Our teams competed for a period of only one year, limiting the amount of experimental data that we were able to collect. And the portfolios submitted were not particularly representative of those typically found in the institutional settings, in our view. In general, investment decisions would have to be tested for a much longer period of time in order to truly separate skill from luck. Nevertheless, given the practical difficulties present in running a competition for several consecutive years and maintaining a sufficiently large number of active teams, we take the M6 competition as a stylized testing ground for the EMH.

To evaluate the EMH, we focus on the 148 teams included in the global leaderboard whose investment submissions were not identical to the benchmark. We did this in order to exclude from the analysis any teams that did not submit original investment decisions.

Table 4 summarizes the performance of the benchmark and the teams in terms of returns, risk, and IR across the 12 submission points, individually and in total. Evidently, although the vast majority of the teams (75%) managed to construct less risky portfolios than the benchmark (this is not particularly challenging, as participants

[4] The other two teams made two and three submissions.



(a) RPS.

(b) IR.

Fig. 3. Number of teams included in the global leaderboard that managed to outperform the benchmark in terms of the RPS (a) or IR (b) in *N* months out of the 12 months the competition covered.

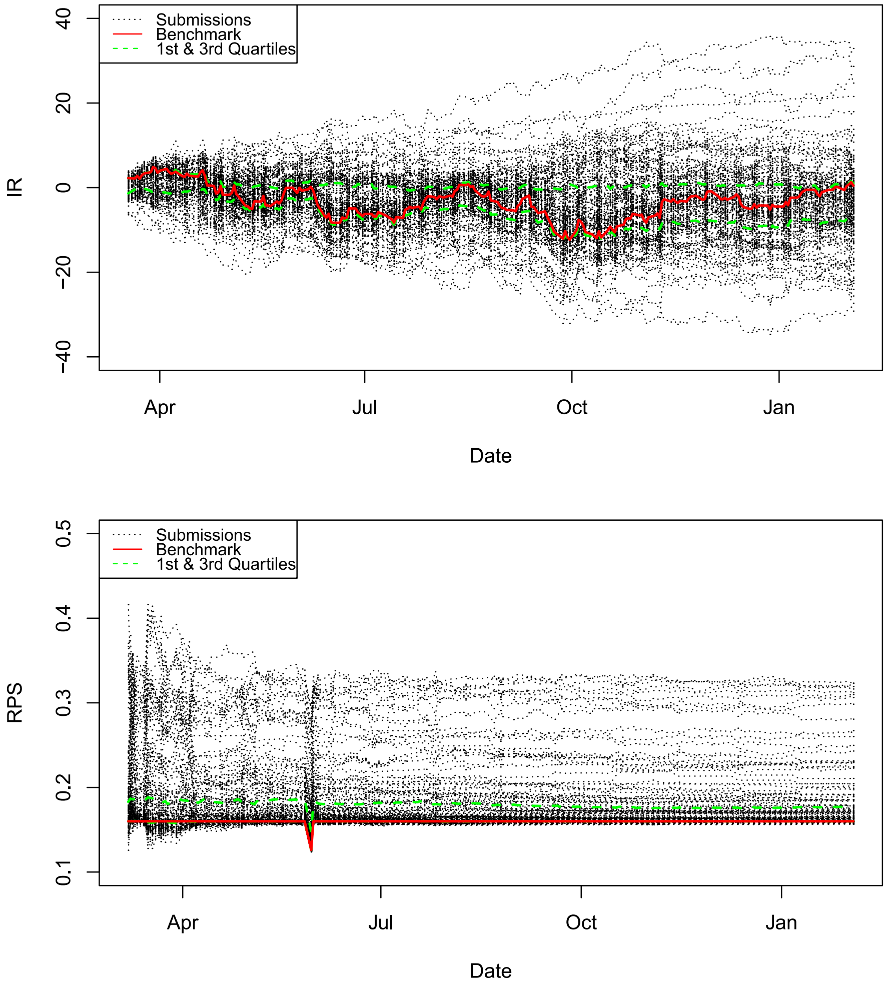


Fig. 4. Daily evolution of the RPS and IR scores of the 163 teams included in the global leaderboard. The performance of the benchmark as well as the 1st and 3rd quartiles of the submission scores are also reported. Blue vertical lines indicate the end points of the 12 evaluation rounds. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Statistics summarizing the performance of the benchmark and the teams (mean and standard deviation) in terms of returns, risk, and IR across the 12 submission points and in total. The percentage of teams that outperformed the benchmark for each measure is also reported. The results focus on the 148 teams included in the global leaderboard whose investment submissions were not identical to the benchmark.

Period	Better than the benchmark (%)			Benchmark			Teams – mean(st. deviation)		
	Returns	Risk	IR	Returns	Risk	IR	Returns	Risk	IR
1st Submission	59.46	72.97	59.46	0.044	0.011	3.990	0.015(0.032)	0.010(0.006)	1.285(3.467)
2nd Submission	18.92	77.70	24.32	−0.063	0.010	−5.972	−0.028(0.043)	0.008(0.005)	−2.957(4.433)
3rd Submission	55.41	56.76	58.11	0.018	0.015	1.215	0.006(0.036)	0.011(0.007)	0.649(3.319)
4th Submission	20.95	56.08	22.30	−0.063	0.015	−4.139	−0.029(0.049)	0.011(0.007)	−2.186(3.609)
5th Submission	27.03	89.86	35.14	0.005	0.009	0.577	−0.003(0.016)	0.007(0.005)	−0.361(2.342)
6th Submission	64.19	91.22	66.89	0.051	0.008	6.060	0.019(0.036)	0.007(0.005)	2.658(4.526)
7th Submission	25.00	73.65	29.73	−0.064	0.012	−5.273	−0.022(0.037)	0.008(0.005)	−1.891(4.858)
8th Submission	30.41	58.11	33.78	−0.073	0.015	−4.834	−0.019(0.048)	0.010(0.006)	−1.020(4.679)
9th Submission	63.51	61.49	66.22	0.110	0.014	7.839	0.028(0.067)	0.010(0.006)	2.223(5.968)
10th Submission	27.03	95.95	38.51	0.000	0.008	−0.017	−0.004(0.028)	0.006(0.003)	−0.529(4.007)
11th Submission	35.14	84.46	47.30	0.006	0.011	0.570	0.001(0.020)	0.008(0.005)	−0.015(2.245)
12th Submission	50.00	91.22	54.73	0.034	0.007	5.122	0.005(0.049)	0.006(0.005)	0.021(5.754)
Global	31.08	75.00	31.76	0.005	0.012	0.453	−0.031(0.087)	0.009(0.004)	−3.421(9.832)

were able to hedge away market risk via short positions), only 31% realized higher returns and IR. Moreover, we observe that the percentage of teams that outperformed the benchmark was usually higher when the benchmark return was positive, meaning that many teams adopted a directional bias. Therefore, it is unsurprising that, overall, the benchmark did better than the average team. Still, we also find that some teams managed to beat the market to a significant extent. Focusing on global scores, where the benchmark realized an IR of 0.453, the teams performed with a mean IR of −3.421 and a standard deviation of 9.832. In other words, assuming a normal distribution, about 16% of the teams (one standard deviation higher than the mean) should have managed to obtain an IR higher than 6.411,[5] which can be regarded as a notable improvement over the benchmark (i.e. the returns increasing from 0.5% to more than 5.6%).

Our finding that the EMH holds despite some teams managing to beat the benchmark is shown clearly in Fig. 5. In agreement with the evidence presented in Table 4, we observe that although the mean and median performance of the teams is worse than the benchmark, a small group of teams achieved strongly positive IR values corresponding to an impressive rate of return of about 30%. Furthermore, it is evident that the improvements in terms of IR grow exponentially as we move from worse to the top-performing teams. At the same time, the performance of the teams is rather symmetric around the mean in the sense that more than one-fourth of the teams realized losses that exceeded 7%, reaching up to 46%.

Hypothesis No. 2. *There will be a small group of participants that clearly outperform the average in terms of both forecast accuracy and portfolio returns.*

In order to assess this hypothesis, we focus on the 162 teams included in the global leaderboard whose average IR and RPS were −3.087 and 0.179, respectively. The medians of their scores were −1.473 and 0.162, respectively. From these teams, 75 (46.3%) managed to outperform the average submission, in terms of both forecast accuracy and portfolio returns, and 41 (25.3%) outperformed the median (refer also to the results depicted in Fig. 6).

By inspecting Fig. 5, we see that only a small group of participants clearly outperform the average and the median submissions. Also, 19 of the 75 teams (and five of the 41 for the median) report negative returns, while the average forecast accuracy improvement is less than 10% (and less than 4% for the median). At the same time, the maximum accuracy improvement is 12.7% for the average submission and 3.7% for the median submission. The number of out-performing teams is even smaller when the benchmark is used as a point of reference. Specifically, just 11 teams report a better IR and RPS than the benchmark, and although notable improvements can be identified in the investment challenge, forecast accuracy improvements do not surpass 2.2%. Given this evidence, we posit that the hypothesis is true.

Hypothesis No. 3. *There will be a weak link between the ability of teams to accurately forecast individual rankings of assets and risk-adjusted returns on investment. The strength of this link will increase in tandem with team rankings, on average. Additionally, team portfolios will in general be more concentrated and risky than can be theoretically justified given the accuracy of their forecasts.*

In order to evaluate this hypothesis, we first compute the correlation coefficient, r, between IR and RPS. We focus on the 138 teams included in the global leaderboard whose forecast submissions were not identical to the benchmark. No connection is identified between the two variables, since we find that $r = 0.04$, with significance of $p = 0.601$, as shown in Fig. 7. Nevertheless, since the M6 was a duathlon competition, it should be the case that at least the top-performing teams, based on their OR scores, have managed to achieve relatively high scores, in terms of both the IR and RPS. Fig. 7 confirms this hypothesis to some extent, suggesting that teams with a higher OR tend to construct more efficient portfolios and at the same time produce more accurate forecasts. However, this link is somewhat weak, being maximized at $r = 0.7$ for the top 20% of the teams, and declining to near zero when

[5] In fact, 16 teams (11%) had an IR higher than 6.411 and 18 teams (12%) had values lower than −13.253.

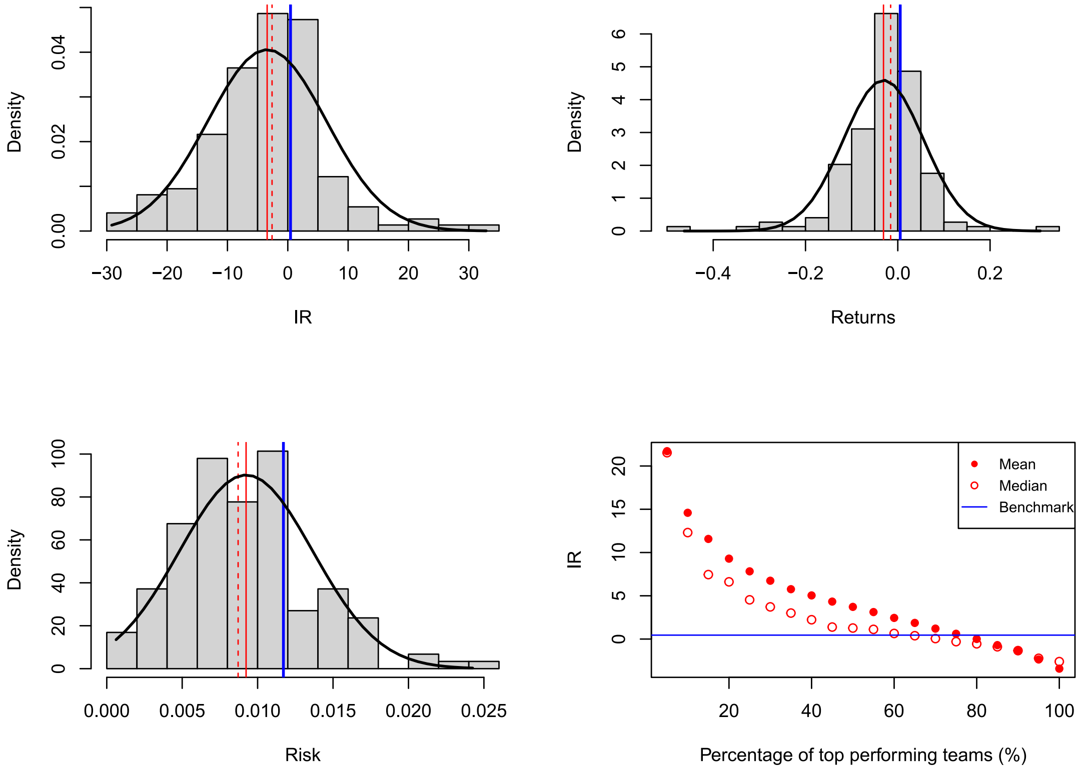


Fig. 5. Distribution of IR, returns, and risk of the 148 teams included in the global leaderboard whose investment submissions were not identical to the benchmark. A normal distribution is fitted over the histograms to facilitate comparisons. In the bottom-right plot, the mean and the median IR of the top-performing teams based on the IR is also presented, along with the benchmark, across various percentages.

more than 40% of the teams are considered. Moreover, there seems to be no association ($r = 0.12$) between the two measures for the top 5% of the teams, meaning that the teams that submitted the very best forecasting submissions did not perform similarly well in terms of investment decisions, and vice versa.

The latter finding is confirmed when the same correlation analysis is conducted, but this time the teams are ranked according to their IR and RPS instead of their OR. As shown in Fig. 7, the top-performing teams in the forecasting challenge constructed relatively inefficient portfolios on average (negative or close to zero coefficients), while the top-performing teams in the investment-decisions challenge submitted forecasts of various accuracy levels, including ones associated with zero or even negative coefficients.

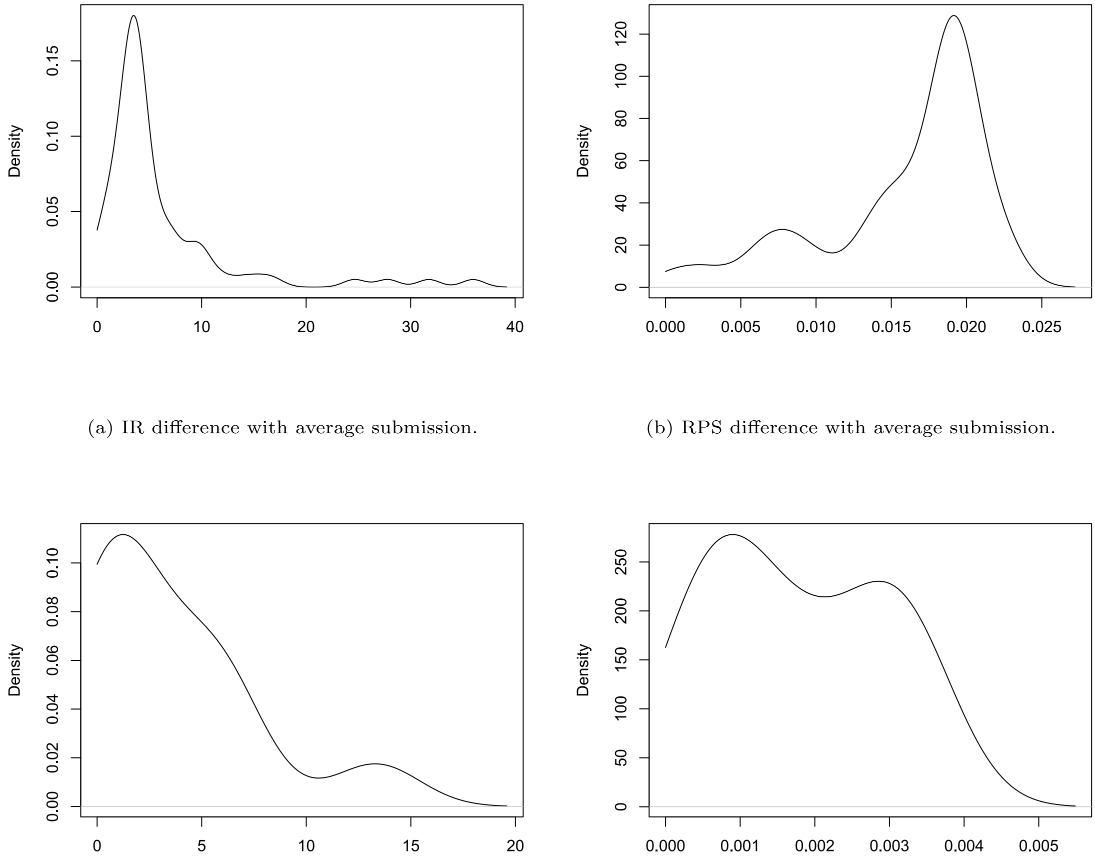
To validate the second part of the hypothesis, we introduce two concentration proxy variables, namely the average number of invested assets and the average absolute investment weight per asset. As their descriptions imply, the first variable measures concentration in terms of the number of assets involved in the constructed portfolios (more assets implies a lower concentration and less risk), while the second variable measures concentration in terms of capital invested per asset (larger investment weights imply a higher concentration and more risk). Our analysis, which involved measuring the correlation between the two concentration proxy variables and RPS, identified small negative correlations ($r = -0.05$), thus confirming that, in general, the risks taken by the teams cannot be justified by the accuracy of their forecasts.

Hypothesis No. 4. *Top-performing teams in the investment challenge will build their portfolios using assets that they can forecast more accurately.*

In order to evaluate this hypothesis, we focus on the 138 teams included in the global leaderboard whose forecast submissions were not identical to the benchmark (i.e. those teams who put some effort into the forecasting challenge of the competition). For each team and submission point, we compute the RPS of each asset separately, as well as the corresponding proportion of invested capital (investment weight). In order for the hypothesis to be true, assets that are assigned with higher investment weight by the top-performing teams in the investment challenge should also display relatively lower RPS values.

To simplify our analysis, we group the forecasts into three classes based on realized accuracy, namely high, moderate, and low accuracy. The first class comprises forecasts with RPS values lower than 0.1. The second class includes forecasts with RPS values in the range from 0.1 to 0.22. And the third class comprises forecasts with RPS values greater than 0.22.[6] Then, the average weight

[6] The thresholds were selected so that they are close to the first (0.10) and third (0.24) quantile of the RPS achieved by the teams, and so that they are symmetric around the RPS of the benchmark (0.16).



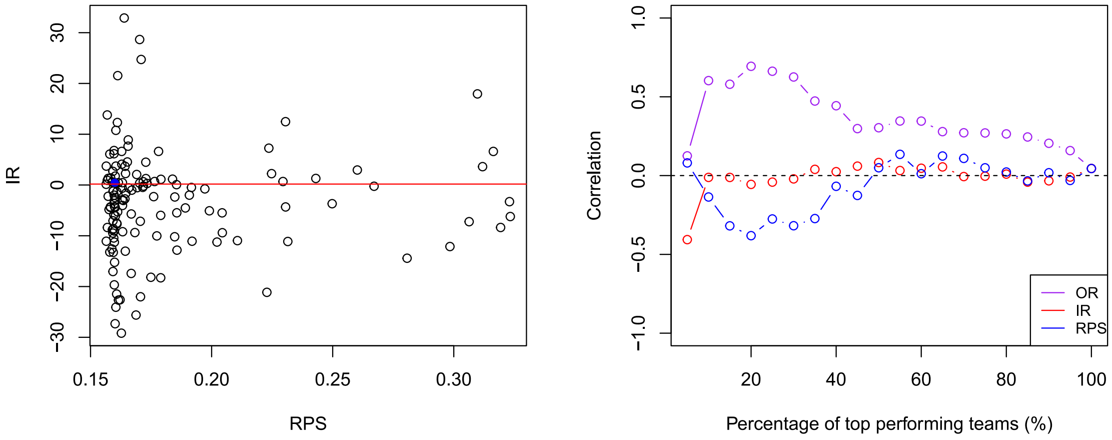
(a) IR difference with average submission.

(b) RPS difference with average submission.

(c) IR difference with benchmark.

(d) RPS difference with benchmark.

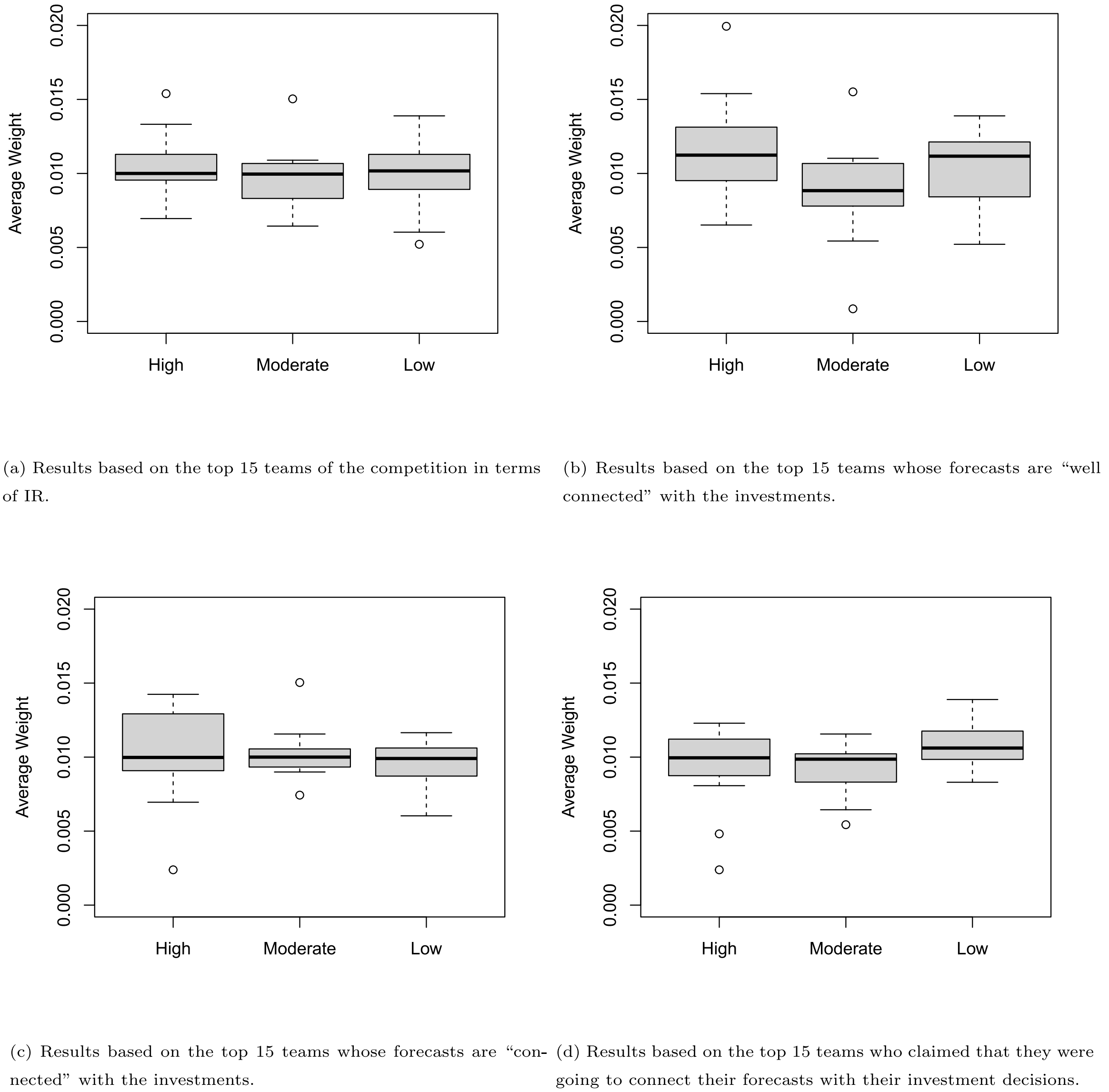
Fig. 6. Performance difference in terms of IR and RPS of the teams that outperformed the average submission and the benchmark.



(a) Results based on the 138 teams included in the global leaderboard whose forecast submissions differed from the benchmark.

(b) Results reported for various percentages of the top performing teams, ranked based on OR, IR, and RPS.

Fig. 7. Correlation between IR and RPS.



(a) Results based on the top 15 teams of the competition in terms of IR.

(b) Results based on the top 15 teams whose forecasts are "well connected" with the investments.

(c) Results based on the top 15 teams whose forecasts are "connected" with the investments.

(d) Results based on the top 15 teams who claimed that they were going to connect their forecasts with their investment decisions.

Fig. 8. Average investment weights used by the 138 teams included in the global leaderboard whose forecast submissions were not identical to the benchmark in assets that were forecast with high (RPS < 0.10), low (RPS > 0.22), or moderate ($0.10 \leq$ RPS ≤ 0.22) accuracy. The relationship between the accuracy reported per asset and the corresponding investment weight is examined for various samples of participants.

assigned to each class of assets is computed for each team across the 12 submission points, as well as the average correlation between the investment weights and RPSs.

Fig. 8 presents the distribution of the investment weights per class of assets for the top 15 teams according to IR values. As seen, there is no evidence that the teams we examined built their portfolios using assets that they could forecast more accurately, a finding that is also supported by the insignificant correlation between the investment weights and the RPSs ($r = 0.06$).

Since we showed (see Fig. 7) that the top-performing teams in the investment challenge did not perform similarly well in the forecast challenge, we might argue that the lack of correlation between the investment weights and the accuracy of the forecasts can be attributed to the general lack of connection between forecasts and the investment decisions. In order to take this possibility into account, for each team we computed the average correlation between the predicted ranks and the investment weights. For this analysis, submissions where higher amounts of capital were invested in assets of higher predicted ranks were classified as "well connected" or "connected", and the other submissions were classified as "weakly connected", "disconnected", or of "opposite connection" (for more details about this classification, refer to the supplementary material in the appendix). Fig. 8 presents the distribution of the investment weights per class of assets for the top 15 teams (according to the IR) whose forecasts were either "well connected" or "connected" with their investment decisions. Once again, the distributions of weights largely overlap across the three classes, indicating that forecast accuracy did not affect the investment decisions of the teams. Similar conclusions can be made if we focus on the top 15 teams (according to the IR) that claimed in the questionnaire that their forecasts would be linked with their investment decisions. The hypothesis is thus rejected.

Hypothesis No. 5. *Teams that employ consistent strategies throughout the competition will perform better than teams that change their strategies significantly from one submission point to another.*

In order to measure the impact of strategy consistency on the investment performance of the teams, we first define the structural elements of an investment strategy. For reasons of brevity, we focus our discussion on the following four elements.

- Exposure: We measure the total amount of capital invested at a given submission point. According to the rules of the competition, exposure can range between 0.25 and 1. Therefore, submissions were classified as lowly [0.25, 0.50), moderately [0.50, 0.80), or highly [0.80, 1.00] exposed.
- Diversification: We measure the concentration of the constructed portfolios in terms of number of invested assets. The submissions were classified as lowly [1, 10), moderately [10, 80), or highly [80, 100] diversified.
- Investment weight range: We measure the range of the investment weights within a portfolio, normalized by exposure. In practice, this measure provides information about whether the portfolio utilized similar investment weights for all assets or focused on a particular set of assets. The weight range can be either small (0, 0.1) or large [0.1, 1].
- Investment direction: The position of each asset can be either long or short, and teams can make all investments long, or all investments short, or some combination thereof. Accordingly, submissions are classified as directional or non-directional, with the latter corresponding to the case where long and short positions are both taken.

Note that the thresholds specified above are based on the distributions of the defined elements and are defined so that any class change will also signify a significant strategy change. Also note that the strategy changes are tracked over the complete duration of the competition and not on a month-by-month basis.

Having measured the exposure, diversification, investment weight range, and investment direction for all of the submissions made by each team, we computed the number of strategy changes and the corresponding investment performance (IR) for the complete duration of the competition. The correlation between these two measures is visualized in Fig. 9, both for the complete sample of teams and for the top 15 teams in the investment challenge.

Inspection of the results depicted in Fig. 9 indicates that there is a weak negative connection ($r = -0.1$) between strategy changes and investment performance. However, it is evident that teams with the same number of strategy changes can have significantly different IR scores, ranging for example from −24 to 22 when four strategy changes occurred. Moreover, the distributions of the IR values for different numbers of strategy changes largely overlap. This is confirmed if we focus on the top 15 teams. Six teams never changed their strategies, four teams changed their strategies only once, and the remaining six teams changed their strategies up to eight times, yet they all achieved comparable IR scores. Interestingly, the winning team of the investment challenge changed their strategy eight times. In addition, for the top-performing teams, the correlation of the measures that we examined becomes slightly positive ($r = 0.38$). In light of this evidence, this hypothesis is rejected.

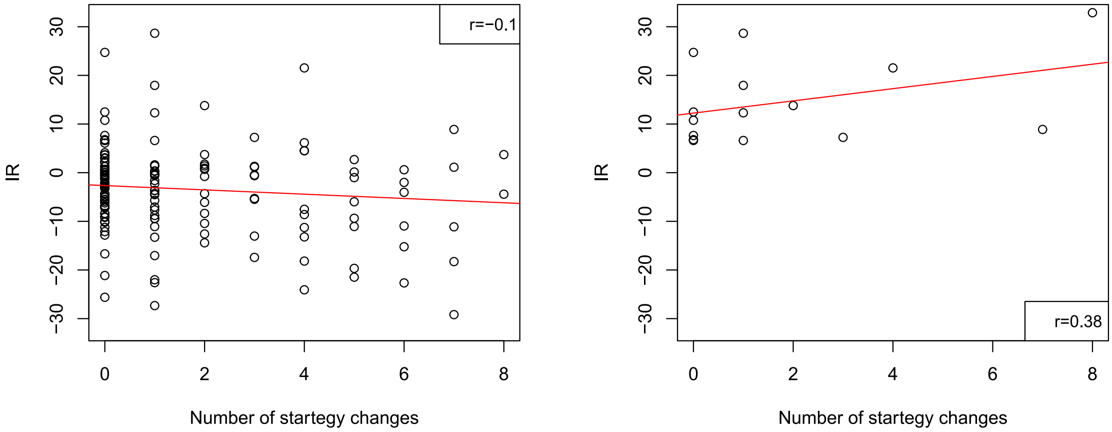
Before turning to our next hypothesis, we first provide a little more insight into the makeup of the strategic elements that defined the winners in the investment challenge. This is done by grouping the top 30 teams (according to IR) based on the exposure, diversification, investment weight range, and investment direction, on average, across the complete competition. Fig. 10 summarizes our results based on this grouping, where it is evident that lower levels of exposure were particularly beneficial for achieving higher IR scores. Moreover, more diversified portfolios of comparable investment weights typically performed better. Finally, it is evident that most of the top-performing teams invested in both short and long positions, and were thus able to effectively adjust to the directional changes of the market.

Hypothesis No. 6. *Team rankings based on information ratios will be different from rankings based on portfolio returns or rankings based on the volatility of portfolio returns.*

The IR is optimized when returns are realized with the minimum possible risk or, equivalently, when risk is realized with the maximum possible returns. In this regard, although it is generally expected that teams ranked higher in the leaderboard constructed portfolios that simultaneously reported higher returns and lower risk, this may not always be the case. For instance, two teams with similarly risky portfolios may ultimately realize significantly different returns and, therefore, IR scores. This is the basis for the present hypothesis. Namely, we investigate whether low risk or high returns contribute to superior information ratios.

Fig. 11 presents the correlation coefficients (i.e. r values), between the IR, returns, and risk in pairwise fashion, computed for various percentages of the top-performing teams (ranked according to the IR). There are several notable observations to make. First, when all teams are considered, the IR and returns are highly correlated ($r = 0.94$), in contrast to the weakly correlated IR and risk ($r = 0.09$). Second, the correlation between the IR and returns decreases significantly for the top-ranked teams, reaching a minimum of 0.55 for the top 20% of teams. Third, the IR and risk are negatively correlated for up to the top 45% of teams, while returns and risk are positively correlated for up to the top 20% of the teams.

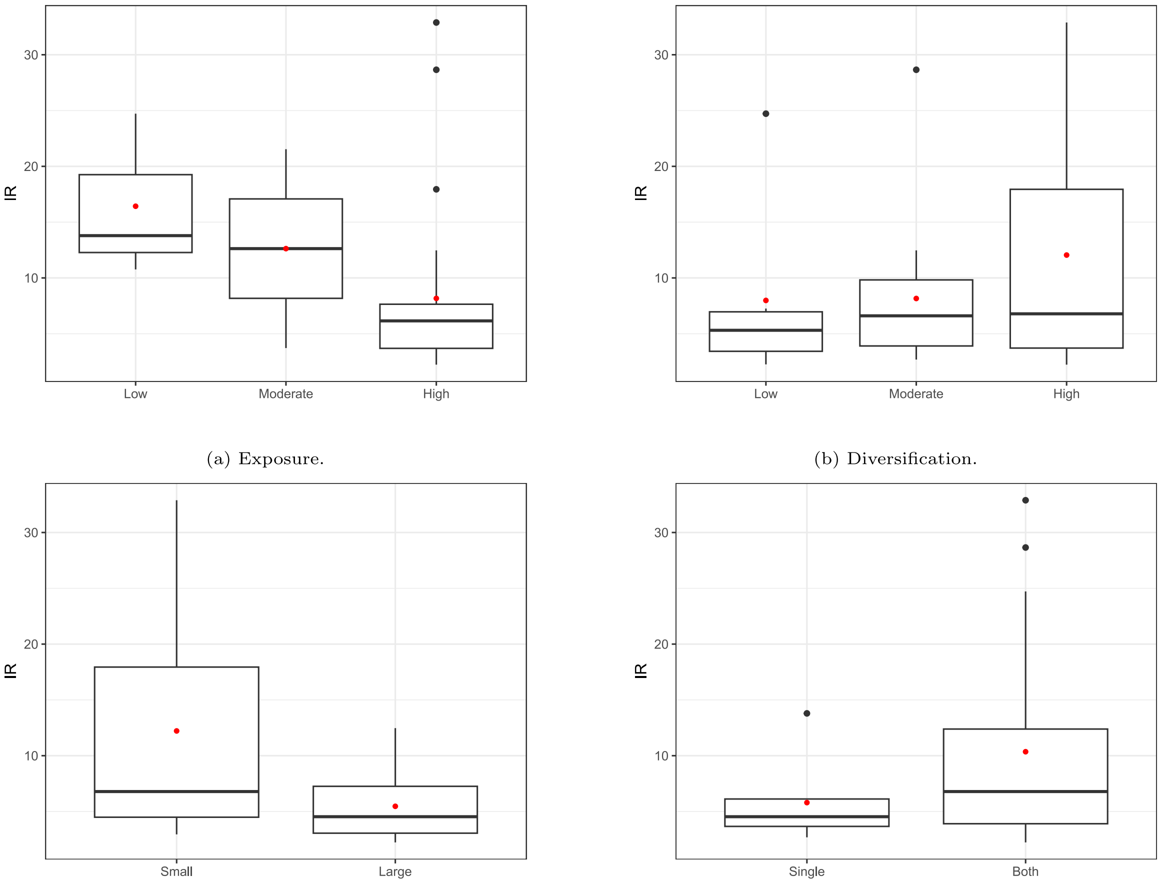
In general we find that successful teams were in a Goldilocks zone where risk was well controlled (not too hot) but sufficient (not too cold) to allow a reasonable excess return to be generated. To elaborate on this point,



(a) Results based on the 148 teams included in the global leaderboard whose investment submissions differed from the benchmark.

(b) Results based on the top 15 performing teams of the investment challenge.

Fig. 9. Correlation between the IR and the number of strategy changes.



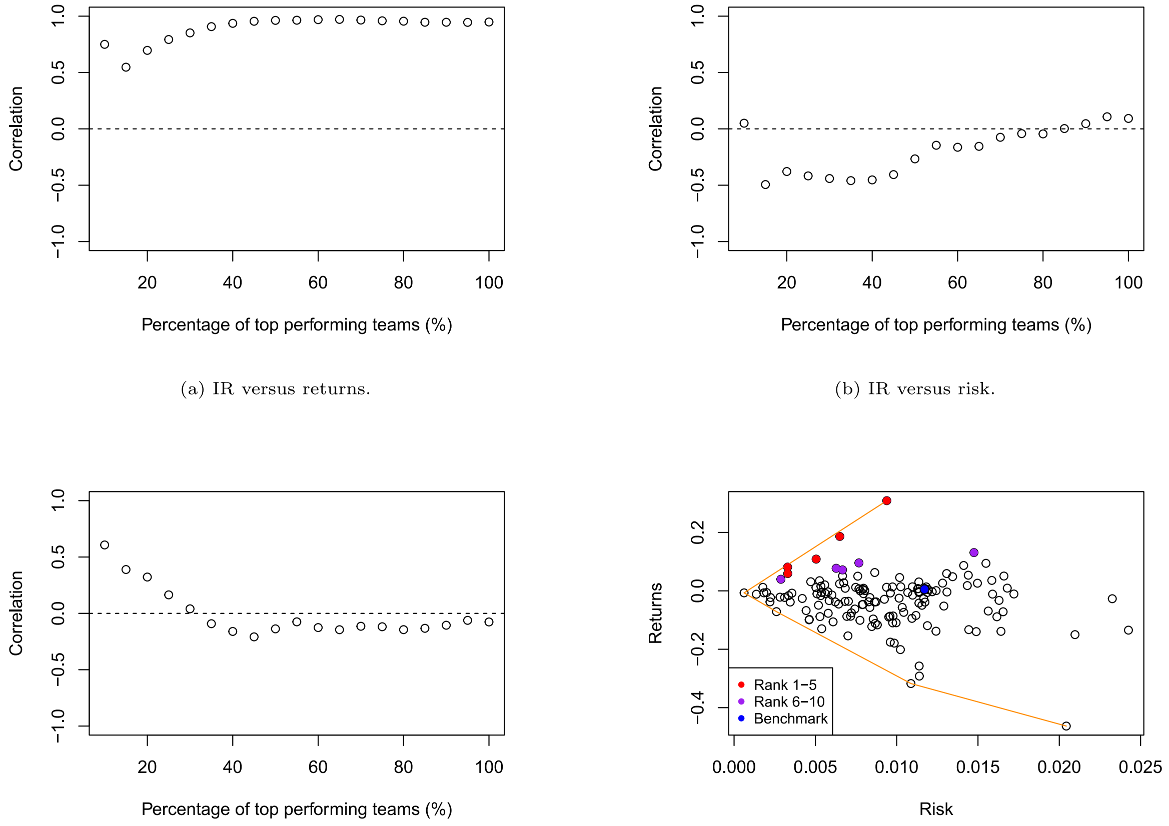
(a) Exposure.

(b) Diversification.

(c) Investment weight range.

(d) Investment direction.

Fig. 10. IR of the top 30 teams in the investment challenge, distinguished based on structural strategic elements—namely exposure, diversification, investment weight range, and investment direction. The classification of the teams is performed based on the average strategy they followed.



(a) IR versus returns.

(b) IR versus risk.

(c) Returns versus risk.

(d) Risk curve.

Fig. 11. Spearman correlation between IR and returns (a), IR and risk (b), and returns and risk (c) of the 148 teams included in the global leaderboard whose investment submissions were not identical to the benchmark. The correlations are presented for various percentages of the top-performing teams, ranked based on the IR. An estimate of the empirical risk curve is also provided (d) using convex hull, i.e. straight lines that optimally enclose every point of the dataset.

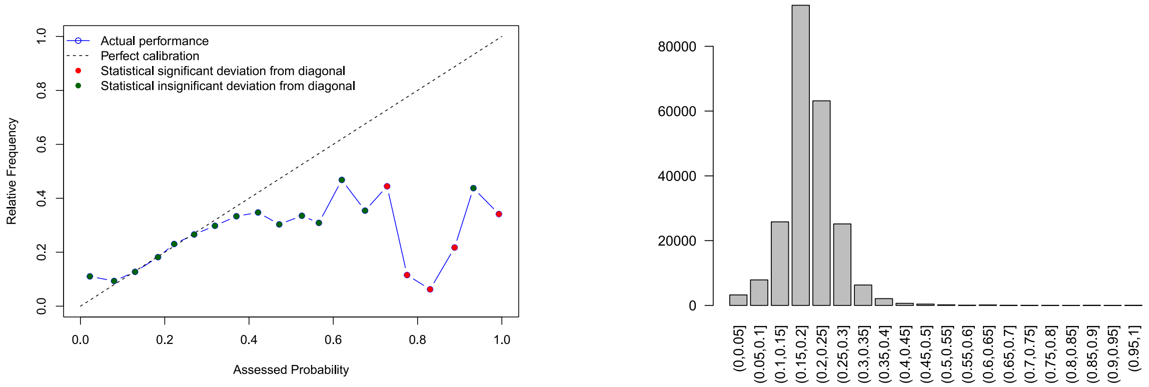
Fig. 11 presents the empirical risk curve estimated based on the portfolio returns and risk measured for each of the participating teams. As seen, with the exception of one team, all of the top 10 teams constructed portfolios that were significantly less risky than the benchmark. Moreover, the top five teams effectively managed to maximize their returns given a certain amount of risk, which was either particularly small (around 0.003) or moderate (around 0.008). Based on the above findings, we conclude that team rankings based on information ratios would be different from rankings based on portfolio returns and particularly based on rankings based on portfolio risk.

Hypothesis No. 7. *Teams will be measurably overconfident in the accuracy of their forecasts, on average. Namely, forecasts will be less dispersed and have smaller variance than observed in the data.*

In order to evaluate this hypothesis, we look at overprecision in the probability forecasts of the teams, in the sense of unwarranted certainty in their forecasts. We expect that for very low probabilities (near zero) the relative frequency of the outcomes will be higher than the probabilities, and for very high probabilities (near one) the relative frequency of the outcomes will be lower than the probability forecasts (see, for example, Lichtenstein et al., 1982).

Recall that for each asset, the teams provided probability forecasts (summing to one) that percentage returns would be within the first, second, third, fourth, or fifth quintile across all assets. These probability forecasts were evaluated using the RPS in our evaluations. We focus on the 38 teams with RPSs less than 0.16 (i.e. the teams that did better than the benchmark when constructing their probability forecasts).

The extent to which there was overconfidence in probability forecasts can be explored through a calibration curve which plots relative frequency of outcomes against forecasted probabilities (which we also call "assessed" probabilities). Fig. 12 depicts the relative frequency of outcomes corresponding to average probability forecasts within intervals of size 0.05 ranging from 0 to 1 across the 38 teams and across all quintiles, for all assets in all the submissions. The dotted diagonal line represents perfect calibration, and the solid line shows actual performance. Observe that for probability forecasts higher than 0.3 (somewhat higher than 0.2 in the benchmark), the relative frequency (based on actual outcomes) is less than the



(a) Assessed probability versus relative frequency.

(b) Number of samples per bin of assessed probability.

Fig. 12. Assessing overconfidence of forecasts.

forecasted probability, and worsens as the assessed probability increases. Similarly, for very low assessed probabilities, the relative frequency is above the 45-degree line. In other words, extreme probabilities (i.e. those that are closest to 0 and 1) show unwarranted certainty in outcomes. This evidence of forecast overconfidence leads us to accept the hypothesis.

Hypothesis No. 8. *Averaging forecast rankings (investment weights) across all teams for each asset will yield rankings (weights) that outperform those of the majority of the teams, except in cases where the very worst teams are removed from the average.*

The "wisdom of crowds" is a popular concept according to which the aggregation of information in groups typically results in better decisions than those made by any individual member of the group (Surowiecki, 2005). The benefits of combining forecasts have been confirmed in all of the previous M competitions, as well as multiple other forecasting studies (Petropoulos et al., 2022). Similar conclusions have been reached in several financial applications (Chau et al., 2020; Dai et al., 2021; Gottschlich & Hinz, 2014). In this context, this hypothesis aims to validate the value of combining, in both the forecasting and investment domains.

Our analysis focuses on the 138 teams included in the global leaderboard whose forecast submissions were not identical to the benchmark. We proceed by ranking said teams based on their OR, IR, and RPS, and averaging their submissions for different percentages of the sample, including the top 5, 10, . . . , 95, and 100 percentages. This is done while consecutively computing the corresponding IR and RPS of these averages. Note that teams were ranked on an ex post basis, i.e. according to the global scores they realized and not based on their expected performance.

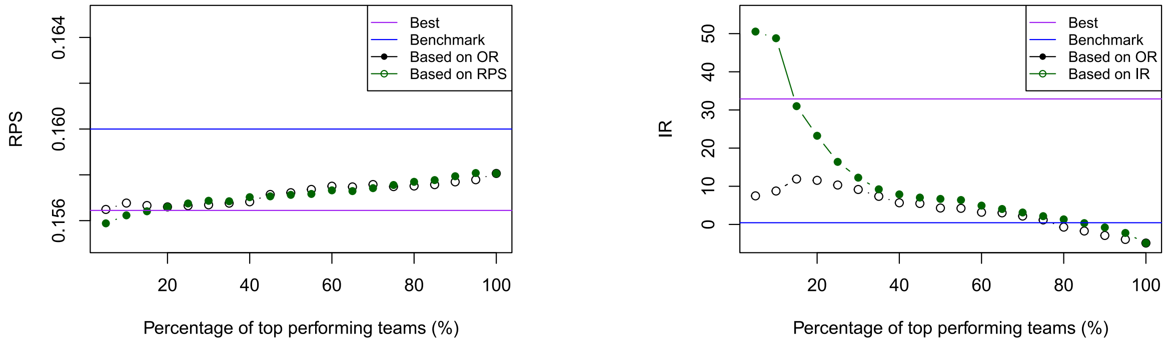
Fig. 13 summarizes the results of this analysis for the RPS and IR measures. In the first plot in this figure, the top N percentage of the teams is ranked based on the OR and RPS measures, while in the second plot, the same is done based on the OR and IR measures. This is done because averaging the submissions of the top-performing teams in the forecasting (investment decisions) challenge is expected to yield better forecasts (investment decisions) than when ranking is performed using the OR.

Focusing first on forecast accuracy, inspection of the plots in Fig. 13 suggests that averaging the forecasts of the top 5%, 10%, or 15% of the teams according to the RPS results in superior performance than the best-performing team. Moreover, when the teams are ranked based on the OR, the accuracy of the combined forecasts is similar to that of the top-performing team, even when the top 40% of the teams are included in the combination. Additionally, when the forecasts are averaged, the forecast error is always lower than that of the benchmark.

The results are similar when we consider investment decisions. Averaging the investment weights of the top 5% or 10% of the teams according to the IR results in significantly better performance than the best-performing team. In addition, the averages of the investment decisions outperform the benchmark, at least when the worst 20% of the teams is excluded from the combination. Finally, similarly to the RPS case, using the IR instead of the OR to decide which submissions should be included in the combination always yields better results. In light of these findings, the hypothesis is confirmed.

Hypothesis No. 9. *Submissions based on pure judgment or that rely heavily on judgment will perform worse than those based on data-driven methods, on average.*

In order to investigate this hypothesis, we classify participating submissions into four categories. This classification was made qualitatively based on the participants' responses to the questionnaire accompanying their submissions, and in particular to their descriptions of the forecasting methods that they implemented. Given that in most cases there were multiple submissions per team, and that between submissions, teams had the option to change their forecasting method, in this analysis we classified the descriptions by considering the most broadly defined approach taken by a team. For instance, when the forecasting approach described by a particular team at a particular submission point was simply data-driven,



(a) Performance in terms of RPS.

(b) Performance in terms of IR.

Fig. 13. Average performance of submissions when various percentages of the top-performing teams are considered either per track (ranked based on the RPS and IR, respectively) or overall (ranked based on the OR). The sample of submissions being averaged involves the 138 teams included in the global leaderboard whose forecast submissions were not identical to the benchmark.

Table 5
Participants' performance per category of forecasting method.

Category	N	%	RPS		IR	
			Mean	90th perc.	Mean	90th perc.
Data-driven	171	68.4	0.182	0.159	−3.374	6.562
Judgment-informed	8	3.2	0.181	0.158	−0.193	7.044
Pure judgment	14	5.6	0.175	0.160	−6.832	0.036
Not specified	57	22.8	0.169	0.160	−1.493	4.555

Table 6
Ranked performance of participants per category of forecasting method.

Category	N	%	RPS rank		IR rank	
			Mean	90th perc.	Mean	90th perc.
Data-driven	171	68.4	84.0	15.1	83.8	16.1
Judgment-informed	8	3.2	79.3	14.5	76.2	35.5
Pure judgment	14	5.6	89.7	49.3	101.8	66.5
Not specified	57	22.8	73.7	27.6	71.5	22.4

whereas later on it was revised also to include judgment, we categorized the overall approach as "judgment-informed". The categories that we decided upon were:

- Data-driven approaches (time series, machine learning, combinations)
- Judgment-informed (data-driven approaches informed by judgment)
- Pure judgment
- Not specified (where participants' descriptions did not allow us to categorize them into one of the above three groups)

The absolute and ranked values of the RPS and IR for each of the above categories of forecasting methods (the mean, as well as 90th percentile) are presented in Tables 5 and 6. Additionally, in the second and third columns of the tables, the counts and percentages of the submissions associated with each category are reported.

It is clear that approaches that were heavily based on pure judgment were in general inferior to those based on data-driven approaches. For example, the 90th percentile of the IR for data-driven approaches is 6.562, compared to 0.036 for pure judgmental approaches. However, it seems that there is some merit in introducing judgment to data-driven forecasting approaches. Judgment-informed forecasting approaches (albeit, there are very few of these) perform on par (if not better) compared to pure data-driven approaches. Thus, when utilized in conjunction with a data-driven approach, for example, judgment can yield good RPS and IR performance. We thus conclude that there is empirical evidence supporting the validity of the hypothesis.

Hypothesis No. 10. *The top-performing teams in the forecasting challenge will employ more sophisticated methods than the top-performing teams in the investment challenge.*

To address this hypothesis, we again focused on the qualitative responses provided by the participating teams regarding their description of their methods used in the competition. In this case, though, our classification focused on separating time series (TS)-based methods from machine learning (ML)-based methods. Also, we pooled all approaches related to the use of judgment. More specifically, we considered the following four categories:

- Judgment-based (either pure judgment or judgment-informed)

Table 7
Frequency of forecasting method categories employed by the top-performing teams in terms of the RPS.

Category	Top 5%	Top 10%	Top 15%	Top 20%
Judgment-based	1	1	2	3
TS-based	4	7	9	12
ML-based	3	7	10	14
Not specified	1	2	4	5

Table 8
Frequency of forecasting methods categories employed by the top-performing teams in terms of the IR.

Category	Top 5%	Top 10%	Top 15%	Top 20%
Judgment-based	1	2	2	3
TS-based	4	8	10	14
ML-based	3	6	9	10
Not specified	1	1	4	6

- TS-based (time series approaches, as well as their combinations)
- ML-based (machine learning approaches, including those integrated with TS approaches, as well as their combinations)
- Not specified (where participants' descriptions did not allow us to categorize them into one of the above three groups)

We present the counts of the top 5%, 10%, 15%, and 20% of the teams that described their forecasting approaches as above in Tables 7 (for RPS) and 8 (for IR). Inspection of the entries in these tables indicates that there is little evidence that teams in the forecasting challenge employed more sophisticated (i.e. ML-based) approaches than the top teams in the investment challenge. We thus have no evidence in favor of this hypothesis.

5. Winning submissions and utilized methods

For several decades, simple time series forecasting methods like exponential smoothing (Hyndman et al., 2002) and Theta (Spiliotis et al., 2020) have been outperforming relatively more sophisticated approaches, such as neural networks and other computational intelligence methods. This finding has been confirmed in numerous empirical studies, including the first three M competitions and, to a great extent, the M4. More recently, improved data availability and algorithmic advances have enabled machine learning approaches to gain gradual ground in the field and even dominate it for many forecasting applications. The results of the M5 and recent Kaggle competitions (Bojer & Meldgaard, 2021) largely confirmed the superiority of these methods and has motivated some of the recent research done to further improve their performance.

Financial forecasting involves very different challenges from those identified in the above forecasting competitions, including among other things issues related to lower signal-to-noise ratios, stochastic trends, seasonalities, and the major impacts of external factors that are either unpredictable or difficult to model. Taking these into consideration, it is not surprising that in the M6, various successful examples of both conventional (e.g. econometric time series methods) and sophisticated (e.g. machine learning methods) were identified among the top-performing submissions.

Focusing on the forecasting challenge, Daniel Weitzenfeld, who ranked 1st, predicted percentage returns by employing a Bayesian dynamic factor model with heteroskedasticity (Nardari & Scruggs, 2007) that accounts for both the expected value of an asset's return and the error variance around that value. The model was extended so that the forecasts could be adjusted when earnings of stocks were announced, also taking into account the hierarchical structure of the assets (e.g. stocks, ETFs, and fixed-income ETFs). In order to derive the required probabilities, samples from the posterior distribution were drawn, each resulting in a different set of point forecasts, which were then ranked and counted per quintile. Sebastian Radlwimmer, ranked 3rd in the forecasting challenge, relied on a similar approach to predict asset returns.[7]. He assumed that asset prices follow a geometric random walk with drift and used a stack of linear models to estimate the conditional mean, as well as Monte Carlo sampling to estimate the covariance matrix. Data augmentation (data from assets outside the M6 universe) was also used.

On the other hand, Miguel Perez Michaus, ranked 2nd in the forecasting and duathlon challenges, derived forecasts using an XGBoost model that defined the prediction problem as a multi-label classification task. The forecasting pipeline used in his method involved smoothing the close price data and extracting features that capture the volatility of the daily ranking of each asset over a number of rolling windows. Filip Staněk, ranked 1st in the duathlon challenge and 4th in the forecasting part of the competition, employed a meta-learning model (Hospedales et al., 2022) that leverages hypernetworks to make quintile predictions. The model was trained on pooled data to model general features of the dataset, but was specialized for each series using a function fitted using a latent parameter specific to each asset. This was done by connecting a neural network used for processing the data and generating the predictions with an encoder–decoder network responsible for identifying the parameters of the former network based on the asset being predicted. To enhance forecast accuracy, data augmentation and sampling were also carried out. Hongfeng Ai, ranked 4th in the duathlon challenge, also relied on neural networks, coupled with denoising autoencoders and a robust feature-selection mechanism. The networks were designed to predict the probability that an asset would fall in each quintile. Thus, he approached the challenge as a multi-class classification problem.

In the investment challenge, it seems that most of the participants relied on relatively simpler approaches and heuristics to construct their portfolios, using as a base the probabilities estimated in the forecasting challenge, but also their judgment and various insights obtained during the competition. Approaches included averaging

[7] https://github.com/sebrad/M6

the bounds of prediction intervals and ranking the assets based on their expected returns, among other things. Interestingly, and in line with the findings of Hypotheses 2 and 3, the forecasts used for supporting the top-performing teams in the investment challenge were of moderate accuracy, at best.

Colin Catlin, who won 1st prize in the investment challenge (92nd position in the forecasting challenge), predicted percentage returns by utilizing *AutoTS*,[8] an open-source Python package designed for probabilistic time series forecasting. The package includes an extensive set of forecasting models (from naive and conventional statistical methods to machine and deep learning regression approaches), supplemented by numerous ensembling, hyperparameter tuning, and data processing (e.g. transformations, detrending, and differencing) utilities that can be applied both before and after generating predictions to improve the overall performance of the forecasting process. In addition, the package offers genetic algorithms for testing combinations of said models and utilities using cross-validation, thus allowing for the identification of suitable, specialized forecasting approaches for each series separately, in an almost automated fashion. The percentage return forecasts were transformed into investment decisions by averaging the upper and lower bounds of the prediction intervals and normalizing said averages to provide an absolute sum of unity. Hanife Taylan Selamlar, who won 2nd prize in the investment challenge (110th position in the forecasting challenge), used the ATA method, which resembles models of the exponential smoothing family, to forecast (on a monthly basis) the daily frequency of each asset's percentage returns across the five quintiles. The frequency forecasts were then normalized by asset to derive probabilities for each quintile, which were in turn used to identify assets of relatively higher and lower expected returns for the specification of long and short positions, respectively. Judgment was also used to define the number of assets to be included in the portfolios.

Many teams also relied on pure optimization methods to construct their portfolios, including traditional methods like the Markowitz model, and various other global and dynamic optimization algorithms. For example, Hongfeng Ai, ranked 13th in the investment challenge, used a differential evolution algorithm (Storn & Price, 1997) to optimize asset allocation and maximize returns under risk constraints. Interestingly, some participants decided to build their investment strategies by taking advantage of the competitive nature of the M6 competition and basing their approaches on the observed behavior of other participants. Filip Staněk, 1st prize winner of the duathlon challenge and 6th best in the investment challenge, observed that assuming the minimum exposure allowed in the competition (absolute sum of weights equal to 0.25) could provide an advantage, as it naturally reduced the penalty of extreme returns to the IR measure (ek & F, 2024). Similarly, due to the high uncertainty involved, he realized that assuming equal investment weights (absolute values) could also contribute towards more stable performance. To that end, he optimized only the number of short versus long positions submitted, which was mostly done based on the relative performance of his team over the other participants.[9]

6. An investment risk model for the M6 competition

"Quantitative active management is the poor relation of modern portfolio theory. It has the power and structure of modern portfolio theory without the legitimacy". This quote is from a foundational reference book by Grinold and Kahn (1999) that was written for quantitatively orientated portfolio managers who were learning their trade in the 1990s and 2000s, a period which saw a significant increase in practitioner interest in more systematic approaches to investment decision making. The authors seem to acknowledge that while the quantitative approach is academically rigorous, the entire premise of active investment decision making is called into question by modern portfolio theory. The M6 competition was in part motivated by this debate.

In this section we take a quantitative approach to analyzing the results of the M6 competition. In particular, we describe an investment risk model that is designed to analyze the *investment* submissions made by competition participants. Our model is relatively simple in the sense that we fit and use data which were readily available to competition participants, while taking advantage of recent advances in the understanding of the structure of multivariate volatility across asset classes. Applying our model to competition submissions, one of our key findings is that most participants were measurably overconfident: they assumed much more investment risk than was justified by the accuracy of the submissions made in the *forecasting* part of the competition. In what follows, we first summarize key features of the investment part of the M6 competition. We then dive more deeply into the measurement of investment risk. Thereafter, we outline our investment risk model. Finally, we summarize a number of findings based on the implementation of the said model.

6.1. Key features of the investment part of the competition

Recall that participants in the M6 were asked to submit their entries in two parts. The first part of each submission comprised a set of *forecasts* summarizing their expected probability distributions for the returns in the universe of investment assets specified in the competition rules. The second part of each submission comprised a set of *investment decisions* which were expected to be made on the basis of these forecasts. Participants were given a clear mandate: to maximize the Sharpe ratio of the resulting portfolio.

[8] https://pypi.org/project/autots/

[9] Given that the vast majority of the participating teams considered solely long positions, submitting a portfolio with a high number of equally weighted long positions would effectively secure for a team a future position close to its previous one. In contrast, submitting a portfolio with a higher number of equally weighted short positions would effectively allow for some diversification and risk, i.e. a potentially higher/lower position on the leaderboard.

In order to explain why we use the Sharpe ratio as our metric for assessing investment performance, recall that economic theory posits that investors choose portfolios (represented by a set of weights allocating capital to positions in some subset of the universe of potential investments) by maximizing a function representing their utility. The procedure to do this was set out in the Nobel-winning work of Markowitz (1959). The idea is that investors make capital allocation decisions by forecasting returns and variance/covariance in some universe of investment assets. For each investor, the set of optimal investment weights are then chosen by maximizing a utility function, given a parameter summarizing the investor's level of risk tolerance. In technical terms, and assuming a quadratic utility function, the investor chooses a weight vector, w, that maximizes their expected utility function:

$$U_w = \boldsymbol{\alpha}\mathbf{w} - \lambda\mathbf{w}\boldsymbol{\Sigma}\mathbf{w}', \tag{8}$$

where the risk-aversion parameter, λ, is a scalar value chosen by the investor.

While this is a simple and elegant theory, there are significant hurdles to overcome in practical applications. Aside from various conceptual (and well-documented behavioral) difficulties that non-technical investors may experience when specifying a utility function, as well as difficulties choosing an investment time horizon and calibrating risk aversion, implementing a Markowitz-style approach requires sensible *estimates* of α and Σ. In general, producing accurate return (α) and covariance (Σ) forecasts can be challenging. For example, while estimating univariate variance terms (i.e. the diagonal component of Σ) is relatively easy, since volatility is somewhat persistent, deriving reliable estimates of the off-diagonal elements of this covariance matrix is somewhat more challenging, given that investors often select from a large number of potential investments (n), and given that covariance changes over time, so that the effective time window (t) for constructing reasonably accurate estimates tends to be rather small. Moreover, experience has shown that using noisy and/or badly calibrated estimates of risk and return leads to poorly structured and poorly performing investment portfolios. Optimization can easily become an estimation error rather than a utility-maximizing tool.

Notwithstanding the forecasting challenges discussed above, the structure of the M6 competition enabled participants to abstract away from the issues discussed above regarding the choice of utility function, time horizon, and the calibration of risk aversion. This was done by giving a clear investment mandate to participants; we asked them to produce portfolios designed to maximize a variant of the well-known Sharpe ratio measured over the four trading weeks commencing on the Monday following the weekend during which they submitted their portfolios. That is, participants were asked to submit a weight vector w which would maximize the Sharpe ratio:

$$SR_w = \frac{\boldsymbol{\alpha}\mathbf{w}}{\mathbf{w}\boldsymbol{\Sigma}\mathbf{w}'} \tag{9}$$

This ratio of portfolio return to portfolio risk is commonly used by practitioners and academics as a measure of historical risk-adjusted investment performance. Notice that in the above expression, the portfolio return is simply the weighted sum of the returns on the individual assets in the portfolio. In this sense, return forecasts made in the first part of the competition are directly relevant. In addition, in order to maximize the Sharpe ratio, participants needed to model or at least make assumptions regarding portfolio risk, given that Σ appears in SR_w. The advantage of posing the M6 investment objective in this way was that no risk-aversion parameter was required. The disadvantages were that submitted portfolios varied significantly in the level of risk assumed (although this was also interesting, academically) and that the optimization objective was somewhat non-standard (although doing so is well within the capabilities of many freely available software packages). For a full discussion of Sharpe ratio optimization and its relationship to the traditional Markowitz approach, the reader is referred to Lassance (2022).

6.2. Measuring and managing investment risk

In this subsection, we discuss our approach to measuring investment risk. We begin with a brief discussion of the stylized facts of stock market volatility, illustrating these with examples from the competition and elsewhere. We then briefly summarize some of the models commonly used to capture these features of the data.

Perhaps the most important stylized fact is that return variance is not constant over time. Fig. 14 plots daily returns of US equities beginning in 1926. Note the clear evidence of clusters of volatility in the 1930s and 1940s associated with the Great Depression and the Second World War. Note also the short period of extreme volatility associated with the 1987 stock market crash and more recent episodes of volatility associated with the Asian debt/Long-Term Capital Management crisis, the financial crisis of 2007–2008, and the Covid-19 pandemic. Between these periods, volatility tends to return to lower levels.

Another stylized fact is that markets are characterized by volatility spillover across assets and asset classes. To illustrate this feature, we plot in Fig. 15 the daily returns for several ETFs selected from the M6 investment universe (including the iShares Core S&P 500 ETF (IVV), the iShares Commodity Indexed ETF (GSG), the iShares MSCI Europe Small Cap ETF (IEUS), and the iShares iBoxx $ Investment Grade ETF (LQD)). Note that the time series of changing volatilities are similar across assets. This is particularly evident for the spike in volatility associated with the Covid-19 pandemic, and increased volatility during the M6 competition period associated with the Russian invasion of Ukraine.

A further stylized fact concerns the common practice of fitting volatility models on an asset-by-asset basis. This is done despite empirical evidence suggesting that fitted model parameters often cluster in a reasonably tight range. The models described in Bollerslev et al. (2018), for example, take advantage of this feature. They justify the use of a common volatility model for many assets and asset classes by showing how similar the distribution of realized volatility becomes when scaled by each asset's

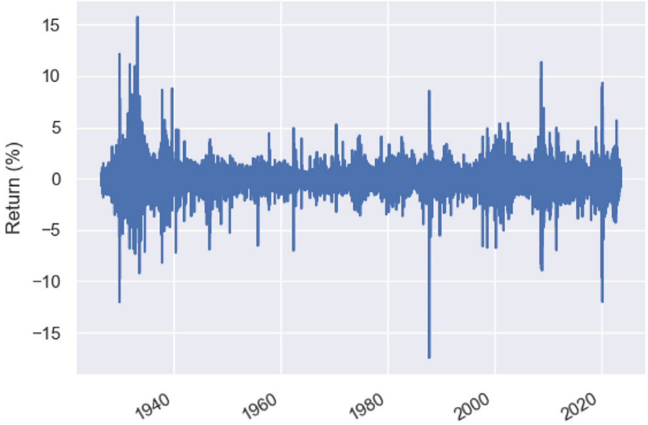


Fig. 14. Daily US stock market returns from 1926. The chart shows daily returns for the US market from 1926.
Source: Sourced from the data library of Kenneth R. French (2023).

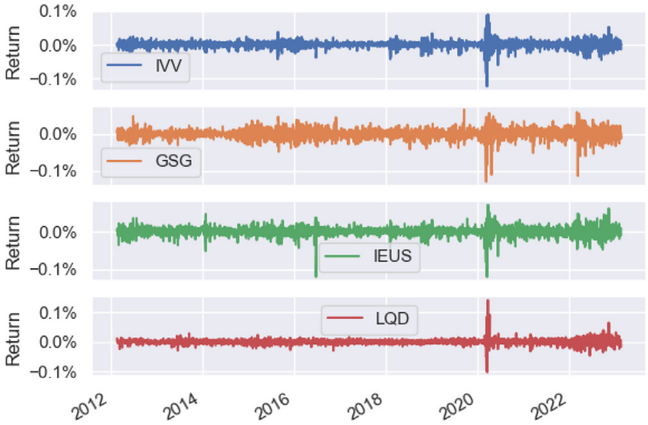


Fig. 15. Daily return time series from 2012 for selected M6 ETF securities: iShares Core S&P 500 ETF (IVV), iShares Commodity Indexed ETF (GSG), iShares MSCI Europe Small Cap ETF (IEUS), and iShares iBoxx $ Investment Grade ETF (LQD).

long-term or "expected" volatility. We repeat their analysis for the same subset of four M6 ETF assets discussed above, and plot our results in Fig. 16. Note that the resulting realized volatility distributions closely mirror those examined by Bollerslev et al. (2018).

Turning now to the specification of volatility models, we first consider the case of univariate models. These are models used to describe the risk patterns associated with an individual asset or market aggregate. A simple (perhaps overly simplistic) approach to this is to use a short-term (for example, 20-day or 100-day) moving average of squared returns. A related but more sophisticated approach is to use an exponential-smoothing-type model, again on squared daily returns. Here, variance is modeled as an exponentially weighted moving average, so that for instance, if the location of a time series y_t is modeled as a function of some vector of predictors x_t, then

$$y_t|F_{t-1} \sim [x_t'\beta, \sigma_t^2] \text{ and } \sigma_t^2 = (1-\gamma)\sigma_{t-1}^2 + \gamma e_{t-1}^2, \quad (10)$$

where F_{t-1} denotes a conditioning information set that includes data up to time period $t-1$, $[x_t'\beta, \sigma_t^2]$ denotes a distribution with mean μ and standard deviation of σ_t, β and γ are fixed parameters that must be estimated, and e_t is a stochastic disturbance term (error). Such models are often fitted assuming that the errors are non-Gaussian (see Jondeau et al. (2007) for a comprehensive treatment of these models). A major (and Nobel Prize-winning) advance in univariate modeling of volatility was made with the introduction of the auto-regressive conditional heteroskedasticity (ARCH) model introduced by Engle (1982). ARCH models (and generalized ARCH models, called GARCH models) are time varying, capture volatility clustering, and are widely used in industry. The ARCH(p) model is specified as above, but with

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i e_{t-i}^2. \quad (11)$$

The GARCH(p,q) generalization of this model is due to Bollerslev (1986), who noted that current conditional volatility (σ_t^2) is likely to depend not only on lagged squared errors (e_t^2) but also on lagged conditional volatility, leading to the following formulation for conditional volatility:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i e_{t-i}^2 + \sum_{i=1}^{q} \beta_j \sigma_{t-i}^2. \quad (12)$$

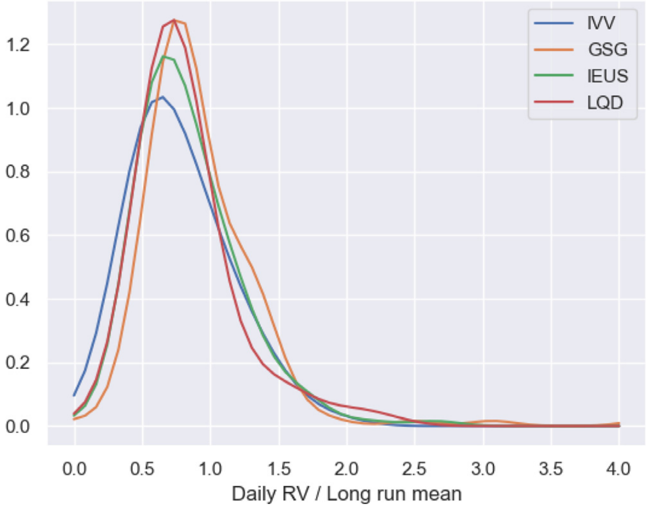


Fig. 16. Kernel density estimates for the four ETF securities displayed in Fig. 15, above. The chart plots kernel density estimates of the distribution of daily realized variance scaled by its long-run average for each asset.

The GARCH model has been generalized considerably, with many useful variants having been developed by many authors, and the GARCH(1,1) variant thereof has seen considerable empirical success when fitted to historical data (see Hansen and Lunde (2005)). In general, Bollersev's 1986 paper led to voluminous literature on time series volatility estimation and an alphabet soup of ARCH/GARCH and related models for modeling both univariate and multivariate time series; see Bollerslev (2009).

Other univariate volatility models are also used by some practitioners. A key example is the stochastic volatility model in which a system of stochastic differential equations is used to describe the return and volatility of an asset. These models are specified in continuous time, and require high-frequency intra-daily data for estimation. They are discussed in detail in survey paper appearing in this special issue; see also Andersen et al. (2001). For a discussion of stochastic volatility modeling using Bayesian methods, see Triantafyllopoulos (2021) and Prado et al. (2021). For further discussion that focuses on the granularity of data used to estimate volatility models, see Garman and Klass (1980), Parkinson (1980), Rogers and Satchell (1991), and Yang and Zhang (2000).

Needless to say, multivariate volatility models are also crucial to portfolio management. In particular, it is important to consider cases where n is large and t may be small. Possibly the simplest approach to multivariate modeling is to use exponential smoothing on the cross-products of daily returns for the universe of assets (see Brockwell (1991)). This approach has the attraction of needing only one exponential smoothing parameter, which can be estimated from the data or chosen by the researcher. Conceptually similar approaches extend the univariate GARCH-type models discussed above to the multivariate setting. For example, see the Baba, Engle, Kraft, and Kroner (BEKK) model developed in Engle and Kroner (1995) and the dynamic conditional correlation (DCC) model of Engle (2002) and Engle (2009). Both of these models can be parameterized heavily or simplified so that very few fitted parameters are required.

A particular modeling approach which considerably simplifies model fitting and leads to parsimony in the underlying model is to use a strategy known as variance targeting (univariate models) or covariance targeting (multivariate models), as described by Engle and Mezrich (1996). To give a univariate example, consider the GARCH(1,1) model:

$$\sigma_t^2 = (1-\alpha-\beta)\sigma_0^2 + \alpha e_{t-1}^2 + \beta\sigma_{t-1}^2, \tag{13}$$

where σ_0^2 is a long-run estimate of the variance to which the process tends to revert. We utilize a variant of this model in our analysis below.

When n is large, components of modern multivariate model building involve key (Bayesian) ideas including variable selection, dimension reduction, and parameter shrinkage that are also common in the machine learning literature. Indeed, modern portfolio optimization approaches which use the full covariance matrix without the application of some or all of these ideas are virtually nonexistent, to the best of our knowledge; see Black (1992) for an early discussion of this topic.

Arguably the earliest and perhaps currently the most popular and successful approach to achieve dimension reduction involves specifying and estimating a so-called factor model. In the context of factor models, common factors are assumed to underlie the co-movements of a set of variables, such as asset volatility, where the number of factors is k, with $k \ll n$. The idea is to achieve dimension reduction by modeling the covariance matrix of n assets as a function of k factors. When carrying out dimension reduction for the purpose of volatility estimation using returns, factor analysis yields a set of factor returns (sometimes orthogonal), and can be used to easily estimate a covariance matrix. For example, one might consider a model such as the following:

$$y_t = F_t\Delta + e_t, \tag{14}$$

where y_t is an $n \times 1$ vector of asset returns, F_t is an $n \times r$ matrix in which each row collects the r values of each of the r common factors that are associated with each of the n returns, Δ is vector of factor loadings, and e_t is a vector of stochastic disturbance terms, such that $e_t \sim [0, \Omega]$, where estimates of Ω can be constructed after jointly estimating the factors and loading coefficients. This framework is convenient. For example, when Ω is diagonal, it is easy to estimate $\Sigma_y = \mathbf{F}\Sigma_F\mathbf{F}' + \Omega$. A simple one-factor model (i.e. set $r = 1$), where the single factor is a latent "market" portfolio and a security's loading is referred to as its "beta" underpins most of modern portfolio theory. In practice, more useful models tend to have more factors, and complicated dynamics can be assumed to characterize the stochastic disturbance term. For further discussion of estimation of these models, as well as more sophisticated variants thereof, refer to Swanson and Xiong (2018), Cheng et al. (2021), and Liao and Fan (2011). In addition, Connor (2019) describes several other conceptual approaches to building such models.

A parallel approach to calibrating return forecasts is outlined by Grinold and Kahn (1999). The idea is to use a linear Bayes approach (Goldstein & Wooff, 2007) to preprocess forecasts into a form compatible with a given risk model. This approach borrows from the analysis of expert opinion; see West (1992) and West and Crosse (1992). Here, relative return forecasts are shrunk towards those of a benchmark. (In practice, the benchmark is usually set up to have zero expected return.) We adopt this approach in the sequel.

For complete technical details describing the specification and estimation of our investment risk model, refer to the appendix.

6.3. Findings based on application of our investment risk model to the M6 competition

Bringing it all together, we list a set of key features of the M6 competition based on the application of forecasts associated with our investment risk model. First, however, we explore how well the risk model captures the realized returns of the submitted investment portfolios. We do this by calculating the realized variance of each submitted portfolio over each 20-day test period, and then comparing this with our (model-based) forecast volatility as at the submission date. Fig. 17 plots outcome volatility vs. actual volatility for each portfolio at all submission points. We note that there is a relatively stable relationship between risk forecast and risk outcome, with a margin of error increasing steadily with the level of risk assumed. It is also clear from this figure that participants submitted portfolios with risk profiles substantially higher than those typically associated with institutionally managed portfolios.

Now we consider the distribution of forecast portfolio volatility forecast errors (i.e. forecast volatility – actual volatility) set out in Table 9. We do this separately for all submitted portfolios, and for portfolios with a more conventional risk profile. (Here, we choose a realized volatility of less than 10% to denote conventional risk.)

Table 9
Distribution of volatility forecast errors for M6 portfolios.

Portfolios	Mean	SD	25%	50%	75%
All	1.97	5.36	−1.3	1.37	4.89
Conventional risk	−0.05	2.35	−1.36	0.07	1.39

These results demonstrate a negative bias (portfolio risk is on average under-forecast by 1.95%) across all portfolios, albeit with a substantial margin of error. Focusing on the more conventional portfolios, we note that this bias all but disappears.

We now explore how well our model forecasts volatility for individual assets. An interesting exercise in this context involves examining (annualized) volatility forecasts for the S&P 500 ETF (IVV), as these can be compared to a "market" forecast (i.e. the Short Term VIX Futures ETF (VXX) provides a market price for S&P 500 annualized volatility approximately one month ahead).[10] Consider the volatility plots in Fig. 18. Evidently, the volatility forecasts embedded in the VXX index were somewhat too high for the period, as were, to a lesser extent, forecasts from our risk model.

We now summarize some of the key features of the M6 competition that arise when comparing the results from our model with the participants' competition entries.

Risk profile of the M6 portfolio submissions

As shown above, the participants in general assumed substantial levels of investment risk in their portfolio submissions. It is of interest to examine the levels of risk assumed in more detail. Throughout this discussion, and unless otherwise noted, we focus on active M6 *submissions* made as at a particular point (as opposed to portfolio entries carried forward from previous submission points).

First, we examine the levels of market exposure. Note that competition guidelines were such that, to be eligible for prizes, participants needed to have an absolute level of between 25% and 100% of their notional assets invested at each submission point. Without use of a risk model, we can examine the degree of net market exposure (the sum of all positions) and gross market exposure (the sum of the absolute value of all positions) as a first approximation to the level of aggressiveness of a portfolio. Managing gross portfolio exposure is generally the simplest tool to control risk in a hedge-fund-style portfolio. Exposure figures are contained in Fig. 19.

Inspection of this bar chart indicates that participants generally maintained a significant gross portfolio exposure throughout the competition, although total market exposure tended to decline throughout, over time. Recall that the participants were not required to assume any directional market exposure. Focusing on cash exposure takes no account of predictable variation in the level of volatility of each asset, nor of the correlation between assets. We can sharpen this analysis by using the risk model to make a direct forecast of portfolio risk using our estimates of these quantiles. Fig. 20 shows the evolution

[10] https://www.spglobal.com/spdji/en/indices/indicators/sp-500-vix-short-term-index-mcap/#overview.

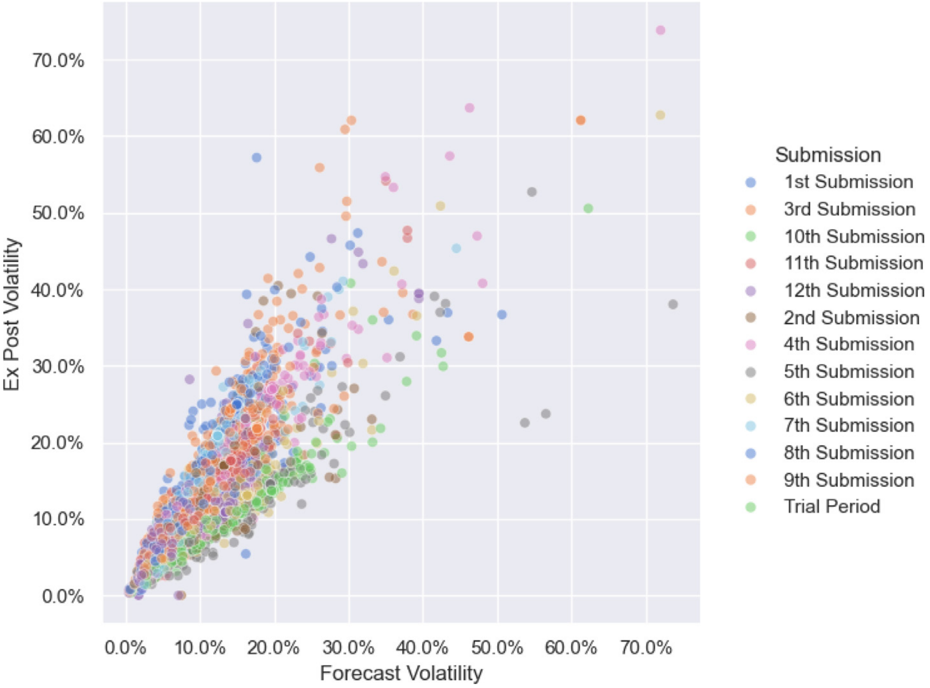


Fig. 17. Forecast M6 portfolio volatility compared to realized 20-day volatility for all M6 submissions. We estimate the M6 portfolio ex ante using our risk model, and compare this to ex post volatility 20-day portfolio returns.

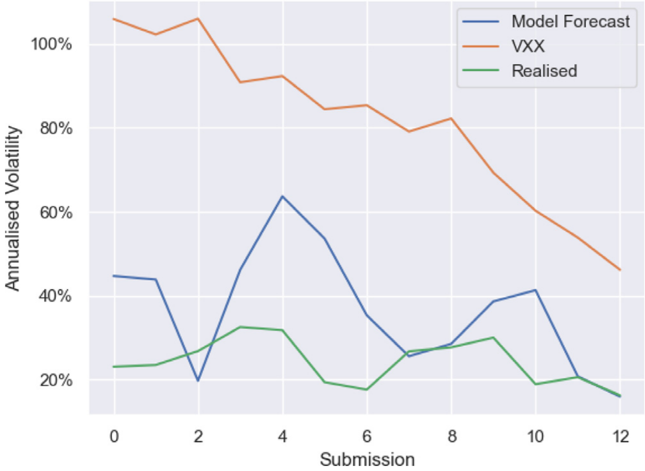


Fig. 18. Volatility forecast for the IVV S&P 500 ETF as at each submission point from our risk model. "Market" forecast of S&P 500 volatility based on the price of the VXX ETF, along with realized ex post daily volatility (measured over each separate 20-day evaluation period).

of the level of risk in the submissions across the course of the competition. Here, we plot the level of risk assumed for the 25th, 50th, and 75th quantiles of the distribution of all portfolios (as opposed to fresh submissions made at each point). It is evident that the risk taken by the participants is constantly much higher than the predicted one.

Further insight can be gleaned from analyzing portfolio risk decomposition using our model. We are able to decompose forecast portfolio variance by the source of risk. In the following, we do so by splitting *variance* (which is additive, whereas of course standard deviation is not) into four components: exposure to the M6M factor, exposure to other systematic factors, specific risk (uncorrelated with systematic factors), and a covariance effect. These results are summarized in Fig. 21. On initial inspection of these results, we suspected that the change in risk profile during the competition may have been driven by the rolling forward of previous submissions that were not being actively updated by participants (see Table 3). However, although we repeated the analysis for explicitly updated submissions, we discovered a similar pattern to that shown in Fig. 21. We see from this chart that risk contribution from diversification tended to increase across

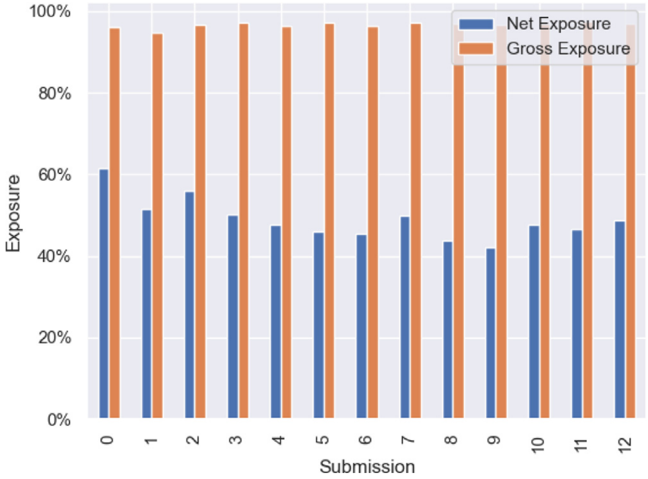


Fig. 19. Net and gross cash exposure of M6 portfolios by submission date. Net exposure is the sum of all portfolio weights. Gross exposure is the sum of all absolute portfolio weights. Submission 0 corresponds to the trial period.

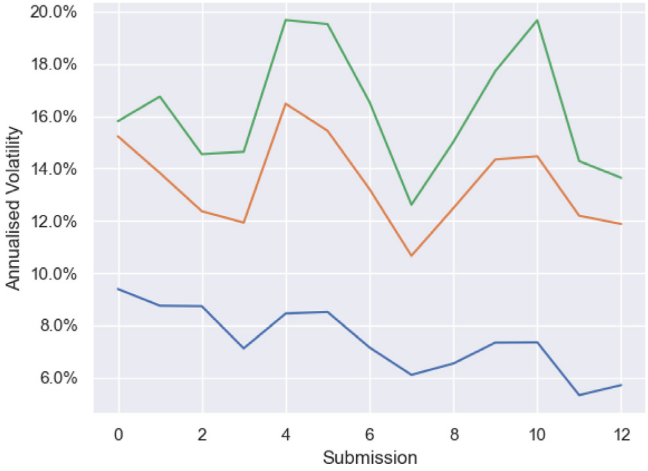


Fig. 20. Evolution of forecast volatility for M6 portfolio submissions. At each submission point, we produce a risk forecast for each M6 portfolio. The chart illustrates the 25th, 50th (median), and 75th percentile of this cross-sectional distribution at each submission point. Submission 0 corresponds to the trial period.

the course of the competition, whereas the proportion of specific risk assumed by participants remained broadly constant. This finding leads to three questions about how participants chose to structure their portfolios: How much investment risk did participants assume? Where did this risk come from? And how did this change over time? We address these questions next.

Portfolio optimization based on forecast submissions

Here, we examine whether participants could have improved the performance/Sharpe ratios of the portfolios they submitted using the risk information from our model. We construct notional portfolios using only the forecast information submitted by participants and the risk estimates from our model. We compare the characteristics and out-of-sample performance of these notional portfolios to the portfolio submissions made by competition participants. To do so, we proceed as follows, largely adopting the procedure set out by Grinold and Kahn (1999).

First, we score each asset at each submission point by taking the product of the vector $[1, 2, 3, 4, 5]$ with the vectors containing the forecast submissions of each participating team. We then standardize these scores across assets for each participant. In several cases, participants submitted identical scores for each asset, meaning that the cross-sectional standard deviation of the scores becomes equal to zero. In such cases, we replaced the standard deviations of the scores by the standard deviation of the set $[1, 2, 3, 4, 5] \approx 1.41$.

Second, we assume an investment coefficient (IC). This is a number representing the expected correlation between scores and outcomes. It is a measure of the "edge" of an investor—the extent to which the investor's forecasts add value relative to the consensus—and plays a key role in the mechanics set out by Grinold and Kahn (1999).

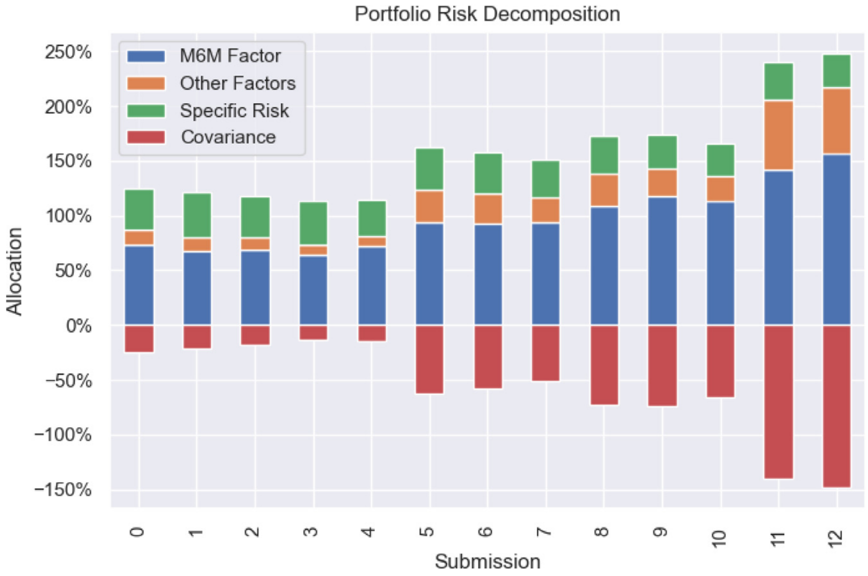


Fig. 21. Alongside each risk forecast used in Fig. 20, we decompose the corresponding portfolio *variance* in to four components, as described in the main text. The chart shows the time series evolution of this decomposition of risk for the distribution of M6 portfolios. Submission 0 corresponds to the trial period.

Because competition among investors makes active management difficult, ICs tend to be close to zero in practice, although efficient implementation and careful risk control can extract value from even a small positive IC. The interested reader is referred to Grinold and Kahn (1999) for extensive further discussion. For the analysis below, we assume that the IC takes three values—0.05 (moderate), 0.1 (good), and 0.15 (exceptional)—and produce portfolios based on each value.

Third, we take the (annualized) expected 20-day volatility $\hat{\sigma}_{20}$ for each asset. The linear Bayes refined return estimate for each asset $E(r_i|g_i)$ is now a function of the prior expectation r_i (which we take to be zero), and its forecast g_i for each asset/participant is given by

$$\alpha_i = \mathbb{E}[r_i|g_i] = \mathbb{E}[r_i] + Cov[r_i, g_i]Var^{-1}[g_i](g_i - \mathbb{E}[g_i]). \quad (15)$$

We have $E(r_i) = 0$, and the second component on the right-hand side of Eq. (15) can be written as $IC \times \sigma_{20} \times Score_i$, where $Score_i$ is the standardized forecast $((g_i - E(g_i))/Std(g_i))$, and IC is the assumed correlation between r_i and g_i, as described above. This procedure has the effect of shrinking the return forecasts for the assets towards zero. The degree of shrinkage is governed by the value of IC, and appropriate choices of this hyperparameter guard against overly risky portfolios. We then choose a set of weights, w, to optimize the portfolio:

$$\arg\max_{w}(\mathbf{w}\boldsymbol{\alpha}/\mathbf{w}\boldsymbol{\Sigma}\mathbf{w}'), \quad (16)$$

where all quantities are annualized. We use the sequential least-squares algorithm in the Scipy Python library, with first derivative of the objective function calculated via the JAX Python library to perform our calculations. We illustrate the results of this exercise by displaying statistics for portfolios aggregated according to the realized IC for each set of submitted forecasts. We estimate the IC for each submission by calculating the rank correlation of the scores with the subsequent returns across each evaluation period for all competition assets. We then group all submissions into quintiles based on realized IC, for each submission period. In Table 10, we report the median risk, returns, and information ratios for each group, for all cases where our optimization software returned a flag reporting that the procedure had terminated successfully.

It is clear from a cursory inspection of these results (see Table 10) that the optimization routines produced portfolios with substantially lower levels of risk than those assumed by the competitors, and, consequently, returns of lower magnitude. In cases where the original forecasts were of low quality, the portfolio optimization routine faithfully mapped these forecasts into underperforming portfolios, with substantially lower information ratios than the original portfolios. It appears that if underlying forecasts are poor, then sub-optimal implementation is, in-fact, optimal. Nevertheless, Table 10 shows that, given a set of reasonably accurate forecasts (IC ≥ 0.05), the IR can be substantially improved.

Portfolio optimization based on forecast and portfolio submissions

As noted, the original set of optimal portfolios described above differed materially in risk profile from the submissions made by the M6 participants, and although the optimal portfolios delivered superior risk-adjusted returns, their absolute returns were relatively low in magnitude, reflecting the level of risk assumed. In an institutional context, portfolios with such low levels of risk are unlikely to generate returns justifying an economically relevant management fee. We therefore undertook a second optimization-based analysis. For this exercise (see Table 11) we again took as inputs the return forecasts submitted by the participants, but supplemented

Table 10
Sharpe ratio optimized portfolios. The table sets out the (median) realized returns and volatility of portfolios constructed using our risk model and optimization procedure, compared to those submitted by competition participants. For full details, see the main text.

IC quintile	Realized IC	Submission ex ante risk (%)	Optimal ex ante risk (%)	Submission return (%)	Optimal return (%)	Submission IR	Optimal IR
1	−0.24	10.8	0.8	−12.4	−3.9	−1.4	−3.2
2	−0.08	11.7	0.9	−5.2	−1.6	−0.8	−1.4
3	0.00	11.9	0.9	2.5	0.0	0.5	0.5
4	0.08	11.5	0.9	4.7	1.4	0.7	1.4
5	0.22	10.3	0.9	16.5	3.7	2.5	3.3

Table 11
Sharpe ratio optimized portfolios, with risk-level targeting. The table sets out the (median) realized returns and volatility of portfolios constructed using our risk model and optimization procedure, compared to those submitted by competition participants, grouped by the quintile ranking of the submission IC. In this instance, the optimizer was configured to target the levels of portfolio volatility assumed by participants. For full details, see the main text.

IC quintile	Realized IC	Submission ex ante risk (%)	Optimal ex ante risk (%)	Submission return (%)	Optimal return (%)	Submission IR	Optimal IR
1	−0.24	10.8	6.2	−12.4	−20.9	−1.4	−2.8
2	−0.08	11.7	5.9	−5.2	−7.9	−0.8	−1.2
3	0.00	11.9	5.9	2.5	−0.1	0.5	−0.1
4	0.08	11.5	5.8	4.7	8.3	0.7	1.6
5	0.22	10.3	6.1	16.5	25.4	2.5	4.3

this information with the estimated investment risk of each corresponding submitted portfolio. We then added an additional constraint to the optimization routine, such that the resulting optimal portfolio had a risk level at least equal to that of the submission made by each participant. Note that the optimization problem as set up for this exercise is more challenging than that conducted above, and submissions corresponding to more extreme portfolios are removed from this sample. Moreover, note that the results of Table 11 involve optimal portfolios that were constructed with returns calibrated using a "good" IC value of 0.1. (The results for "moderate" and "exceptional" ICs differ very little from those reported below.)

Evidently, the optimizer failed to obtain the objective of matching the risk profile of the submitted portfolios in many cases. An investigation revealed that this was the case mostly for instances when submitted portfolios had high or very high ($>$ 10%) risk forecasts. Despite risk profiles which remained substantially lower than those of the portfolio submissions, where underlying forecasts were of good quality, the optimal portfolios outperformed the submissions in terms of return, and consequently exhibited higher information ratios. In this sense, the optimizer displays the same efficiency noted above in turning poor-quality forecasts into wealth-damaging portfolios.

Reverse optimization

Finally, we attempt to deduce the set of asset return forecasts that would, given the risk model, make each set of portfolio submissions made by the M6 participants optimal. We then compare these sets of implied asset return forecasts to the forecast submissions. In order to do this, we set sensible upper and lower bounds on implied alphas for each asset. We calculate these using an implied IC of 0.3, a score of three standard deviations, and a volatility estimate from the risk model. This ensures that for a given score, more volatile assets have higher expected returns. We then optimize to choose a set of α based on the following functional optimization:

Table 12
Reverse-optimized portfolios. Implied portfolio α correlation with actual submitted forecasts. See the main text for full details.

Mean	SD	25%	50%	75%
0.32	0.38	−0.02	0.29	0.64

$$\arg\max_{\alpha}(\mathbf{w}\boldsymbol{\alpha}/\mathbf{w}\boldsymbol{\Sigma}\mathbf{w}'). \quad (17)$$

Subsequently, we take the values returned from this optimization, rank them from 1 to 5, and set our forecast submission equal to 1 for the rank of each asset. We first examine the cross-sectional correlation of the reverse-optimized scores with those actually submitted by participants. In particular, for each submission, we calculate the correlation coefficient between the submitted and reverse-optimized rankings. The results are given in Table 12. These findings indicate a relatively weak but positive relationship between the two sets of rankings, suggesting that the participants' portfolios reflect, on average, their forecasts. This is confirmed by a highly significant t-statistic of 117 on the reverse-optimized α obtained by regressing the submitted forecasts against this (plus a constant) across all submissions.

6.4. Analysis of winning submissions

In this section, we briefly analyze the performance of the winning submissions. We separately analyze the top 10 competitors by RPS and IR. We find that the top-ranked forecasters (best RPS) were in general well calibrated in terms of investment decision making. Namely, they submitted portfolios with lower-than-average levels of investment risk. Our optimized portfolios were still able to demonstrate significant improvements in IR, however, by virtue of both reduced risk and increased returns (see Table 13).

Table 13
Statistics for the top 10 competitors ranked by RPS across the entire competition. The table shows the median ex ante & ex ante volatility, ex post return, and IR for the top 10 competitors raked by RPS (across the entire submission history), compared to re-optimized portfolios constructed using the risk model and the submitted set of forecasts.

	Ex ante volatility (%)	Ex post volatility (%)	Ex post return (%)	Ex post IR
Submission (Mean)	7.2	8.4	5.3	0.23
Re-optimized (Mean)	5.9	6.5	5.6	0.87
Submission (Median)	4.5	5.7	0.9	0.37
Re-optimized (Median)	4.9	5.8	2.6	0.50

Table 14
Statistics for the top 10 competitors ranked by IR across the entire competition. The table shows the median ex ante & ex ante volatility, ex post return, and IR for the top 10 competitors raked by IR (across the entire submission history), compared to re-optimized portfolios constructed using the risk model and the submitted set of forecasts.

	Ex ante volatility (%)	Ex post volatility (%)	Ex post return (%)	Ex post IR
Submission (Mean)	9.4	10.4	32.0	2.05
Re-optimized (Mean)	5.8	6.7	5.2	0.65
Reverse α (Mean)	6.3	6.1	6.7	0.92
Submission (Median)	6.5	8.2	6.0	1.15
Re-optimized (Median)	5.3	5.7	−0.0	−0.12
Reverse α (Mean)	5.5	5.3	1.0	0.18

We now turn to the winning IR portfolios. We compare the submitted portfolios to the re-optimized portfolios constructed in two different ways. First, we re-optimize based on submitted return forecasts for each submission. Second, we re-optimize based on the reverse-optimized α, as described above. Once more, winning teams assumed lower levels of portfolio risk than the average participant. We also find that the performance of these portfolios was in general not well explained either by the explicit return forecasts submitted by participants, or by the forecasts implied by the starting portfolio weights. To an extent, these participants may have benefited from luck in their investment decision making: there is little evidence to support the hypothesis that their positive IRs were generated by superior investment insights (see Table 14).

6.5. Final remarks based on the findings of our investment risk model

The M6 competition guidelines explicitly asked competitors for return forecasts on a selection of assets, and for investment portfolio decisions. However, the link between the two parts of the competition was not specified, and competitors were left to make their own choices as to how to tie forecasts to investment decisions. Still, some incentive was given to participants to properly link the two parts of their submission, given that fiduciary prizes were allotted to the best forecasts, investment decisions, and overall performance. In this section, we discussed a risk model that allowed us to assess the strength of the connection between the participants' forecasts and investment decisions. Our findings can be summarized as follows. We find that participants assumed significantly greater risk in their portfolios than was justified by the accuracy of their forecasts. For the most accurate sets of return forecasts, portfolio optimization along with our risk model was able to produce portfolios with substantially better returns and much lower levels of risk. On the other hand, when return forecasts were poor, our model-based optimization efficiently translated these into wealth-destroying portfolios. This is because assigning appropriate levels of risk to assets that are poorly forecast leads to substantially sub-optimal investment-portfolio weight selection.

Our overall characterization of these results is that sub-optimal portfolio construction added a (large) additional random component to the achieved portfolio returns. Some "lucky" participants made relatively poor forecasts but benefited when this random component turned out to be beneficial, ex post. Other competitors made relatively accurate forecasts, but poor implementation and "bad luck" (negative ex post return realizations) led to poorly performing portfolios.

7. Major findings and insights

Below is a summary of our key findings based on the analysis of the M6 submission data, the evaluation of the 10 hypotheses, the performance of the top-ranked methods, and the results obtained through our risk model:

Finding 1: The challenging task of forecasting the relative performance of assets.

Due to the volatile nature of tradable assets, predicting future prices is a hard task. The M6 competition focused on a potentially easier task—that of forecasting the relative performance of the assets. Unfortunately, our results clearly demonstrate that producing such forecasts is also far from straightforward.

The benchmark selected for the forecasting track of the competition assigned equal probabilities to all five quintiles, assuming that all assets are equally likely to yield relatively higher or lower percentage returns. Although simplistic, less than 25% of the teams managed to estimate the requested probabilities more precisely than the benchmark. Moreover, those that did improved accuracy by less than 2.5% and found it difficult to outperform the benchmark consistently. Not surprisingly, then, we found that only three teams performed better than the benchmark in all 12 months of the competition, and less than 15 teams in more than nine months. Our analysis also showed that participants had unwarranted certainty concerning the precision of their forecasts.

Note that this finding does not suggest that the accuracy improvements achieved by the top-performing teams were minor. On the contrary, it highlights that the inherit uncertainty of the requested forecasts makes any improvement challenging to achieve, particularly in the long run. It is encouraging that even minor accuracy gains can result in significant improvements in terms of IR. For instance, our analysis showed that re-optimizing the portfolios of the top 10 teams according to RPS based on the developed risk model improved average IR values from 0.23 to 0.87. Similarly, our results indicated that re-optimizing portfolios with high IC values can result in better IR values, which can be realized either by taking more reasonable risks or by optimally translating forecasts into investment weights, given a certain level of risk. Interestingly, in the latter case, both the IR and the returns can be improved.

Finding 2: The difficulty of consistently outperforming the market.

Although testing the EMH using a single sample of teams and a set of submissions that cover a period of one year is challenging, our results confirm that beating the market is particularly difficult.

Focusing on the global investment performance of the teams, about 60% of the participants realized negative returns that usually exceeded 7% and reached up to 46%. In addition, less than one-third of the teams managed to outperform the benchmark, and only a very small group of participants managed to consistently outperform the market. It is interesting that none of the teams achieved higher IR values than the benchmark in all 12 months of the competition, and just four beat the market in more than eight months.

Our analysis, apart from confirming the EMH, also demonstrates that the risks taken by participants were rarely justified by the returns of the constructed portfolios, and that when there were any positive returns, these were often marginal. The fact that the majority of the teams tended to perform better than the benchmark when the market was bearish is also interesting, as it suggests that most of the participants failed to effectively short assets when losing value or to construct portfolios that were robust in bear markets.

On the positive side, some teams managed to beat the market to a significant extent, realizing impressive returns that surpassed 30%. These rare exceptions may have been due to luck, though.

Finding 3: The limited connection of the submitted forecasts and investment decisions, as well as the potential benefits of their association.

One of the primary objectives of the M6 competition was to investigate the value that accurate forecasts can add to investment decisions. In this regard, participants were asked to predict the relative performance of tradable assets and, possibly based on said predictions, construct portfolios.

The analysis conducted to test Hypothesis No. 3 demonstrates that a limited number of teams actually chose to exploit their forecasts to define the weights of their investments. When analyzing the complete set of submissions, we found no connection between the RPS and IR. Moreover, although we found some association between the two scores for the top 20% of the teams (according to their OR), this connection diminished when more than 40% of the teams were considered or when considering only the duathlon winners. Even more interesting is that the top-performing teams in the forecasting challenge constructed relatively inefficient portfolios on average, while the top-performing teams in the investment-decisions challenge submitted forecasts of various accuracy levels.

The analysis of Hypothesis No. 4 is also relevant to this topic. By investigating the investment weights assigned to assets that were predicted more accurately, we found that there is no evidence that the top-performing teams (according to the IR) built their portfolios primarily using assets that they could forecast more accurately. In fact, the weights assigned to all assets were similar on average, regardless of how accurately the participants were able to forecast their relative performance. Interestingly, this was true even for the teams that, overall, tended to assign higher/lower weights to assets that were predicted to have higher/lower returns.

These findings indicate that many of the teams probably decided to work on the forecasting and investment tracks of the competition separately, using approaches that were not connected. Nevertheless, they do not imply that better forecasts cannot be used to yield more profitable investments. In fact, based on the analysis conducted using our risk model, it is evident that investment decisions can be improved, provided of course that risk is properly measured and that the forecasts are of adequate quality.

Finding 4: The value added by information exchange and the "wisdom of crowds".

Aggregating information from groups has been shown to result in better investment decisions than those made by individuals, and combining forecasts has long been regarded as one of the most successful strategies for improving predictive accuracy. The analysis conducted when testing Hypothesis No. 8 clearly supports these statements.

Our results show that the averages of forecasts submitted by the top-performing teams (according to the RPS) consistently outperform forecasts made by the best-performing participant. In addition, the accuracy of such combinations remains similar to that of the most accurate submission, even in extreme cases where about half of the teams are included in the combination. Finally, combinations always outperform the benchmark. In similar fashion, averaging the investment decisions of the top-performing teams (according to the IR) results in significantly better performance than the best-performing team, while averaging the investment weights of multiple teams outperforms the benchmark, at least when the worst-performing teams are excluded.

These findings highlight the robustness of forecast combination and decision aggregation, confirming the importance of the "wisdom of crowds".

Finding 5: The positive effect of adapting to change.

Before the competition started, we hypothesized that teams that employed consistent forecasting and investing strategies throughout the competition would perform better than those that changed their strategies significantly from one submission point to another. However, our results suggest otherwise. By focusing on the top-performing teams, we found that changing strategies proved to be beneficial. For example, the winning team of the investment challenge was found to have changed its strategy eight times, and most the top-performing teams changed their strategies more than once. We also found that teams that updated their submissions (i.e. their forecasts and investment weights) on a regular basis tended to perform better than those that did not. In particular, all five winners in the forecasting track and four of the five winners in the investment track updated their submissions every single round, while the same was true for most of the duathlon winners. Adapting based on external information and judgment was also shown to have a positive effect on forecasting and investment performance. Although few in number, judgment-informed forecasting approaches were found to perform better or similarly to pure data-driven approaches, suggesting that when judgment is utilized correctly, it can lead to good performance. Although these findings cannot be generalized for the complete set of teams and submissions, they do provide empirical evidence that adapting to change can have a positive effect on both forecasting and investment performance.

8. Concluding remarks and directions for future research

Following the rich and successful tradition of past Makridakis (M) competitions, M6 was an overall success. In this latest competition, our objective was to empirically investigate an interesting and specific question: How well do forecasters and investors predict various tradable assets and construct investment portfolios? The Makridakis forecasting competitions have always been at the forefront of forecasting research, pushing boundaries, setting research agendas, and shaping the field. The M6 competition was no different: its impact on the community is already evident.

8.1. Discussion

Compared to past forecasting competitions (Makridakis et al., 2021), we implemented three key changes in our experimental design. First, the M6 competition was live, meaning that participants collected the most up-to-date data and submitted forecasts and investment decisions in real time. Their forecasts were also evaluated in real time, prior to participants submitting subsequent predictions and investment weights. This cycle was repeated 12 times (over a calendar year) to achieve a certain level of robustness. Second, the nature of the forecasting task differed compared to past competitions. In the past, the objective was to submit forecasts for the future values of each target variable. In the M6 competition, the forecasting task involved forecasting the relative performance of different tradable assets. Third, this competition not only involved an evaluation of performance based on the forecasting task, but also directly evaluated the financial implications associated with using (or not) the forecasts in real time to make investment decisions.

One of the key results of the competition was the inability of participants to significantly outperform the forecasting benchmark (that assumed equal forecast probabilities) when predicting the probabilities of falling within returns quintiles for the various assets. Despite this fact, participants were able to form efficient portfolios that balanced returns with risk. Additionally, we observed a considerable disconnect between forecasting performance and investment performance, with almost zero correlation between the two, when including all participants in our calculations. Our empirical results confirmed the EMH: only one team outperformed the investment benchmark (which assumed an equally weighted portfolio, with only long positions) at the end of the competition, and none of the teams managed to outperform the investment benchmark across each of the 12 months of the competition. Another key finding that reiterated the results of past forecasting competitions is that the use of forecast combination (or aggregation, in the sense of using the “wisdom of crowds” when making investment decisions) is a simple but efficient way to improve performance. Finally, and again in line with insights from past M competitions, relatively simple, standard methods often result in good overall performance.

The disconnect between the performance when comparing results from the two tasks (i.e. the forecasting and investment challenges) was an interesting finding that was discussed during the presentation of our results at the 2023 International Symposium on Forecasting. Some attendees argued that it is likely that participants simply focused on making informed investments and avoided producing any forecasts, for example. We disagree with this argument. We believe that any decision, in this case the investment weights for the different assets, requires a forecast. This forecast is sometimes implicit and thus not fully quantified. However, it is still a forecast: one opts for a high weight and long position in a particular asset only because one implicitly forecasts that the asset will perform better relative to other available assets. In fact, by construction the M6 enabled the submission of either qualitative (i.e. pure judgmental) forecasts, quantitative forecasts, or any combination thereof. Not surprisingly, some participants bypassed the forecasting task completely, failing to efficiently record their forecasts, which implicitly or explicitly informed their investment decisions. Our results in Section 6 clearly show that if participants used an explicitly formulated and realistic risk model, then there would be a clear link between forecasting and investment performance.

8.2. Limitations

One of the limitations of this competition was its duration. The competition involved 12 non-overlapping submission points, with each being 28 days after the previous one. As a result, the competition ran for almost one year. This had a negative impact on the total number of participating teams as well as on the level of engagement of the participants. We anticipated this, and we attempted to mitigate its effects by offering substantial monetary awards for top-performing teams (both overall and per quarter). However, it was still the case that only a small percentage of the teams updated their forecasts regularly during the duration of the competition; many teams simply submitted once or twice at the very beginning of the competition. Could we have designed the competition such that it lasted less than a year to increase engagement and decrease dropouts? This might have been done by increasing the ''skin in the game'' by having participants invest money in their selected assets such that they benefited from any profits instead of setting prizes based on relative performance. However, we believe that for evaluating the EMH, it would have been better to run the competition for a longer time period, possibly 2–5 years, in order to be able to include significant cycles in the economy. In the end, we believe that while neither the competition's duration nor the achieved participation was perfect, we did achieve a good balance between these two features of the M6.

The second limitation of the M6 competition's design has to do with the selection of 100 assets used to form the investment universe. As with any empirical exercise that uses a finite and finely defined set of data, one can argue that the results of the competition are limited based on the data. In an ideal scenario, we could have included all possible investable assets available in the world. This, though, would clearly render the exercise impractical from the participants' point of view, as the forecasting and investment task would be vastly complex. Another possibility would be to just allow participants to focus (and provide forecasts of) specific subsets of the 100 assets (for instance, only stocks or only ETFs). We decided against doing so, as this would have inevitably led us to have more distinct evaluation categories (resulting in a further split of the prizes), as the forecasting performance of one subset (e.g. stocks) would not be comparable to that of another subset (e.g. ETFs).

The third limitation of the M6 competition was that, in addition to the duathlon prizes, we separately offered prizes for each of the tasks/challenges. This allowed participants to focus exclusively on one of the challenges (forecasting or investment), receive significant awards, and ignore the other challenge. Did this happen? Perhaps, as the top-performing teams in the forecasting task did not perform very well in the investment task. As a consequence, this limited the usefulness of our measures of the connectedness between forecasting and investment performance. Nevertheless, our post-competition analysis suggests that if the top performers in one challenge made use of state-of-the-art risk models, then they could have performed better in the other challenge.

The final limitation of the M6 competition has to do with the limited opportunities it offers for reproducing or even replicating the forecasts and investment decisions of the participating teams. Given the nature of the competition (financial forecasting), and in order to attract participation from expert financial analysts, we decided to ask only for a brief bird's-eye-view description of the participants' forecasting and investment approaches. The main reason for this was our belief that expert analysts would not participate in such a competition if they had to fully expose their approaches. Requiring the submission of a detailed description of each team's approach, or even the submission of the respective code, as a prerequisite for participation would have allowed us to replicate the submissions but at the same time would have decreased participation.

8.3. Directions for future research

We believe that the live component of the M6 forecasting competition offers a unique take on forecasting competitions. As organizers, we did not need to conceal the data or their source and did not need to define a suitable hold-out sample. Additionally, participants had access to all information related to the data and were able to apply apart both quantitative and judgmental approaches to the problem, either exclusively or in conjunction with formal models. An additional advantage of live competitions is that they allow us to be in a position to evaluate the domain expertise of participants in addition to their technical expertise. We feel that more forecasting competitions should move away from concealed data and hold-out samples, and should endeavor to set up live experiments that use real-time data. Associated challenges include data availability and the timed nature of the task (as forecasts need to be produced very quickly after actual data have been published).

In this and past M forecasting competitions, participants were required to submit outputs of their approaches (i.e. forecasts and decisions). An alternative way forward would be to ask participants to submit their solutions directly (i.e. code or executable files). While this would limit the scope of their approaches to strictly quantitative ones, it would offer a considerable advantage. Competition organizers would be able to run the participants' solutions directly, being in a position to evaluate their effectiveness over longer time periods, in a rolling origin manner. Also, the submission of solutions instead of outputs would guarantee the reproducibility of the submitted forecasts and decisions.

In this competition, we made a careful selection of 100 assets to create a representative and diverse universe of investment options (see Section 2.2 for discussion). This was a design choice that we made in order to render the competition viable by allowing participants to invest resources even in a judgmental fashion, should they wish to do so. Future work might analyze other datasets, including for example major indices such as the S&P 500 and NASDAQ.

Three of the most cited M forecasting competitions, M1, M3, and M4, offered diverse forecasting challenges, encompassing data from various domains (micro, macro, industry, finance, demographic, etc.) and a variety of frequencies (yearly, quarterly, monthly, weekly, daily, and hourly). Such competitions allow us to identify methods and approaches that are robust and efficient over a variety of settings and contexts. The M5 and M6 competitions instead focused on specific problems (i.e. retail and financial forecasting). We believe that focusing on particular industries allowed us to better identify the ''horses for the course''. We envisage future competitions focusing on specific industries, such as energy forecasting, pharmaceutical forecasting, and medical forecasting. Equally importantly, more generalist competitions could be designed that push the boundaries of innovation and engender the development of generically robust forecasting methods.

In closing, we would like to stress the value of the M forecasting competitions for bridging the gap between theory and practice, and we reiterate the importance of carrying out robust experiments and empirical evaluations such as those reported on in this paper. As we move forward, we invite companies and industries that face challenges with their forecasting tasks to consider M forecasting competitions as a platform of innovation designed to offer them a diverse pool of efficient forecasting solutions that are uniquely engineered for their data.

CRediT authorship contribution statement

Spyros Makridakis: Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Evangelos Spiliotis:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Ross Hollyman:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Formal analysis. **Fotios Petropoulos:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis. **Norman Swanson:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Formal analysis. **Anil Gaba:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Construction of the M6 competition universe of assets

The 50 stocks were selected so that each GICS (Global Industry Classification Standard) sector is represented in M6 in a similar proportion with that in the S&P500 index. In this regard, the number of assets to be sampled per sector was defined as shown in Table A.1.

In November 2021, the 9 key features presented in Section 2.2 were computed for each of the stocks, taking into account the average price of the stock and its volatility, the daily and compound returns, the volatility of the returns, as well as the trading volume. These features where then used to group the stocks of each sector into diverse clusters and make sure that the selected assets would sufficiently represent the market.

To do that, the silhouette method was first employed to define the optimal number of clusters per sector. Then, based on the population size of each cluster, an appropriate number of stocks was randomly sampled per cluster. As an example, Fig. A.1 showcases that the 21 stocks of the 'Energy' sector were grouped into two clusters of 5 and 16 stocks, respectively. Given that the examined sector should contribute 2 stocks in total to the M6 universe, the first cluster should provide $(5/21) * 2 = 0.48 \approx 0$ stocks, while the second $(16/21) * 2 = 1.52 \approx 2$ stocks. By randomly sampling two stocks from the second cluster, the COP and XOM stocks were included among the M6 assets.

ETFs were selected qualitatively with the objective of capturing the overall returns of major markets (e.g. US, UK, Germany, Japan, China, India, Australia, Brazil), involve short- and long-term treasury and corporate bond options, allow investing in major sectors (e.g. energy, infrastructures, technology, health), provide precious metals and commodity options (e.g. gold and silver), and include equity style options (e.g. value, momentum and minimum volume for the United States of America and Europe). The selected ETFs can be categorized according to their type as shown in Table A.2.

The 100 assets that were included in the M6 investment universe are presented in Table A.3.

Appendix B. Connection of forecasts and investment decisions

In order to quantify the degree to which the forecasts of the participating teams were connected with the corresponding investment decisions, we introduced a correlation measure, r_{con}, computed as follows:

- For each submission, we multiply the investment weight of each asset with the corresponding forecasts (probability that the asset will be ranked within the first, second, third, fourth and fifth quintiles). By doing this, we end up with 100 vectors, each containing 5 elements, {Rank1, Rank2, Rank3, Rank4, Rank5}, that represent the expected value of the investment. Elements of high positive and negative values suggest that the uncertainty around the forecast was relatively low or that the invested amount (either long or short position) was relatively large. Accordingly, elements that are close to zero indicate that uncertainty around the forecast was relatively high or that the invested amount was relatively small.
- We compute the average of the {Rank1, Rank2, Rank3, Rank4, Rank5} elements by aggregating their values across all assets and submissions made by a team throughout the competition.
- We calculate the correlation, r_{con}, between the averages computed in the previous step and a vector containing the following integer numbers: {1,2,3,4,5}.

If the forecasts of a team are connected with its investment decisions, r_{con} should be close to unity (larger amounts of capital are invested at higher ranked assets

Table A.1
Stocks included in the S&P500 and the M6 investment universe per sector.

Sector	S&P500		M6	
	Count	Proportion (%)	Count	Proportion (%)
Communication Services	27	5.3	3	6.0
Consumer Discretionary	63	12.5	6	12.0
Consumer Staples	32	6.3	3	6.0
Energy	21	4.2	2	4.0
Financial	65	12.9	7	14.0
Health Care	64	12.7	6	12.0
Industrial	74	14.7	7	14.0
Information Technology	74	14.7	7	14.0
Materials	28	5.5	3	6.0
Real Estate	29	5.7	3	6.0
Utilities	28	5.5	3	6.0

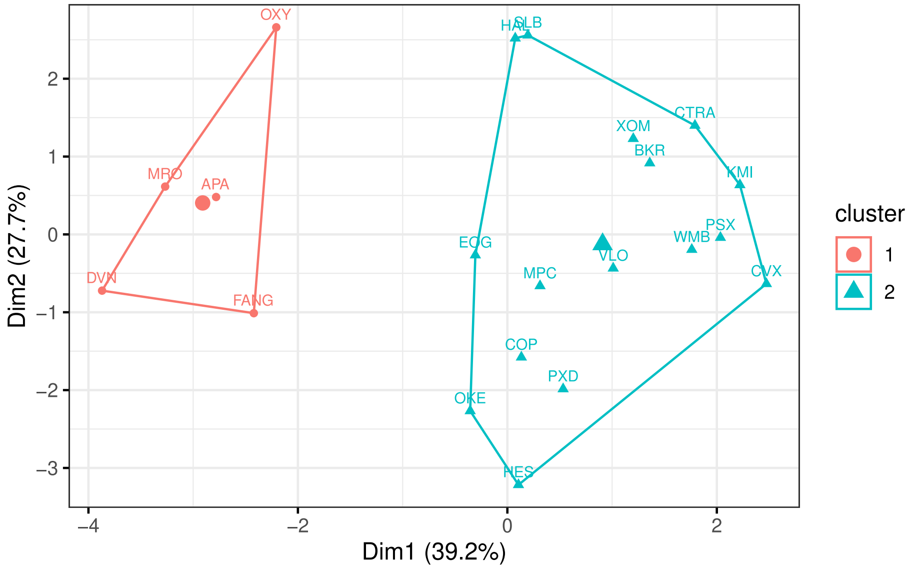


Fig. A.1. Example illustrating the clustering process taking place for the case of the "Energy" GICS sector.

Table A.2
ETFs included in the M6 investment universe per type.

ETF type	Count
Large Cap	15
Small Cap	2
Sector	14
Equity Style	6
Government	5
Credit	4
Precious Metals	2
Diversified Commodities	1
Volatility	1

and lower—or even negative—amounts of capital are invested at lower ranked assets). For instance, the following vector {−0.036, −0.022, 0.004, 0.019, 0.073} would score $r_{\text{con}} = 0.96$, while the following one {0.003, −0.019, −0.019, −0.007, 0.027} would score $r_{\text{con}} = 0.29$. In light of this, team submissions were classified into the following categories:

- 'Well connected' ($r_{\text{con}} \geq 0.75$);
- 'Connected' ($0.50 \leq r_{\text{con}} < 0.75$);
- 'Weekly connected' ($0.25 \leq r_{\text{con}} < 0.50$);
- 'Disconnected' ($-0.25 \leq r_{\text{con}} < 0.25$);
- 'Opposite connection' ($r_{\text{con}} \leq -0.25$);
- 'NA' (Teams whose forecasts were identical to those of the benchmark and, therefore, r_{con} could not be computed).

From the 163 teams included in the global leaderboard, 39 fell into the 'well connected' category, 20 into the 'connected' category. Additionally, 19 were classified as 'weakly connected', 42 as 'disconnected', and 16 had an 'opposite connection'. This result suggests that the majority of the teams developed separate approaches for preparing their submissions for the two challenges of the M6 competition, often making investments that can scarcely be justified based on their provided forecasts. Fig. B.1 further supports this finding, indicating that although some teams shorted assets that fell according to their predictions in the worst category (Rank1) and invested more capital into assets that fell according to their predictions in the best category (Rank5), in most of the cases similar investments were made across all five asset classes.

Table A.3
Assets included in the M6 investment universe.

The M6 Investment Universe (Stocks and ETFs)									
ABBV	ACN	AEP	AIZ	ALLE	AMAT	AMP	AMZN	AVB	AVY
AXP	BDX	BF-B	BMY	BR	CARR	CDW	CE	CHTR	CNC
CNP	COP	CTAS	CZR	DG	DPZ	DRE[a]	DXC	EWA	EWC
EWG	EWH	EWJ	EWL	EWQ	EWT	EWU	EWY	EWZ	FB[b]
FTV	GOOG	GPC	GSG	HIG	HIGH.L	HST	HYG	IAU	ICLN
IEAA.L	IEF	IEFM.L	IEMG	IEUS	IEVL.L	IGF	INDA	IUMO.L	IUVL.L
IVV	IWM	IXN	JPEA.L	JPM	KR	LQD	MCHI	MVEU.L	OGN
PG	PPL	PRU	PYPL	RE	REET	ROL	ROST	SEGA.L	SHY
SLV	SPMV.L	TLT	UNH	URI	V	VRSK	VXX	WRK	XLB
XLC	XLE	XLF	XLI	XLK	XLP	XLU	XLV	XLY	XOM

[a] On October 3, 2022 the DRE price stopped being updated due to the company being acquired by PLD. Therefore, participants had to assume a return of zero for DRE (any investment to DRE would have no effect on portfolio returns and the rank of the asset in terms of returns would have to be forecast assuming no price change).

[b] On June 9, 2022 the FB identifier changed into META.

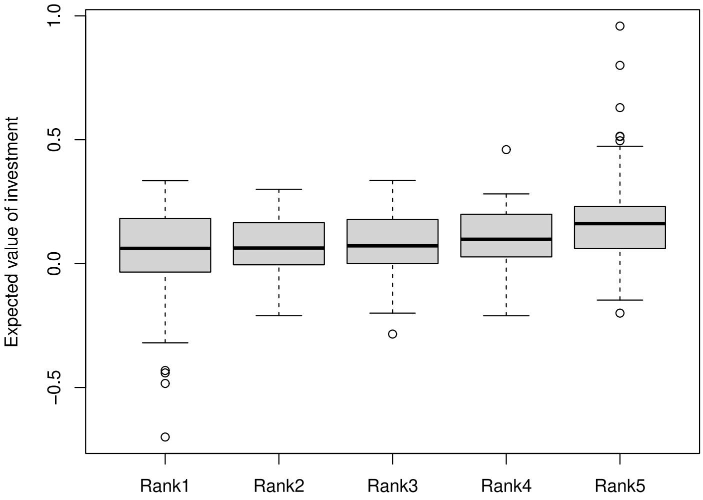


Fig. B.1. Distribution of the values computed for the {Rank1, Rank2, Rank3, Rank4, Rank5} elements (expected value of investment) for the 163 teams included in the global leaderboard.

Fig. B.2 provides further insights into the impact that the degree of connection between the forecasts and investment decisions had on team performance. Evidently, 'well connected' submissions tend to involve better portfolios than 'disconnected' and 'opposite connection' portfolios. However, 'weekly connected' submissions perform similarly well and the IR values largely overlap across the individual classes. The differences in terms of performance are even smaller when it comes to RPS. In fact, we observe that 'well connected' submissions involve some of the worst sets of forecasts, in contrast to 'weekly connected' submissions, that perform much better on average.

Appendix C. Technical details of the risk model

In this appendix, we discuss some of the technical details associated with the estimation of our investment risk model. Before discussing estimation of the key building blocks of the model (i.e., volatility forecasts) that are used in portfolio optimization when creating pseudo-true portfolios for comparison with M6 participant portfolios, however, we first briefly discuss the historical dataset used for this purpose. Our data include daily prices for the cross section of M6 assets, collected from Yahoo Finance. More specifically, we analyze 10 calendar years of daily data, ending on the start date of the competition, (i.e., February 2, 2022). The data comprise Open, Close, High and Low prices, as well as Adjusted Prices, accounting for splits/corporate actions, dividends etc. When building our dataset, we attempted to select a broad cross section of underlying asset classes for our ETF universe, and unfortunately this meant that we are unable to obtain a balanced panel of data; several ETF securities were launched subsequent to our data start date. Additionally there are several days with missing data, due to market holidays and other reasons. We *forward* filled missing data with the last available set of prices, effectively assigning a return of zero to market holidays, for example.

Modeling asset level volatility

To model asset level volatility we use Heterogeneous Exponential Realized Volatility (HEXP) models, as described in Bollerslev et al. (2018). We select these models for several reasons. First, the approach taken by the authors is to fit a 'global' model for the entire panel of

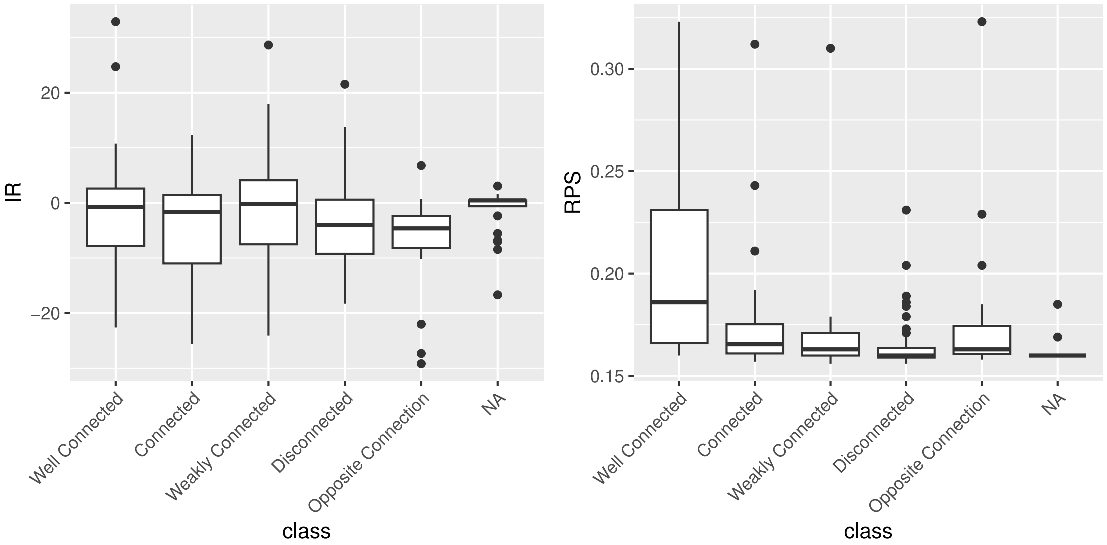


Fig. B.2. Distribution of the IR and RPS scores achieved by the 163 teams included in the global leaderboard, grouped based on the degree the forecasts are connected with the investment decisions.

assets, which is inherently parsimonious and robust, as it uses the same set of parameters for each underlying asset (see above discussion). Such global models reflect a Bayesian philosophy, where the (scaled) time series of each asset is regarded as exchangeable, resulting in significant reductions in estimation error. Empirical evidence summarized in Bollerslev et al. (2018) indicates that these models are appropriate for the wide range of underling asset classes, including those examined here. Second, the model can be easily fitted using OLS, making the approach accessible to competitors without sophisticated econometric knowledge and software. Third, the model offers a simple and intuitive route to account for volatility spillovers and common dynamics in volatility across asset/asset classes. Because the model uses smooth, exponentially weighted averages with several look back periods, it avoids potential variance estimation issues to which its closely related predecessor (i.e., the heterogeneous autoregression (HAR) model) is prone.[11] Of further note is that the HEXP model described in Bollerslev et al. (2018) uses a database of intra-day realized volatility estimates from a cross section of futures contracts. We do not assume that competition participants have access to such data, and instead use daily volatility estimates based on our intra-day high/low/open/close data. Evidence presented in Molnár (2012) suggests that such estimates are reasonably interchangeable with those obtained from intra-day data.

Recall, that competitors are asked to produce portfolios which maximize the Sharpe Ratio over the 20 trading days following each submission date. We therefore build our risk model to make direct single step variance estimates with a 20 day horizon. Separately, we fit a model with a 1 day horizon as a filter prior to constructing our covariance matrix estimates. The HAR structure of our model is readily adaptable to both requirements.

To fit the variance model we begin by calculating historic variances in three ways, adopting the methodology described in Yang and Zhang (2000). We call this the YZ estimator in the sequel. Namely, for each asset define:

$$\{O_t, C_t, H_t, L_t\} = \{Open, Close, High, Low\} \text{ Prices, day t.}$$

$$o_t = log(O_t) - log(C_{t-1}) \qquad u_t = log(H_t) - log(O_t)$$

$$d_t = log(L_t) - log(O_t) \qquad c_t = log(C_t) - np.log(O_t)$$

$$V_P = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{4log2}(u_t - d_t)^2 \qquad V_{RS} = \frac{1}{T}\sum_{t=1}^{T}[u_t(u_t - c_t) + d_t(d_t - c_t)]$$

Here, V_P and V_{RS} are Parkinson (1980) and Rogers and Satchell (1991) variance estimators. When a variance estimate is required for a given day (these values are used as a basis for the exponentially weighted moving average values which are the regressors in our model, and as dependent variables in our 1 day horizon model described below) we use the Garman-Klass estimator:

$$V_{GK} = V_O' - 0.383V_C' + 1.364V_P + 0.019V_{RS},$$

with $V_O' = o_t^2$, $V_C' = c_t^2$, and with V_P and V_{RS} defined as above, setting $T = 1$. When a variance estimate is required for a 20 day period (these form the dependent variables in our models) we use the YZ estimator which accounts for drift in the asset price over the relevant time period. The model also requires a long-run volatility estimate for each asset. For this we utilize the YZ estimator based on an expanding window of data commencing with the first recorded price for each asset. The YZ estimator is

[11] For further discussion of the HAR model, see Doung and Swanson (2015), and the references cited therein. In addition, note that the HEXP model also has some features that draw on the MIDAS model discussed in Ghysels et al. (2007).

defined as[12]:

$$V = V_O + kV_C + (1-k)V_{RS}$$

with,

$$V_O = \frac{1}{T-1}\sum_{t=1}^{T}(o_t - \bar{o})^2 \qquad V_C = \frac{1}{T-1}\sum_{t=1}^{T}(c_t - \bar{c})^2$$

$$\bar{o} = (1/T)\sum_{t=1}^{T} o_t \qquad \bar{c} = (1/T)\sum_{t=1}^{T} c_t$$

$$k = \frac{.34}{1.34 + \frac{(T+1)}{(T-1)}}.$$

At each estimation date, we follow Bollerslev et al. (2018), and form 4 (centered) exponentially weighted averages of daily variance, called $ExpRV_T^K$, with centers of mass (CoM) at 1, 5, 25, and 125 days. Namely, we define:

$$ExpRV_T^K = \sum_{t=1}^{T} w_t RV_t, \quad w_t = \frac{e^{-\lambda t}}{\sum e^{-\lambda t}}, \quad \lambda = CoM[1, 5, 25, 125],$$

with $\lambda = log(1 + 1/CoM)$. We also calculate a universe 'average' volatility factor for our entire investment universe (using assets available at each historic point in time) and re-scale this back to the native scale of each asset as described in Bollerslev et al. (2018). For our h step ahead forecast ($h = [1, 20]$) we then stack all assets and all historic h day time windows in to one vector, and run a single LS regression on the entire data set. We estimate the model as at each competition entry point. The LS model is:

$$\begin{aligned}(\hat{V}_{t:t+h} - RV_t^{LR}) &= \kappa_1(ExpRV_t^1 - RV_t^{LR}) + \kappa_2(ExpRV_t^5 - RV_t^{LR}) \\ &\quad + \kappa_3(ExpRV_t^{25} - RV_t^{LR}) + \kappa_4(ExpRV_t^{125} - RV_t^{LR}) \\ &\quad + \kappa_5(ExpGlRV_t^5 - RV_t^{LR}) + e_t.\end{aligned}$$

Modeling asset level volatility covariance

There are two key components to our estimates of the covariance structure of our asset volatility. First, we assume that M6 asset returns can be described using a factor model. To do this we define factors as linear combinations of standardized returns on M6 assets. Using returns standardized using univariate volatility models is common practice in financial risk management (see Engle (2002)). For example, Alexander (2002) use such returns as the basis for principal component analysis, which is used to estimate factors. We adopt a similar approach, except that we define our factors based on prior knowledge, rather than identifying them from the data. For further details, refer to Connor (2019), who describe several possible approaches to building factor models. These authors note that for 'asset allocation' type analysis, regression models with pre-specified factors tend to be used in practice. Building a bottom up fundamental model for a portfolio of equities would indeed require significant effort, and we assume that the data collection, estimation and testing of such a model was unrealistic for M6 participants. We therefore adopt a purely regression based approach to this problem, solely based on the use of historical returns.

Table C.1
Factor definitions for M6 risk model factors.

Level	Factor	Definition
1	M6M	M6 Market Factor
2	USE	US Equity factor
2	EUE	European equity factor
2	AE	Asian equity factor
2	TERM	Yield curve slope factor
2	CREDIT	Corporate debt - Govt. debt
2	MOM	Momentum factor
2	VAL	Value factor
2	SIZ	Size factor
3	TECH	Technology factor
3	FIN	Financial factor
3	HEALTH	Healthcare factor
3	ENERGY	Energy factor
3	DISC	Consumer discretionary factor
3	IND	Industrial factor
3	COMMS	Communications factor
3	UTIL	Utilities factor
3	STAPLES	Consumer staples factor
3	MATERIAL	Materials factor

More specifically, we define our factors in three hierarchical layers. The layers are fitted sequentially on residuals from the previous layer(s). This enables the use of a very simple identity prior or shrinkage target for our dynamic regression at each stage. For the top level of the hierarchy we use a single factor which we call the 'M6 Market factor' (M6M), as an equally weighted combination of all 50 M6 ETF assets. In some ways this is analogous to a market portfolio for the competition, although obviously it is not market capitalization weighted or appropriate for use outside the context of M6. To construct a market capitalization weighted average we would have needed to either choose some subset of the assets, or obtain incremental data, which we chose not to do. We also decided to exclude individual equity securities, as equities already have significant underlying representation in our ETF securities.

The next layer of the factor model hierarchy is designed to capture the differential dynamics of various asset classes and style factors, which are known from the literature to have significant effects on asset pricing. We again define these using returns on various linear combinations of ETF assets form the M6 universe.

Finally, our level 3 factors are defined to be congruent with the equity market industry ETFs which form a part of the M6 Universe. At this stage, we again adopt a simple identity prior for asset loadings (i.e. we do not attempt to use prior information regarding individual equities to identify them with particular industries, although might be useful). The full set of factors are defined as in Table C.1, with precise weightings available on request from the authors.

The second step of our approach is to estimate time varying loadings for each asset in these pre-specified factors. We do so by using a parsimonious variant of the

[12] Refer to Yang and Zhang (2000) for derivation of V.

BEKK model described above, with an additional shrinkage component, along the lines suggested in Hafner and Reznikova (2012).

More specifically, to estimate asset covariance we proceed as follows. At each submission date we take the in-sample 1 day horizon fitted values from our univariate variance model for each asset. We then standardise the observed asset return for that day using the standard deviations corresponding to these variance estimates. We then use these daily standardized returns to build the hierarchical factor model, using the BEKK approach to estimating the rolling covariance between the M6M factor and each individual asset.[13] As the constant term in the BEKK model, we use the long run whole period covariance between each asset and M6M (estimated via an OLS regression of the standardized asset returns on the factor plus a constant). We specify our BEKK model using two *scalar* parameters, namely γ, the weight on the time $t-1$ error term, and ω, the weight on the long run covariance estimate. The weight on the time $t-1$ covariance estimate is defined as $(1-\omega-\gamma)$. The BEKK model used for this and the following layers of the factor model can be written as:

$$\boldsymbol{\Sigma_t} = \omega \boldsymbol{\Sigma_0} + \gamma \mathbf{e}_{t-1}\mathbf{e}'_{t-1} + (1-\omega-\gamma)\boldsymbol{\Sigma_{t-1}}, \tag{18}$$

where Σ_t denotes our time varying covariance matrix and Σ_0 is our constant long run covariance matrix, the e_{t-1} are previous period errors, and all other terms in the above expression are constants. We then collect the residuals from this exercise, and re-standardize them to have mean zero and standard deviation of 1. We again use a BEKK model (with the same ω and δ) to estimate time varying covariance between all the assets and each factor, with shrinkage towards the identity matrix.

The third step of our approach is to again collect the residuals from the new BEKK model, re-standardize and run the BEKK model once more with the 'level 3' industry factors as regressors.

Finally, to calculate the asset specific risks, we utilize yet another BEKK model, again using the same ω and γ values, with the long run specific variance calculated from the entire sample period. Also, the factor covariance matrix is modeled using a similar structure, again with an identity matrix as the long run shrinkage target and the same BEKK parameters as previously.[14] To choose ω (the weight on the previous day's cross error cross product) and γ (the weight on the long run shrinkage target) we run a grid search, with a test data set comprising the 36 20 day trading periods ending on November 30, 2021. For each date in the period, we estimate our model and compute the log-likelihood of the entire set of M6 asset return observations for following 20 day out of sample period. Prior experience with similar models leads us to test values of ω from the set [.030, .020, .010, .005] and γ from the set [.0025, .0050, .0075]. The values which maximize the log-likelihood for the test data are $\omega = .01$ and $\gamma = .005$. As expected, these values place substantial weight on the shrinkage target. We use these values to estimate the covariance matrix for the rest of the competition.

[13] We note that the M6 assets are traded both in London and in New York, and thus there is the potential for imperfect time synchronicity to affect our results.

[14] At this step, one can introduce a separate set of parameters for the factor covariance matrix, which might be expected to be more stable than the factor/asset loading process. This is left to future research, however.

References

Alexander, C. (2002). Principal component models for generating large garch covariance matrices. *Economic Notes*, *31*, 337–360.

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Ebens, H. (2001). The distribution of realized stock return volatility. *Journal of Financial Economics*, *61*, 43–76.

Armour, B. (2023). Active funds continue to fall short of their passive peers. https://www.morningstar.com/etfs/active-funds-continue-fall-short-their-passive-peers. (Accessed 01 September 2023).

Best, R. (2023). Top 5 positions in warren buffett's portfolio. https://www.investopedia.com/articles/investing/022816/top-5-positions-warren-buffetts-portfolio.asp (Accessed 01 September 2023).

Black, F. (1992). Global portfolio optimization. *Financial Analysts Journal*, *48*, 28–44.

Bojer, C. S., & Meldgaard, J. P. (2021). Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting*, *37*, 587–603.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*, 307–327.

Bollerslev, T. (2009). Glossary to ARCH (GARCH). In T. Bollerslev, J. Russel, & M Watson (Eds.), *Volatility and time series econometrics: Essays in honor of Robert F. Engle*. London: Oxford University Press.

Bollerslev, T., Hood, B., Huss, J., & Pedersen, L. H. (2018). Risk everywhere: Modeling and managing volatility. *The Review of Financial Studies*, *31*, 2729–2773.

Brockwell, P. J. (1991). *Time series: Theory and methods springer series in statistics* (2nd ed.). New York: Springer-Verlag.

Buffett, W. E. (2023). Letter 2022 to Berkshire shareholders.

Chau, M., Lin, C.-Y., & Lin, T.-C. (2020). Wisdom of crowds before the 2007–2009 global financial crisis. *Journal of Financial Stability*, *48*, Article 100741.

Cheng, M., Swanson, N. R., & Yang, X. (2021). Forecasting volatility using double shrinkage methods. *Journal of Empirical Finance*, *62*, 46–61.

Connor, G. (2019). The three types of factor models: A comparison of their explanatory power. *Financial Analysts Journal*, *51*, 42–46.

Dai, M., Jia, Y., & Kou, S. (2021). The wisdom of the crowd and prediction markets. *Journal of Econometrics*, *222*, 561–578.

Doung, D., & Swanson, N. R. (2015). Empirical evidence on the importance of aggregation, asymmetry, and jumps for volatility prediction. *Journal of Econometrics*, *187*, 606–621.

ek, Stanˇ, & F (2024). Designing time-series models with hypernetworks & adversarial portfolios. ArXiv preprint, arXiv:2407.20352.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, *50*(987).

Engle, R. (2002). Dynamic conditional correlation – A simple class of multivariate GARCH models. *Journal of Business & Economic Statistics*, *20*, 339–350.

Engle, R. (2009). *Anticipating correlations: A new paradigm for risk management*. New Jersey: Princeton University Press.

Engle, R. F., & Kroner, K. F. (1995). Multivariate simultaneous generalized ARCH. *Econometric Theory*, *11*, 122–150.

Engle, R., & Mezrich, J. (1996). GARCH for groups. *Risk*, *9*, 36–40.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, *8*, 985–987.

Fama, E. F. (1969). *Papers and proceedings of the twenty-eighth annual meeting of the American Finance Association*. New York.

Frazzini, A., Kabiller, D., & Pedersen, L. H. (2018). Buffett's alpha. *Financial Analysts Journal*, *74*, 35–55.

French, K. R. (2023). Kenneth r. French data library. https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

Garman, M. B., & Klass, M. J. (1980). On the estimation of security price volatilities from historical data. *Journal of Business*, *53*, 67–78.

Ghysels, E., Sinko, A., & Valkanov, R. (2007). MIDAS regressions: Further results and new directions. *Econometric Reviews*, *26*, 53–90.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*, 359–378.

Goldstein, M., & Wooff, D. (2007). *Bayes linear statistics: Theory and methods*. New York: John Wiley & Sons.

Gottschlich, J., & Hinz, O. (2014). A decision support system for stock investment recommendations using collective wisdom. *Decision Support Systems*, *59*, 52–62.

Graham, B. (1949). *The intelligent investor: The definitive book on value investing*. New York: Harper Business.

Grinold, R. C., & Kahn, R. N. (1999). *Active portfolio management (PB)*. New York: McGraw Hill Professional.

Hafner, C. M., & Reznikova, O. (2012). On the estimation of dynamic conditional correlation models. *Computational Statistics & Data Analysis*, *56*, 3533–3545.

Hansen, P. R., & Lunde, A. (2005). A forecast comparison of volatility models: does anything beat a GARCH(1, 1)? *Journal of Applied Econometrics*, *20*, 873–889.

Hospedales, T., Antoniou, A., Micaelli, P., & Storkey, A. (2022). Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*, 5149–5169.

Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, *18*, 439–454.

Jondeau, E., Poon, S.-H., & Rockinger, M. (2007). *Financial modeling under non-Gaussian distributions*. New York: Springer Science & Business Media.

Lassance, N. (2022). Maximizing the out-of-sample Sharpe ratio. Available at SSRN.

Liao, Y., & Fan, J. (2011). High dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics*, *39*, 3320–3356.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge University Press.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, *1*, 111–153.

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., & Simmons, L. F. (1993). The M2-competition: A real-time judgmentally based forecasting study. *International Journal of Forecasting*, *9*, 5–22.

Makridakis, S., Fry, C., Petropoulos, F., & Spiliotis, E. (2021). The future of forecasting competitions: Design attributes and principles. *INFORMS Journal on Data Science*, *1*, 96–113.

Makridakis, S., & Hibon, M. (2000). The M3-Competition: Results, conclusions and implications. *International Journal of Forecasting*, *16*, 451–476.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020a). Predicting/hypothesizing the findings of the M4 Competition. *International Journal of Forecasting*, *36*, 29–36.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020b). The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*, 54–74.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022a). Predicting/hypothesizing the findings of the M5 competition. *International Journal of Forecasting*, *38*, 1337–1345.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022b). The M5 competition: Background, organization, and implementation. *International Journal of Forecasting*, *38*, 1325–1336.

Malkiel, B. G. (1973). *A random walk down wall street: The time-tested strategy for successful investing*. New York: Norton and Company.

Markowitz, H. (1959). *Portfolio selection: Efficient diversification of investments*. New York: Wiley.

Molnár, P. (2012). Properties of range-based volatility estimators. *International Review of Financial Analysis*, *23*, 20–29.

Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., & Talagala, T. S. (2020). FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting*, *36*, 86–92.

Nardari, F., & Scruggs, J. T. (2007). Bayesian analysis of linear factor models with latent factors, multivariate stochastic volatility, and apt pricing restrictions. *The Journal of Financial and Quantitative Analysis*, *42*, 857–891.

Parkinson, M. (1980). The extreme value method for estimating the variance of the rate of return. *Journal of Business*, *53*, 61–65.

Pedersen, L. H. (2015). *Efficiently inefficient: How smart money invests and market prices are determined*. Princeton University Press.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Cyrino Oliveira, F. L., De Baets, S., Dokumentov, A., ... Ziel, F. (2022). Forecasting: Theory and practice. *International Journal of Forecasting*, *38*, 705–871.

Prado, R., Ferreira, M. A. R., & West, M. (2021). *Time series: Modeling, computation, and inference* (2nd ed.). CRC Press.

Rogers, L. C. G., & Satchell, S. E. (1991). Estimating variance from high, low and closing prices. *The Annals of Applied Probability*, *1*, 504–512.

Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, *36*, 75–85.

Spiliotis, E., Assimakopoulos, V., & Makridakis, S. (2020). Generalizing the Theta method for automatic forecasting. *European Journal of Operational Research*, *284*, 550–558.

Storn, R., & Price, K. (1997). Differential evolution – A simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, *11*, 341–359.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor Books.

Swanson, N. R., & Xiong, W. (2018). Big data analytics in economics: What have we learned so far, and where should we go from here? *Canadian Journal of Economics*, *3*, 695–746.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, *16*, 437–450.

Triantafyllopoulos, K. (2021). *Bayesian inference of state space models: Kalman filtering and beyond springer texts in statistics*. Cham: Springer International Publishing AG.

West, M. (1992). Modelling agent forecast distributions. *Journal of the Royal Statistical Society*, *54*, 553–567.

West, M., & Crosse, J. (1992). Modelling probabilistic agent opinion. *Journal of the Royal Statistical Society*, *54*, 285–299.

Yang, D., & Zhang, Q. (2000). Drift-independent volatility estimation based on high, low, open, and close prices. *The Journal of Business*, *73*, 477–492.