The M6 Competition

Hypotheses and key findings

Spyros Makridakis
Evangelos Spiliotis
Ross Hollyman
Fotios Petropoulos
Norman Swanson
Anil Gaba

University of Nicosia
Institute for the Future
National Technical University of Athens
Forecasting & Strategy Unit
University of Bath

43nd International Symposium on Forecasting

The M6 Competition

Innovations

- The first major competition in the area of financial forecasting
- Connecting forecasts with decisions
- Conducted live
- Including multiple evaluation rounds that enhance the objectivity of the results
- Involving real, open-access data and allowing the use of any information or judgment

Objective

Identify practices that would allow investors to **improve the accuracy of their forecasts**, mitigate the uncertainty and bias involved in these forecasts, and exploit their findings to build **robust, profitable portfolios**

The M6 Competition

Setup

- The competition involved 100 assets (50 S&P500 stocks and 50 ETFs), carefully selected to represent all financial sectors

- For each evaluation period (12 in total), participants had to provide forecasts and decisions for all 100 assets. If no new submission was made, the previous one was assumed to continue

- The forecasting horizon was set equal to 4 weeks

- When a new submission was made, the team could specify the data sources and methodological approach used through a questionnaire

- Teams could join the competition at any month. However:
  - Submissions for all months were required to collect the "Global" prizes
  - Submissions for all months of a given quarter were required to collect the prizes of said quarter

Evaluation Measures

Forecasting track

Forecasts refer to the **relative ranks of the assets** in terms of percentage total returns over the evaluation period, divided into quintiles ranking from 1 (worst) to 5 (best)

- The sum of the forecasts for each asset (**probabilities** of being ranked 1 to 5) must sum to 1
- The individual forecasts must be non-negative numbers

Performance was measured based on Ranked Probability Score (RPS):

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

*Score for asset **i** in period **T***

$$RPS_T = 1/100 \sum_{i=1}^{100} RPS_{i,T}$$

*Score for all assets in period **T***

Evaluation Measures

Investing track

Decisions refer to the **proportion (weight)** of capital invested in each asset over the evaluation period

- Positive weights indicate long positions, negative weights short positions, and zero no position
- Exposure must range between 0.25 and 1 (some risk must be taken overall)

Performance was measured based on risk adjusted returns - Information Ratio (IR)

$$IR = \frac{ret}{sdp}$$

__ret__ denotes the continuously compounded portfolio returns and __sdp__ the standard deviation of these returns, measured at a daily frequency

$$ret_t = ln(1 + RET_t)$$

$$RET_t = \sum_{i=1}^{N} w_i \left(\frac{S_{i,t}}{S_{i,t-1}} - 1 \right)$$

Portfolio returns in trading day __t__ are computed based on the weighted average of the price differences of the assets

Evaluation Measures

Duathlon

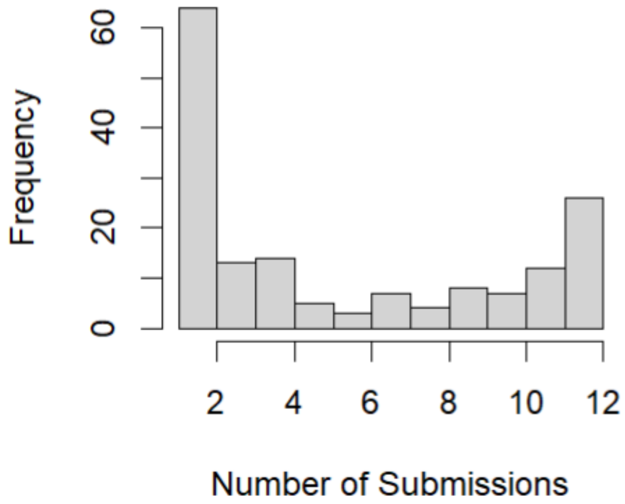
- ✓ **Overall** performance was measured by means of the arithmetic mean of the ranks of the RPS and IR

- ✓ Since the M6 was a duathlon, it assumed equal importance between the two tasks

$$OR = \frac{\mathrm{rank}(RPS) + \mathrm{rank}(IR)}{2}$$

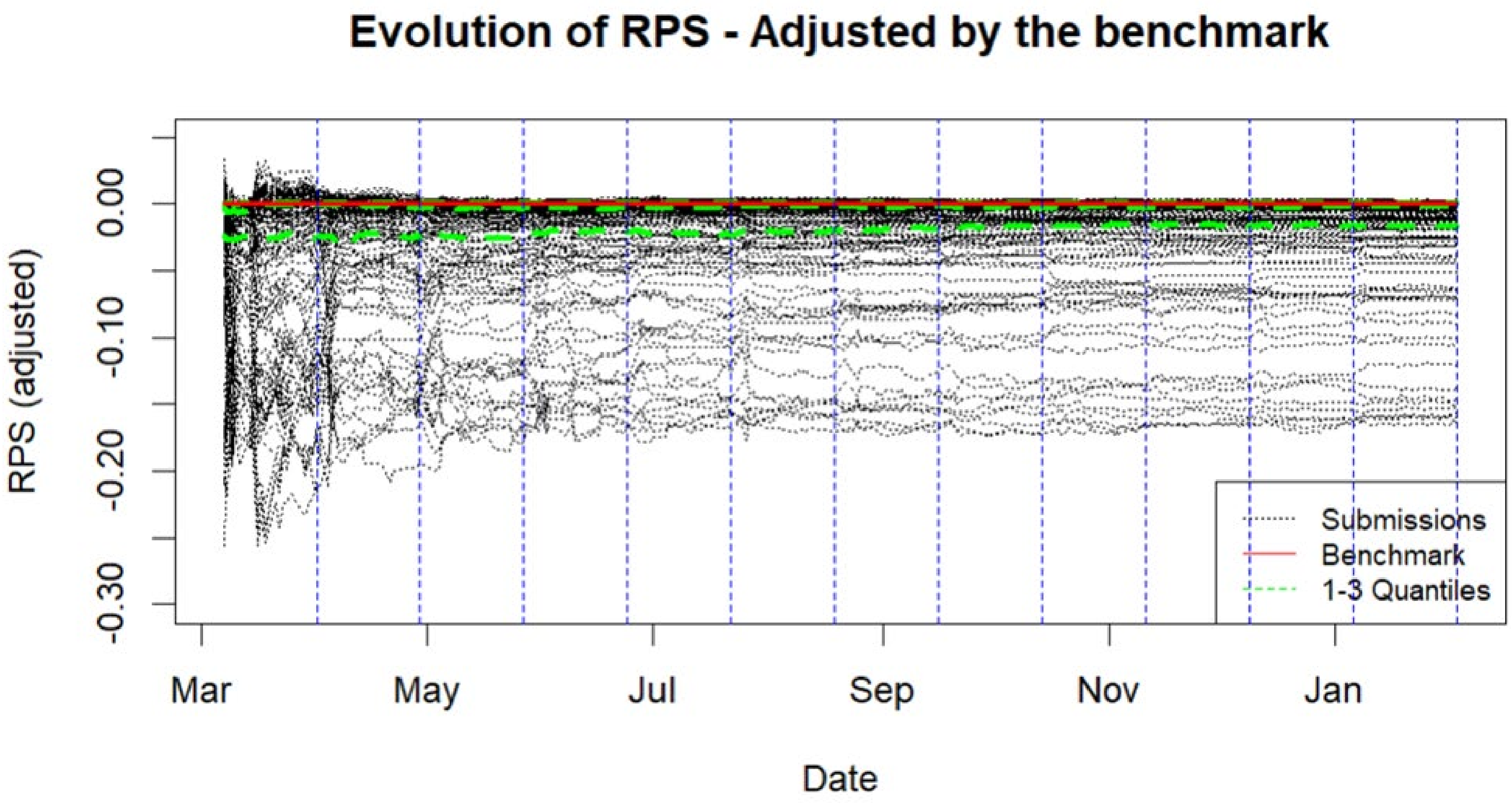
Overview

Participants & Submissions



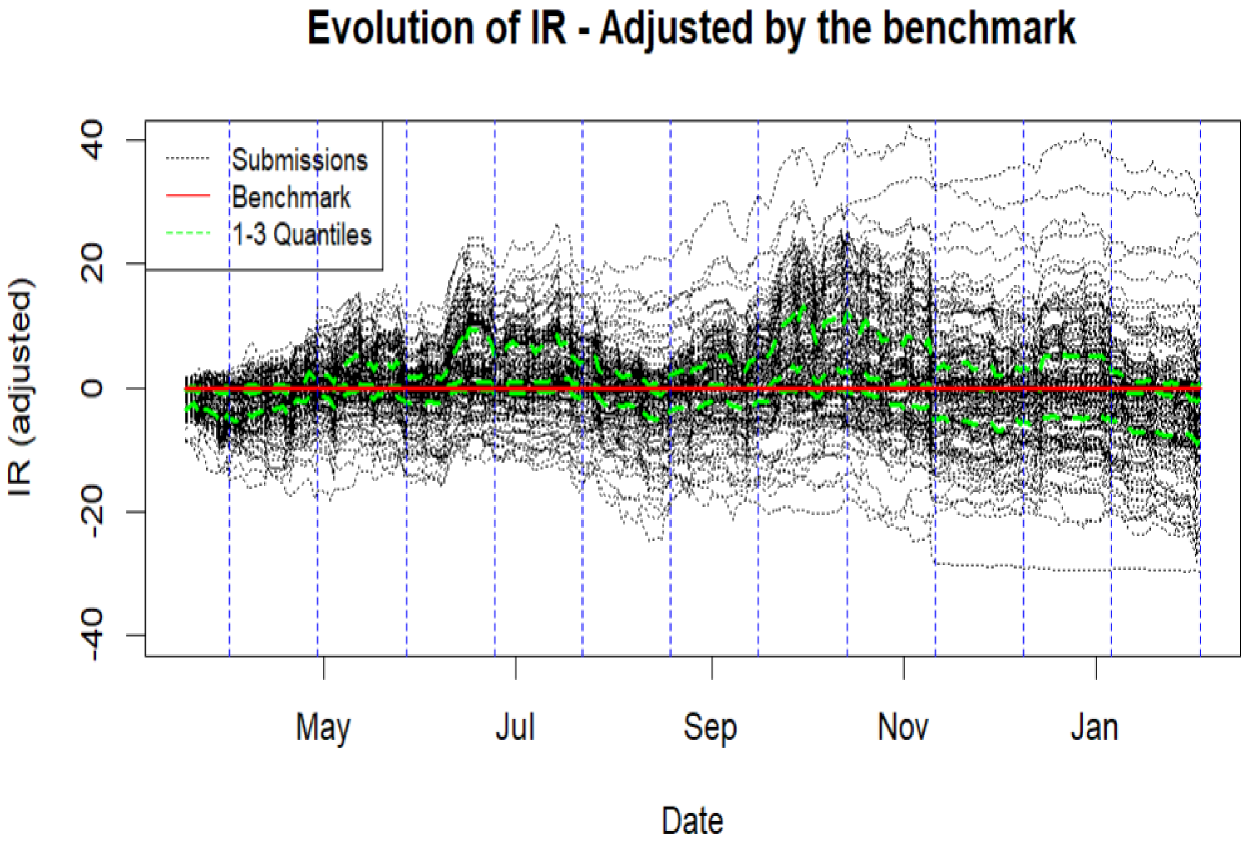
- ✓ 318 participants from 50 countries
- ✓ 226 teams - **163 included in the “Global” leaderboard**
- ✓ On average, the teams included in the “Global” leaderboard made **5 submissions**, mostly within the first 4 months of the competition

Overview



- Globally, 38 teams (**23%**) did better than the "benchmark" (equal probabilities)
- Global improvements reached up to **2.2%**
- Only 13 teams (**8%**) outperformed the "benchmark" in all quarters
- Just 3 teams (**2%**) outperformed the "benchmark" in all months

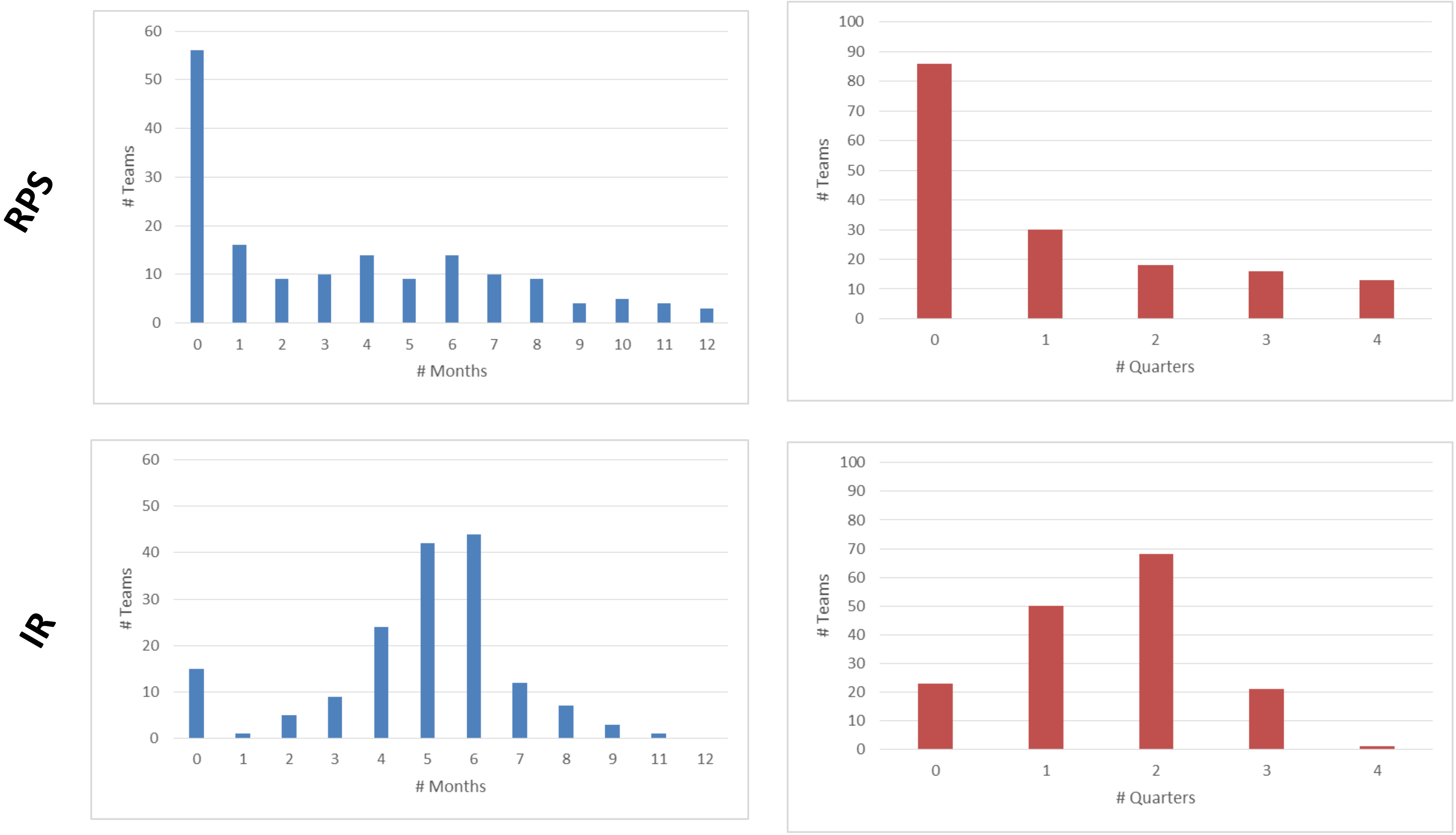
Overview



- Globally, 47 teams (**29%**) did better than the “benchmark” (equal long positions)
- Global improvements were significant, increasing IR by up to **72 times**
- Just 1 team (**0.6%**) outperformed the “benchmark” in all quarters
- No team managed to outperform the “benchmark” in all months

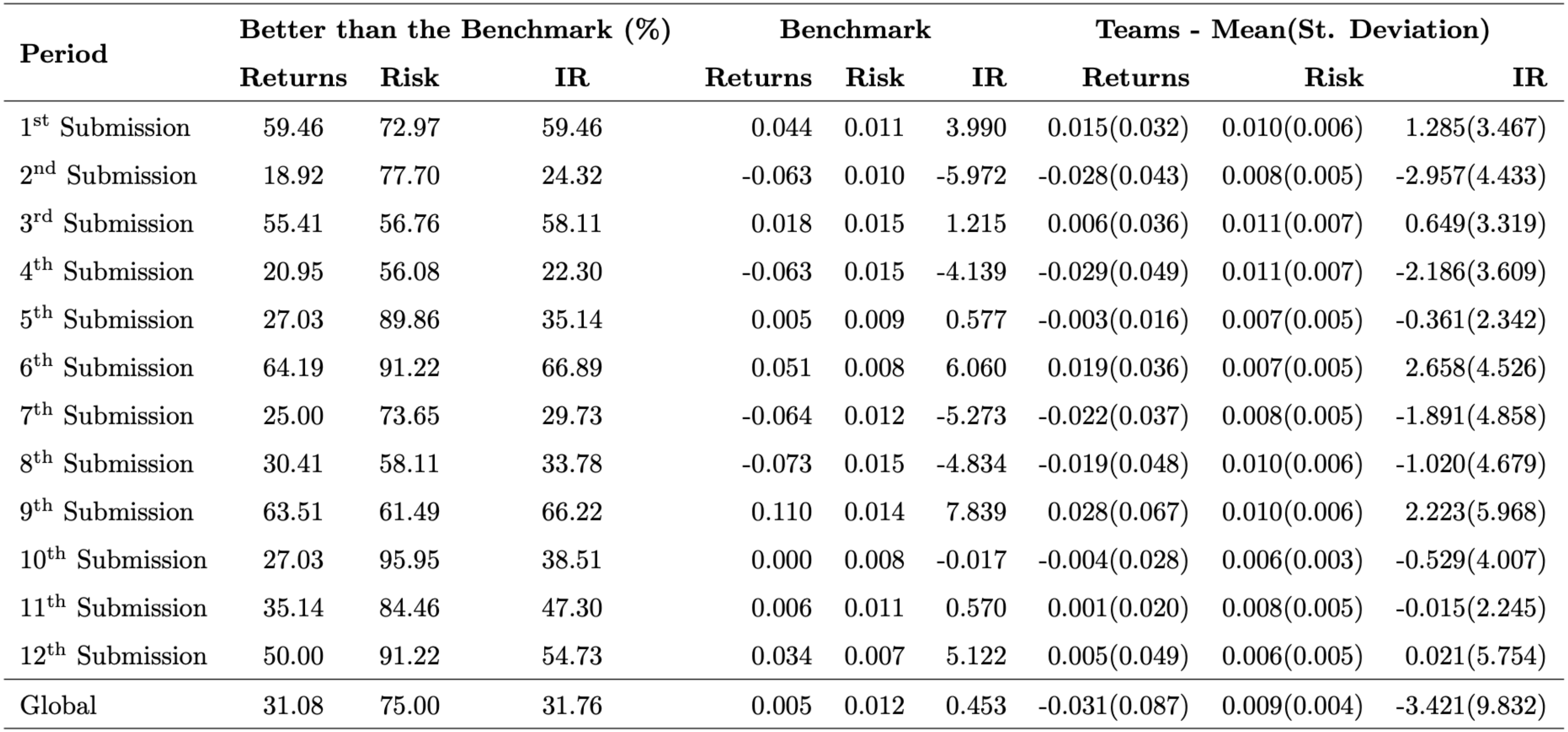
Overview

Number of teams that outperformed the benchmark in *N* months or *M* quarters



Hypotheses Evaluation

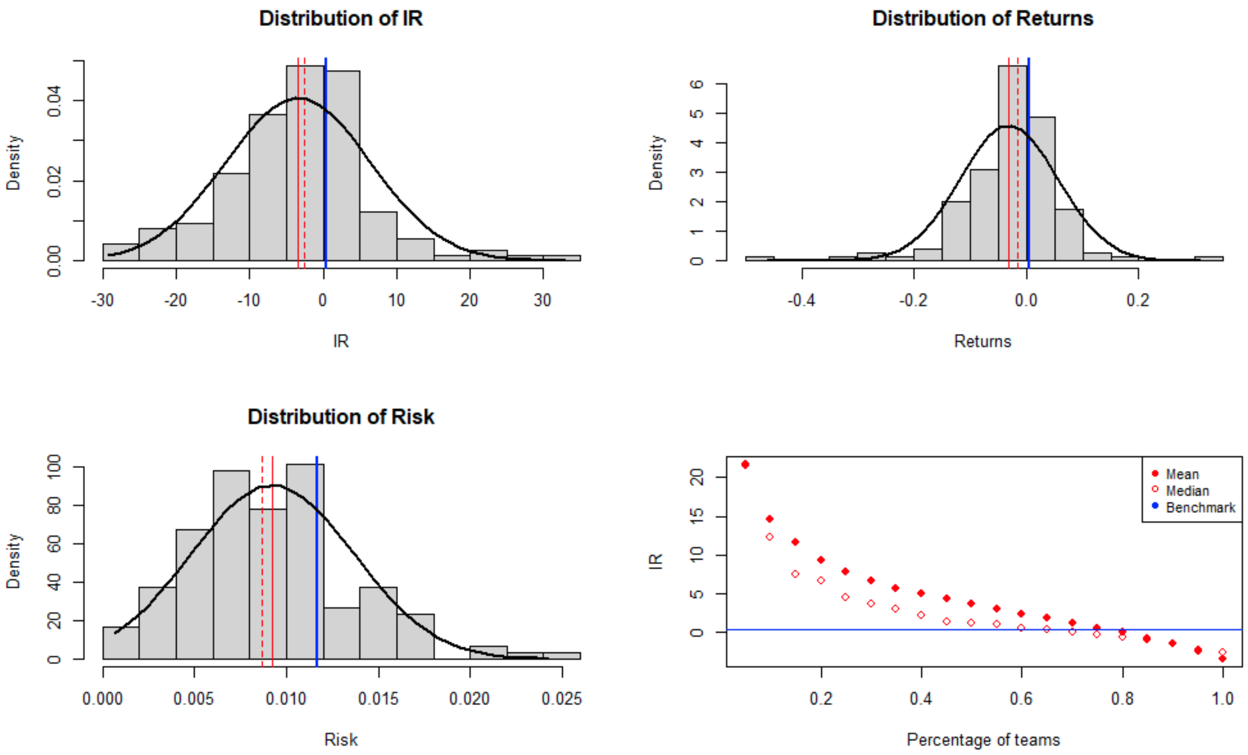
No.1: *The efficient market hypothesis will hold for the great majority of teams but this will not be the case for the top-performing ones.*



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)

Hypotheses Evaluation

No.1: *The efficient market hypothesis will hold for the great majority of teams but this will not be the case for the top-performing ones.*



Results based on the 148 teams included in the "Global" leaderboard whose investment submissions were not identical to the benchmark

ACCEPTED

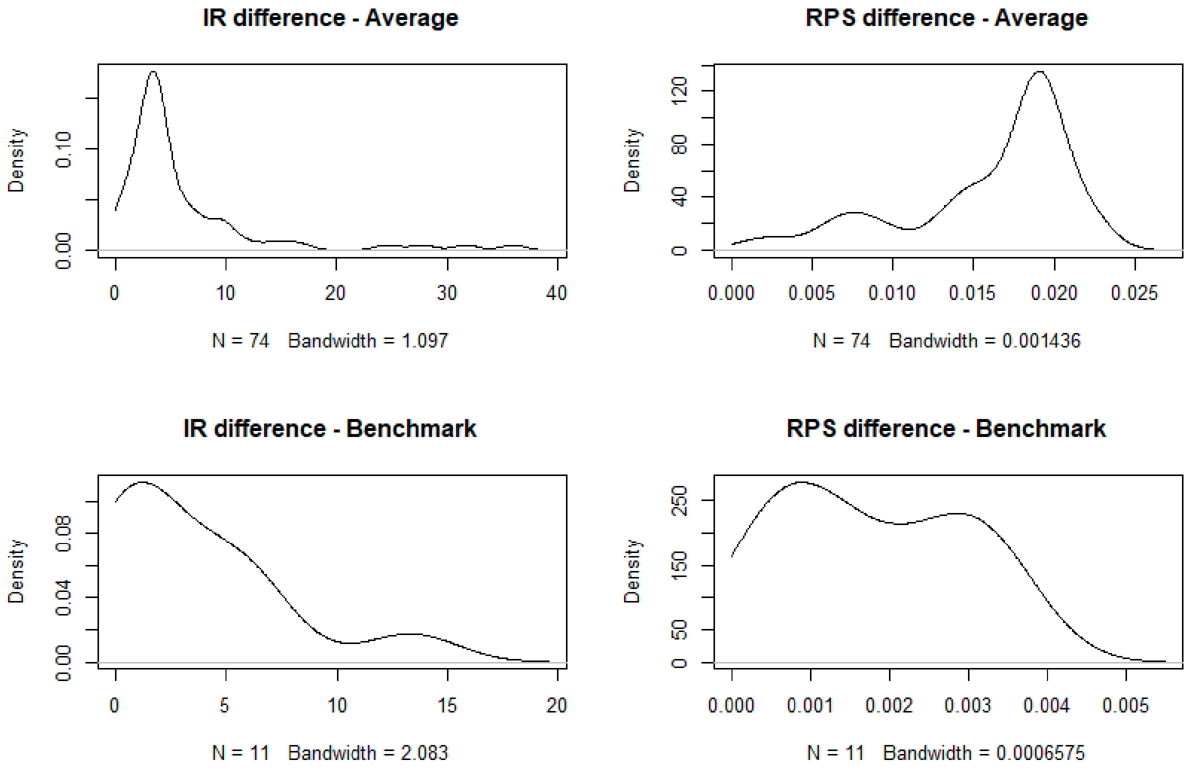
UNIVERSITY of NICOSIA

UNIVERSITY OF BATH

FORECASTING & STRATEGY UNIT

Hypotheses Evaluation

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



Results based on the 162 included in the "Global" leaderboard that outperformed the ***average submission*** *(top) or the* ***benchmark*** *(bottom)*

- Only a limited number of teams did significantly better than the average in both tracks
- Additionally, only 11 teams outperformed the benchmark in both challenges

ACCEPTED

UNIVERSITY of NICOSIA

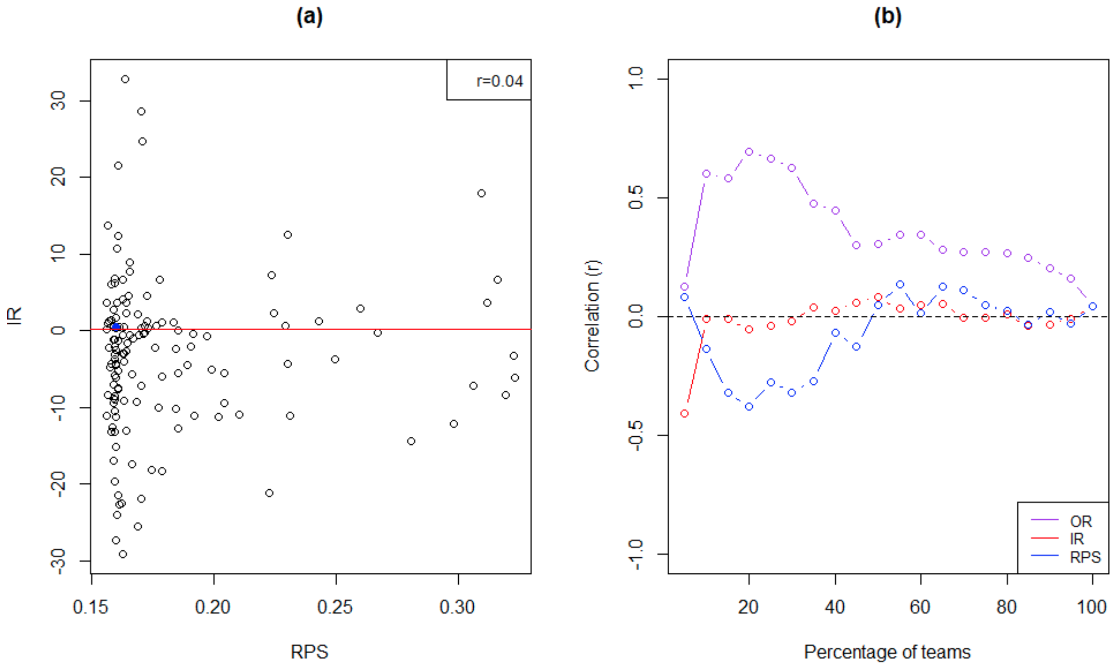
UNIVERSITY OF BATH

43rd International Symposium on Forecasting

FORECASTING & STRATEGY UNIT

Hypotheses Evaluation

No.3 (part 1): *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 magnitude of this link will increase in tandem with team rankings, on average.*

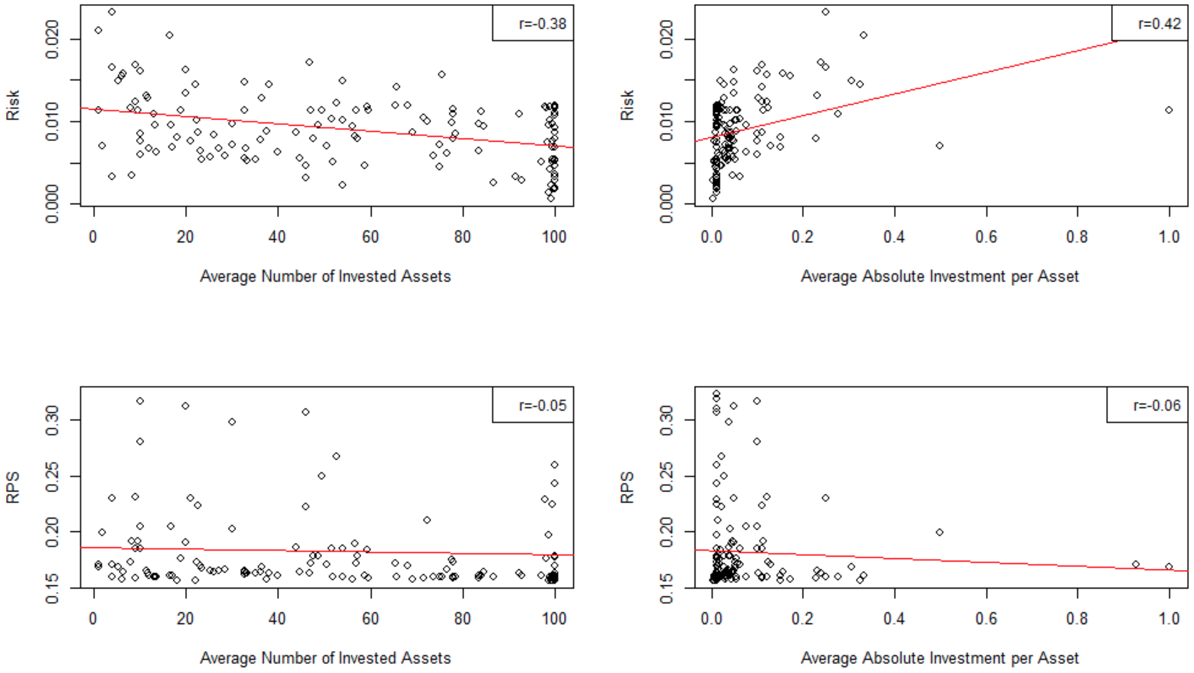


Results based on the 138 teams included in the "Global" leaderboard whose forecast submissions were not identical to the benchmark

- The link is indeed weak on average, but also insignificant for the top-performing teams (or even negative), following the hypothesis when OR is used for ranking the teams
- Many teams focused either on the forecasting or the investment track of the competition, thus rarely performing well in both tracks

Hypotheses Evaluation

No.3 (part 2): *Additionally, team portfolios will in general be more concentrated and risky than can be theoretically justified given the accuracy of their forecasts.*



Results based on the 138 teams included in the "Global" leaderboard whose forecast submissions were not identical to the benchmark

ACCEPTED

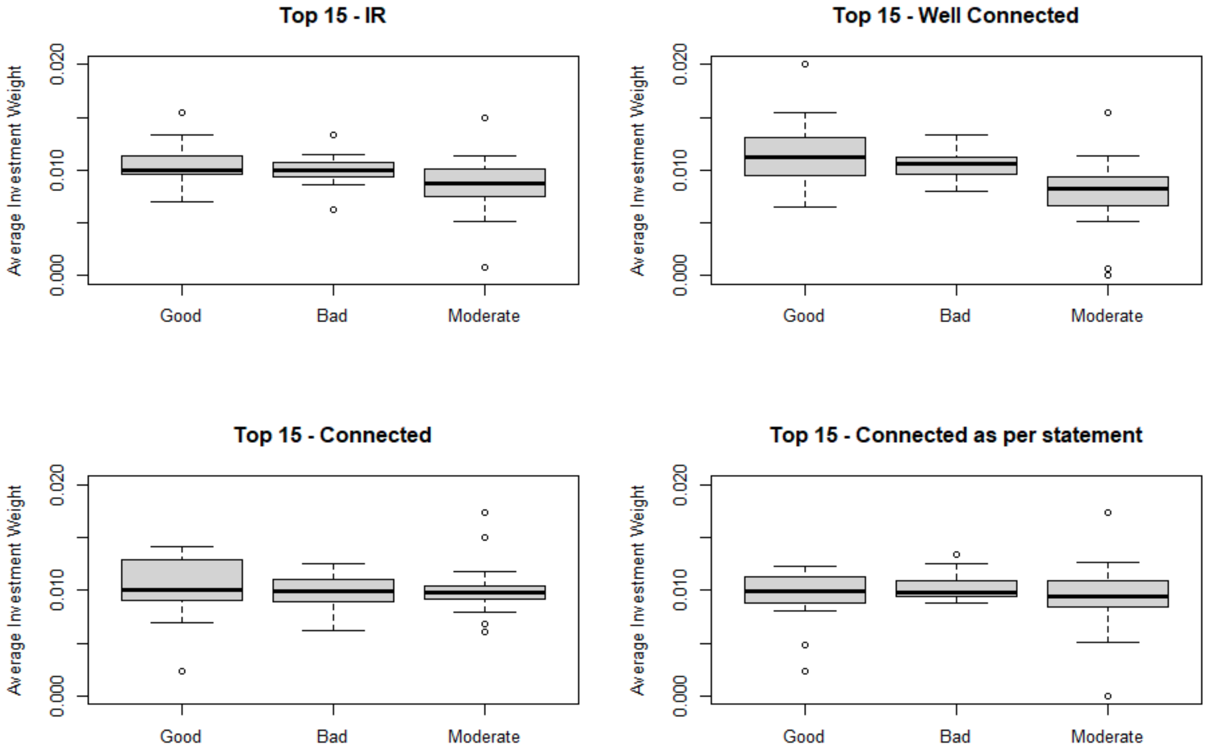
UNIVERSITY of NICOSIA

UNIVERSITY OF BATH

FORECASTING & STRATEGY UNIT

Hypotheses Evaluation

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



Top 15 - IR

Top 15 - Well Connected

Top 15 - Connected

Top 15 - Connected as per statement

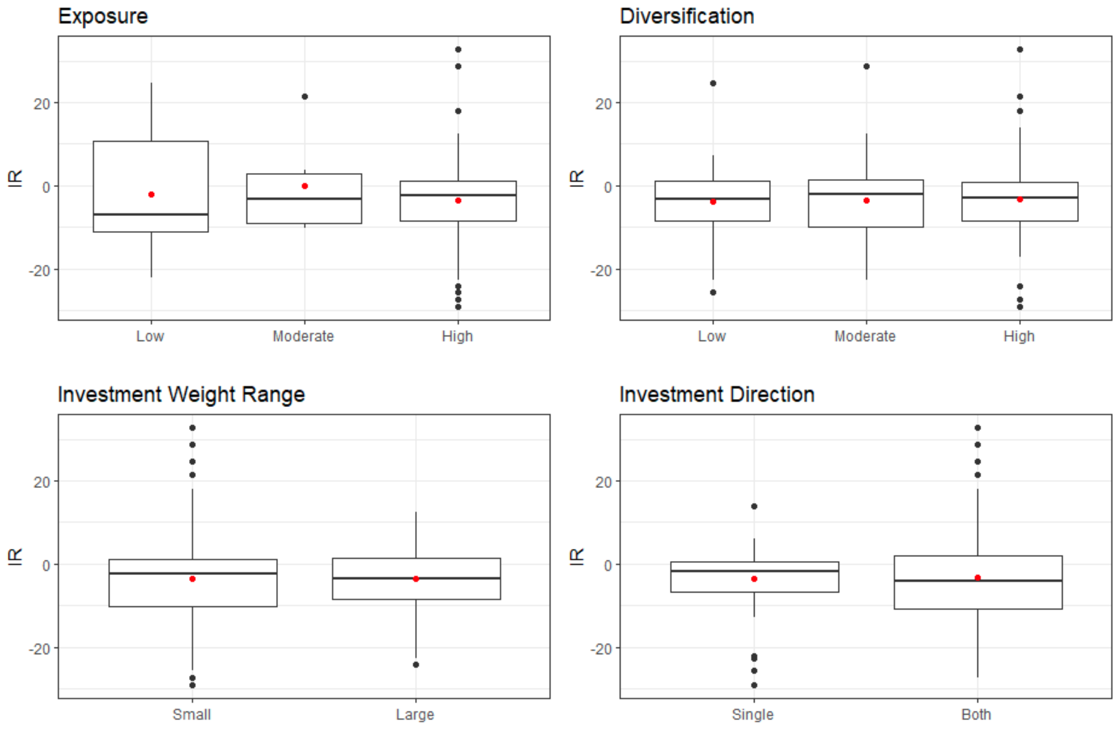
Results based on the 138 teams included in the "Global" leaderboard whose forecast submissions were not identical to the benchmark

RPS	Class
<0.1	Good
0.1-0.2	Moderate
>0.2	Bad

REJECTED

Hypotheses Evaluation

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



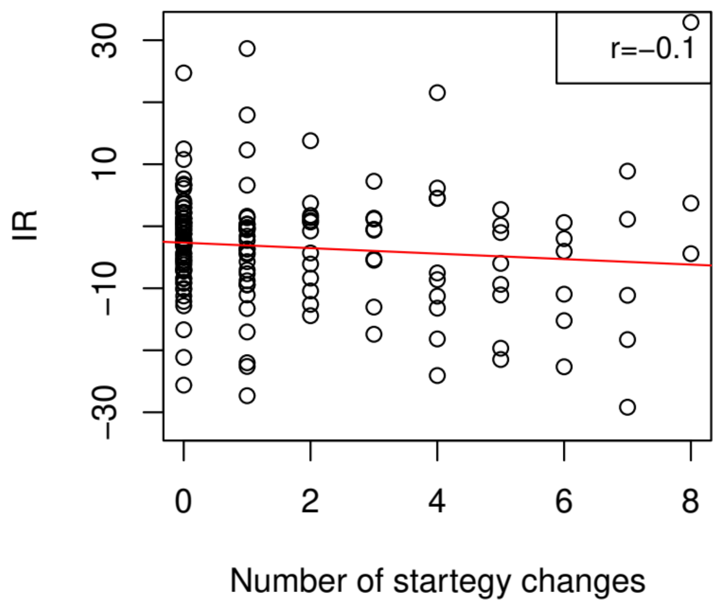
- **Exposure**: Amount of investment
- **Diversification**: Number of assets with an investment
- **Investment Weight Range:** Maximum difference of investment weights across assets
- **Investment Direction:** Going both long/short or not

Results based on the 148 teams included in the "Global" leaderboard whose investment submissions were not identical to the benchmark. Classification based on the average strategy followed.

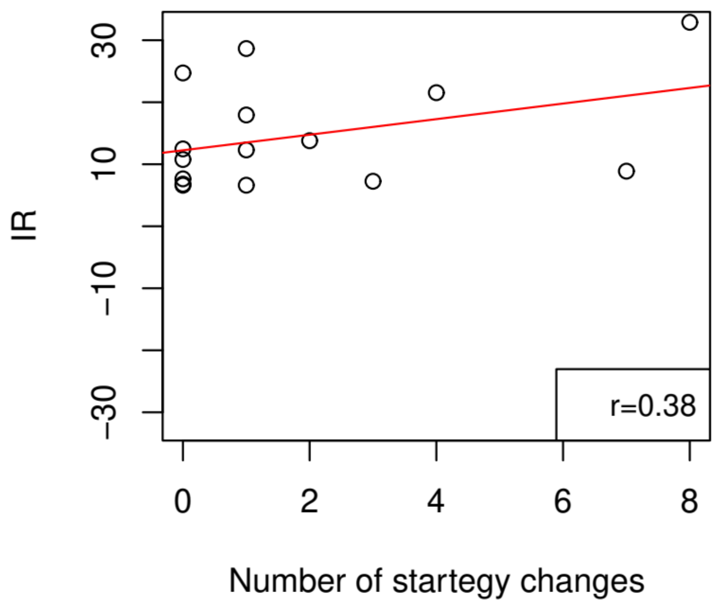
Hypotheses Evaluation

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

All teams

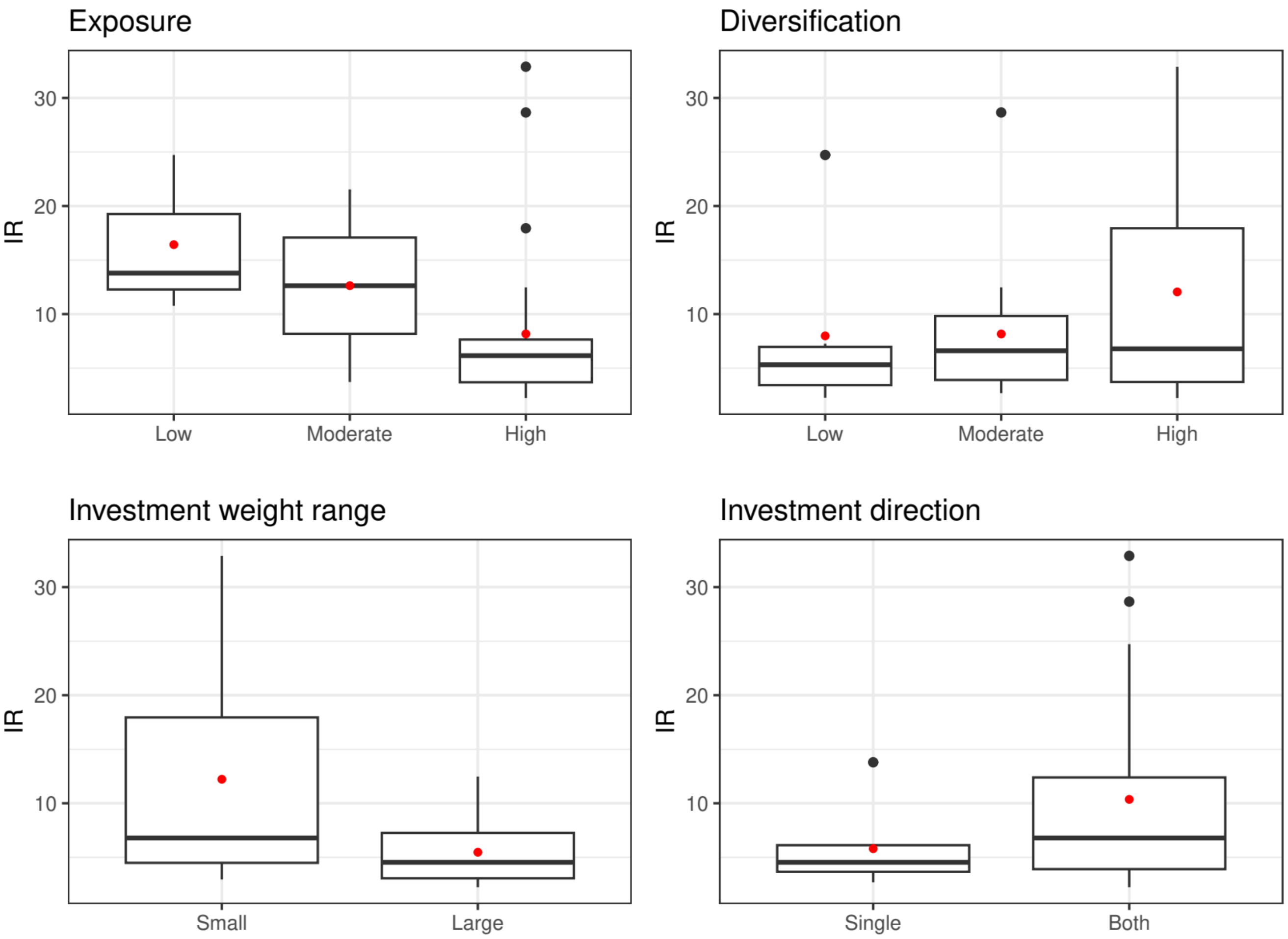


Top 15 teams



REJECTED

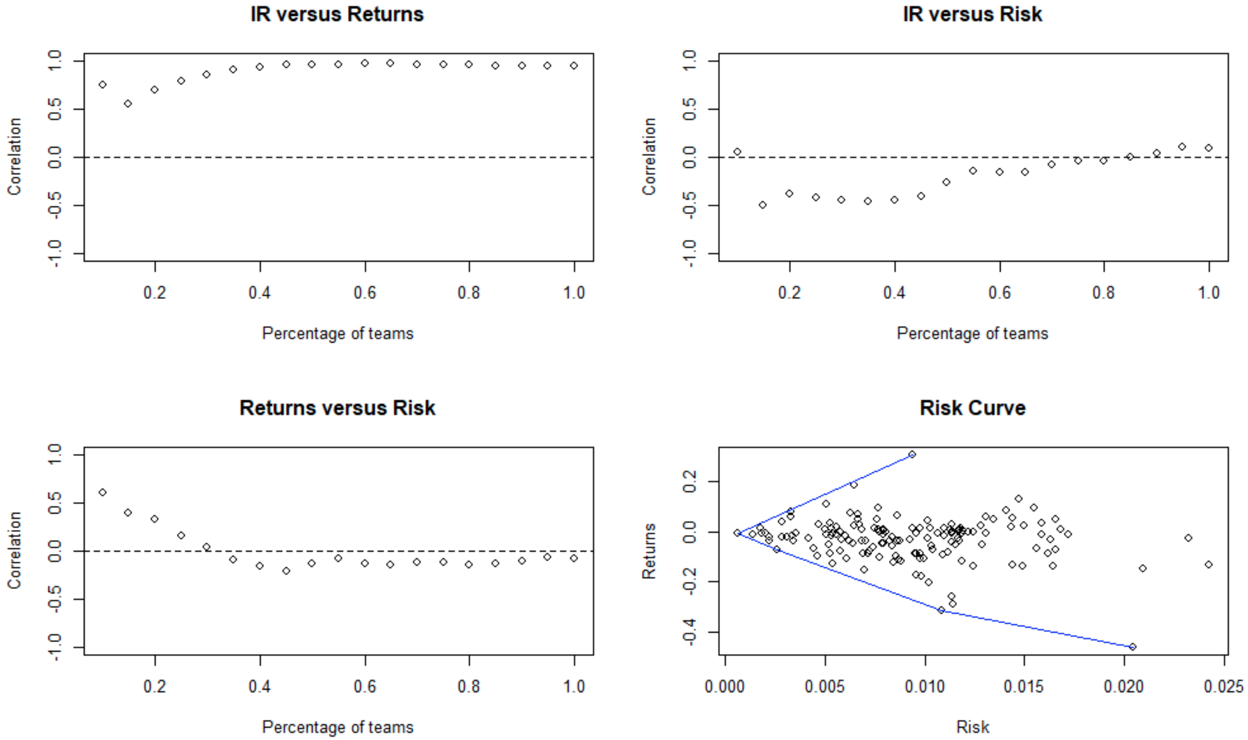
Hypotheses Evaluation



Results based on the top-30 teams

Hypotheses Evaluation

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.*



IR versus Returns

IR versus Risk

Returns versus Risk

Risk Curve

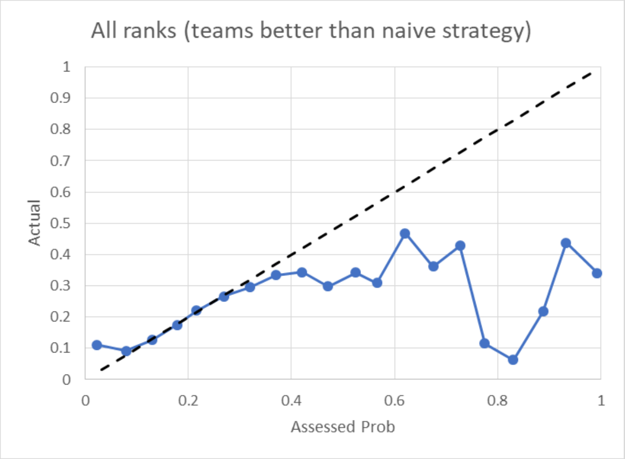
- The correlation between risk and returns is low on average and moderate for the top-performing teams
- There is strong correlation between IR and returns, but it gets weaker for the top-performing teams
- The fact that the risk taken does not always justify the realized returns explains this phenomenon

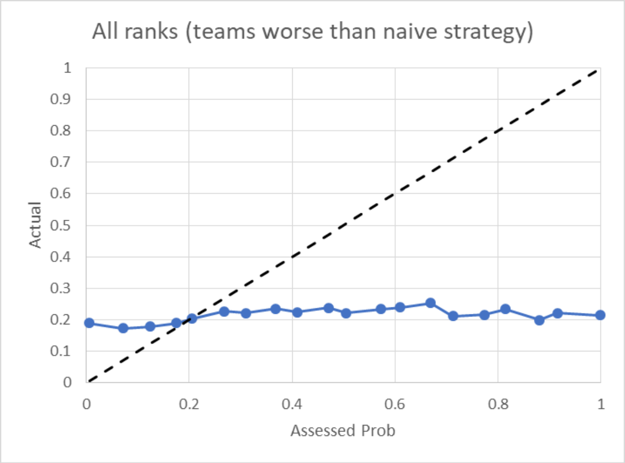
Results based on the 148 teams included in the "Global" leaderboard whose investment submissions were not identical to the benchmark

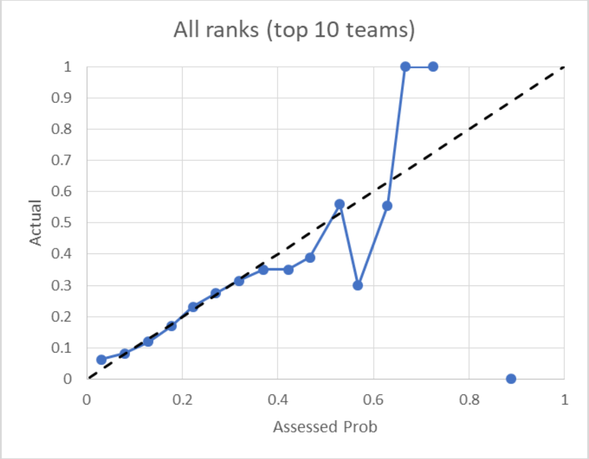
ACCEPTED

Hypotheses Evaluation

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.*





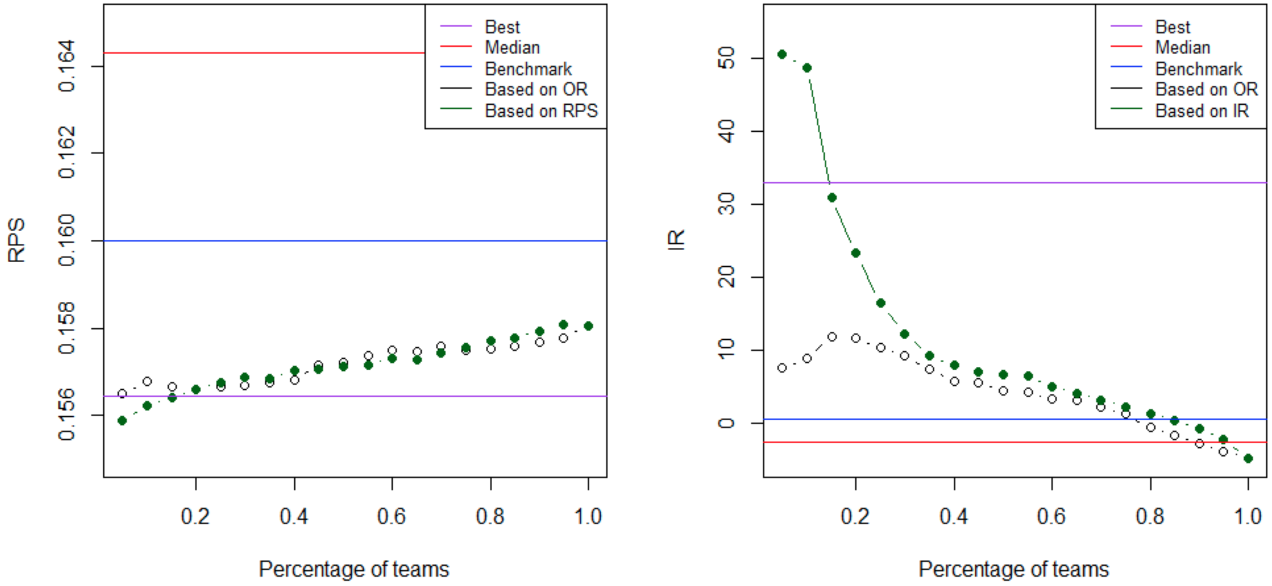


- Overall, the forecasting performance of the teams was low
- There were minor improvements for a few teams and high deteriorations for the majority of the teams
- Most of the teams failed to calibrate their forecasts
- The top performing teams managed to calibrate their forecasts on average but were overconfident is some cases.
- The main reason that the top preforming teams did well was that they avoided submitting large probabilities (>0.7)

ACCEPTED

Hypotheses Evaluation

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.*



- Surprisingly, the hypothesis holds true even when the very worst teams are included in the average

- As expected, the more the average is focused on the top-performing teams, the higher the performance becomes

Results based on the 138 teams included in the "Global" leaderboard whose forecast submissions were not identical to the benchmark

ACCEPTED

Hypotheses Evaluation

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.*

Scores

By forecasting approach	**N**	**%**	**RPS**	**RPS**	**IR**	**IR**
			Mean	**Q90**	**Mean**	**Q90**
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

Ranks

By forecasting approach	**N**	**%**	**RPS**	**RPS**	**IR**	**IR**
			Mean	**Q90**	**Mean**	**Q90**
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

*Time series, ML, and combinations

**Data-driven informed by judgment

- Approaches that were based on pure judgment were significantly inferior to those based on data driven approaches
- There is some merit in introducing judgment to data-driven forecasting approaches
- When judgment is utilized properly, it can offer good performance

ACCEPTED

Hypotheses Evaluation

No.10: *The top-performing teams in the forecasting challenge will employ more sophisticated methods compared to the top-performing teams in the investment challenge.*

RPS	Judgment based	TS* based	ML** based	Not Specified
Top 5%	1	**4**	*3*	1
Top 10%	1	**7**	**7**	2
Top 15%	2	*9*	**10**	4
Top 20%	3	*12*	**14**	5

IR	Judgment based	TS* based	ML ** based	Not Specified
Top 5%	1	**4**	*3*	1
Top 10%	2	**8**	*6*	1
Top 15%	2	**10**	*9*	4
Top 20%	3	**14**	*10*	6

*Including TS combinations

**Including ML integrated with TS and combinations

There is insufficient evidence that teams in the forecasting challenge employed more sophisticated (ML-based) approaches than the top teams in the investment challenge.

CONTROVERSIAL

Major findings

- **Finding 1:** The challenging task of forecasting the relative performance of assets.
- **Finding 2:** The difficulty of consistently outperforming the market.
- **Finding 3:** The limited connection of the submitted forecasts and investment decisions as well as the potential benefits of their association.
- **Finding 4:** The value added by information exchange and the "wisdom of crowds".
- **Finding 5:** The positive effect of adapting to changes.

Platinum Sponsor

Google

Gold Sponsor

Meta

Diamond Sponsor

J.P.Morgan

IIF International Institute of Forecasters

KINAXIS®

Intech®
— JANUS HENDERSON —

causaLens

Erasmus
School of
Economics

sas®

forecast pro®

RUTGERS
School of Arts and Sciences
DEPARTMENT OF ECONOMICS

UNIVERSITY of NICOSIA

FORECASTING & STRATEGY UNIT

INSEAD
The Business School for the World®

MOFC
Learn. Forecast. Compete. Disseminate. Excel.

Thank you for your attention

Questions?

If you would like to learn more about M6 visit

https://mofc.unic.ac.cy/the-m6-competition/