



Munich Personal RePEc Archive

Multiple hypothesis testing of market risk forecasting models

esposito, francesco paolo and cummins, mark



dublin city university, business school

1 March 2015

Online at <https://mpra.ub.uni-muenchen.de/64986/>

MPRA Paper No. 64986, posted 11 Jun 2015 14:00 UTC

Multiple Hypothesis Testing of Market Risk Forecasting Models.

F. P. Esposito * and M. Cummins *

*DCU Business School, Dublin City University

01/03/2015

Abstract

Extending previous risk model backtesting literature, we construct multiple hypothesis testing (MHT) with the stationary bootstrap. We conduct multiple tests which control for the generalized confidence level and employ the bootstrap MHT to design multiple comparison testing. We consider absolute and relative predictive ability to test a range of competing risk models, focusing on Value-at-Risk (VaR) and Expected Shortfall (ExS). In devising the test for the absolute predictive ability, we take the route of recent literature and construct balanced simultaneous confidence sets that control for the generalized family-wise error rate, which is the joint probability of rejecting true hypotheses. We implement a step-down method which increases the power of the MHT in isolating false discoveries. In testing for the ExS model predictive ability, we design a new simple test to draw inference about recursive model forecasting capability. In the second suite of statistical testing, we develop a novel device for measuring the relative predictive ability in the bootstrap MHT framework. The device, we coin multiple comparison mapping, provides a statistically robust instrument designed to answer the question: "which model is the best model?".

Keywords: value-at-risk, expected shortfall, bootstrap multiple hypothesis testing, generalized familywise error rate, multiple comparison map.

1 Introduction

Value-at-Risk, or more simply VaR, has gained popularity among practitioners in the past years because of the increasing exposure to market risk of large financial companies and financial divisions of non-financial firms, and mostly because of the ability of this metric to deliver a readable quantity concerning overall risk borne. This popularity has increased as a result of the many “crises” and large corporate defaults due to market exposure, which have become more frequent since the early 90’s and largely publicised by the media.

VaR¹ is used by risk managers in banks and wealth management companies to monitor the market risk of large and varied portfolios of financial securities and over-the-counter products in order to trigger action by the management on the back of the information packed into this number, summarizing the optimistic loss in a worst case scenario, with a given probability on a certain time horizon. This calculation has also become part of regulatory requirements, e.g., in banking regulations such as in Europe, whereby it is used to determine the amount of regulatory and economic capital. From an operational point of view, however, VaR in general lacks the important property of subadditivity, Artzner et al. (1999). Practically, this means that the VaR of a weighted sum of individual quantities is not equal to the weighted sum of each VaR, hence requiring multiple layers of calculation when aggregating from subsets to the consolidated portfolio level. This feature led to a shift of focus towards the alternative risk measure of expected-shortfall (ExS), see Artzner et al. (1999), which holds the subadditivity property and provides further information: namely, the expected loss in a worst case scenario, with a given probability on a certain time horizon. This measure represents also a complementary indicator that accounts for the magnitude of losses exceeding the VaR threshold and draws attention to the full shape of the tail event distribution.

VaR and ExS are not accounting quantities that come out of a simple algebraic operation. If we were to give some formalism, these measures and in general a risk measure is a functional defined on a space of random variables, mapping the set of events or “scenarios” of concern $\mathcal{A} := (\Omega, \mathcal{F})$, that is a set of events and a complete algebra defined on its terms. The space \mathcal{A} is assumed to possess a probability measure \mathbf{P}^* , that is a measure of the uncertainty of the events belonging to \mathcal{F} , which is unknown. A model is any candidate $\mathbf{P} \sim \mathbf{P}^*$, that is a measure of uncertainty equivalent, in a certain sense, to \mathbf{P}^* , representing an approximation of the true distribution. Essentially, the estimate of a risk measure requires the construction of a probabilistic model for the distribution of values and the establishment of robust procedures to infer those numbers from historical samples.

As a logical consequence, the statistical testing of risk models is an important step towards assessing the ability of these tools to provide reliable output and to contribute to the decision-making process hinged on market risk exposure. From a practitioner perspective, there are serious implications for a financial institution from its choice of risk model in terms of its overall risk management performance and more importantly its capital adequacy requirements. So for industry, the question of which risk model performs best in capturing and forecasting risk exposure is crucial. Historically the first contributions in backtesting the performance of risk-models are those of Kupiec (1995) and Christoffersen (1998), who construct unconditional and conditional tests based on the mere sequence of VaR breaches. Thereafter, research focused on specific issues affecting the VaR prediction ability, such as the time-horizon of the forecast, the inclusion of time-varying volatility and accounting for fat-tailed distributions generated by volatility clustering and jumps in returns, see BIS (2011) for a review. On the other hand, backtesting ExS is more problematic due to the peculiarity of its functional form which in principle requires the estimation of the entire tail distribution. The literature on model prediction of ExS is not as extensive as that on VaR, possibly due to the latter reason. The main contributions in this field are Berkowitz (2001) and Kerkhof and Melenberg (2003), who use the probability transform Rosenblatt (1952) to process the data and construct tests of ExS prediction based, respectively on the likelihood ratio and the δ -functional method Van der Vaart (1998). Both these works focus on the development of a test for ExS.

All of this literature ignores a fundamental issue with the multiple testing of competing models. This issue is the multiple comparisons problem that is inherent in multiple hypothesis testing. The multiple comparisons problem is well established in the statistical and econometrics literature but is largely ignored in the empirical finance literature. The problem arises when performing multiple hypothesis tests simultaneously and leads to the non-negligible likelihood of identifying statistically significant results by pure chance alone, rather than on the basis of true statistical relationships. Without controlling for the multiple comparisons problem, the probability of rejecting true hypotheses, i.e. making erroneous false discoveries, is increased. Romano, Shaikh and Wolf (2010) provide a detailed exposition of the issues

pertaining to multiple hypothesis testing, outlining the main literature in the area. A key contribution of this paper is the application of generalised multiple hypothesis testing procedures to control for the multiple comparison problem in backtesting VaR and ExS models. To date, the authors are only aware of the paper by Bao et al. (2006) who explicitly account for this multiple comparisons bias with the application of the bootstrap reality check of White (2000). However, the use of the bootstrap reality check of White (2000) has some limitations, in particular it is highly conservative in that it seeks to control the probability of making even one false discovery and so lacks power, where power is loosely defined as the ability to reject false null hypotheses, i.e. to make true discoveries. Moreover, the relative comparison test is limited to the benchmarking of the model suite to the forecasting performance of RiskMetrics. We overcome these limitations by controlling for the generalised family-wise error rate and further introducing a new statistic for the relative model performance comparison.

In this article, we build on the work of Bao et al. (2006) in that we use a similar model set and investigate model predictive ability not only for the VaR but also for the ExS models with respect to different time horizons and volatility conditions. We work within the framework developed by Beran, Beran (1988), Politis and Romano (1994b), Romano and Wolf (2005), Romano and Wolf (2007), Romano and Wolf (2010), in that we utilize *generalised bootstrap multiple hypothesis testing*, or MHT for short. This approach offers a robust technique for easily implementing any kind of statistical test, which for our purposes only requires weak stationarity. Sitting on top of a powerful simulation engine, the generalised bootstrap MHT is capable of delivering hypothesis testing that is free from analytic or asymptotic pivotal results and that can incorporate finite sample effects, model misspecification and parameter estimation error. The main contribution of this article lies in the application of the bootstrap MHT framework to the backtesting of market risk measures. We test absolute and relative forecasting performance of market risk models. Within the latter exercise, we develop a novel device to support the cross-comparison of relative predictive performance, a device we coin *multiple comparison map*, or MCM for short.

We construct bootstrap MHT of risk model predictive ability, analyzing the out-of-sample performance over 1-day and 10-day time horizons. We extend the investigation to forecasts that target a time horizon wider than a single day, an exercise that might either confirm the predictive power of a model or highlight situations whereby the forecast deteriorates fast. We also observe the model performance under stressed market scenarios. The inference procedure is accomplished via a direct measure of the VaR predictive ability or rather exploiting the idea first popularised by Diebold et al. (1998) and used by Berkowitz (2001), Kerkhof and Melenberg (2003), Bao et al. (2007), in that we use the probability transform Rosenblatt (1952) to construct statistics which are functionals of the model probability distribution and thereby indirectly test the data via the probability transformed sample. In the latter case we provide a new simple test for the backtesting of the model predictive ability of the expected-shortfall.

The battery of tests draws inference about two aspects of the model forecasting ability, that is each individual model's capability to provide significant predictions, namely the absolute model performance, and secondly the quality of each model in relation to the rest of the competing models, namely the relative model performance. In the absolute performance exercise, we estimate confidence sets for the target statistic and derive joint balanced tests which control for the probability of committing Type I error. This concept is extended further with the introduction of the generalised family-wise error rate (k -FWER), cfr. Romano et al. (2009) for a review. The k -FWER sets a tolerance trigger for false rejections, which on one hand increases the power² of the MHT while tightening the confidence bands, while on the other hand provides a mechanism to assess hypotheses not excessively distant from acceptance. The MHT is further refined by the introduction of a step-down algorithm, a procedure which involves recursive testing that potentially allows for further rejections by altering the critical values at each stage depending on the hypotheses already rejected up to that point, see Romano and Wolf (2010). In the analysis of the relative forecasting ability, we develop a new approach expanding the base bootstrap MHT structure of the test and using the multivariate test distribution generated by the bootstrapping algorithm, which embeds the overall test dependencies, and we produce a thorough comparison of each model with respect to all its competitors, measuring the pair-wise probability that each model is better than any other in the collection. By the suitable extraction of relevant information which is summarised in one simple table, we are able to evaluate the forecasting ability of each model in relation to the performance of the remaining competitors, providing a valuable tool to answer robustly the question concerning the best model in the set of competitors.

The work is organized as follows. In Section 2 we outline the computational engine, the bootstrap, while in Section 3 we present the framework we use to construct the balanced confidence sets approach and the step-down algorithm. In Section 4 we briefly introduce the conditional distribution models that

form the suite of competing market risk forecasting instruments, whereas in Section 5 we define the target risk measures and the sample statistics, further describing the structure of the testing exercise and introducing the MCM. The experimental Section 6 describes the data set, the modelling approach and discusses the empirical evidence exhibited. Section 7 concludes.

2 The Stationary Bootstrap

The bootstrap, Efron (1979), is a versatile method for investigating a general form of functions depending on the full sample history. In the original form of this procedure, we search for an estimate of the statistic $R(\mathbf{X}; \mathbf{P})$ with $\mathbf{X} = \{X_i\}_{i=1, \dots, n}$ and $X_i \stackrel{\text{iid}}{\sim} \mathbf{P}$. The bootstrap method allows one to construct an estimate of the statistic distribution using the sample distribution $\hat{\mathbf{P}}$

$$R^* = R(\mathbf{X}^*; \hat{\mathbf{P}}),$$

which consists of repeatedly drawing with replacement observations $X \in \mathbf{X}$, each weighted with probability $1/n$. The distribution estimate of R is generated through the re-sampling \mathbf{X}^* , (m resamplings of \mathbf{X}). This procedure is valid under the i.i.d. hypothesis for X . A further generalization is achieved, for example, with the methods in Küsch (1989), Liu and Singh (1992) or in Politis and Romano (1994b), whereby the bootstrap delivers robust estimates of the distribution of the root³, R , for stationary and weakly dependent time series. In this work, we adopt the *stationary bootstrap* of Politis and Romano (1994b).

The stationary bootstrap algorithm starts by “wrapping” the data in circle, such that $Y_t = X_{\tilde{t}}, \forall t \in \mathbb{N}$, with $\tilde{t} := (t \bmod n)$ and the convention that $X_0 := X_n$. A pseudo-time series \mathbf{X}^* is produced retaining the stationary properties of the original data sample \mathbf{X} . The re-sampling scheme requires the construction of blocks $B_{i,l} = \{Y_i, Y_{i+1}, \dots, Y_{i+l-1}\}$, generated through the withdrawal of i.i.d. discrete uniform random numbers $I_1, \dots, I_s \in \{1, \dots, n\}$ and geometric random block lengths L_1, \dots, L_s , with distribution function $\mathbf{D}\{L_i = k\} = p(1-p)^{(1-k)}$, $k \in \mathbb{N}$. The generic re-sampled time series is $\mathbf{X}^* := \{B_{I_1, L_1}, \dots, B_{I_s, L_s}\}$.

Although optimally choosing the expected block length $1/p$ does not affect the consistency properties of the bootstrap, the optimal p grants the fastest convergence rate of the estimates and therefore their minimum variability, cfr. Politis and White (2004). In the sequel, the artificial samples are simulated with preconditioning on the optimal p , see Politis and White (2004), Patton et al. (2009). In terms of bias and variability of the variance of the pseudo-time series, the stationary bootstrap of Politis and Romano (1994b) is equivalent to other techniques for bootstrapping stationary and weakly dependent sample data, though not originally noticed in the multiple comparison work of Lahiri (1999), but successively corrected by Nordman (2009).

The most attractive characteristic of the bootstrap approach is its high degree of flexibility; it can be used with parametric and non-parametric models, non-pivotal statistics, that is lacking asymptotic distribution results, and mostly it can capture features of finite sample statistics whose distributions might be sensibly different from asymptotic pivotal results, see Horowitz (2000) for a review on the topic. These features are very appealing in the present context where the ultimate purpose of this work is to examine the predictive ability of different classes of possibly misspecified models, carrying additional model estimation error. It is relatively simple within the bootstrap approach, to design experiments for model selection based on the performance of models with different statistical properties and targeting risk measures that might have unknown finite sample or even unknown asymptotic properties. Ultimately, we construct tests for statistics that are functional of some estimate of the conditional distribution of the random variable modeling the sample observations, cfr. Politis and Romano (1994b), Politis and Romano (1994a).

3 Simultaneous Confidence Sets and the Step-Down Algorithm

In this section we present two subsections describing, respectively, the construction of the balanced confidence set controlling the generalised probability of Type I error across the family of testing hypothesis, k -FWER, and the step-down procedure. Essentially, the step-down procedure consists of an iterative procedure which aims to minimize Type II error probability and hence optimise the statistical power of the MHT.

3.1 Balanced Confidence Sets

Consider a statistical model \mathbf{P}_j of the observations $\mathbf{X} \sim \mathbf{P}$, and a risk measure $\rho(\mathbf{P})$, in general a functional of the probability measure \mathbf{P} . The MHT problem consists of testing the m hypothesis

$$\mathcal{H}_j: |\rho_j(\mathbf{P}_j) - \rho(\mathbf{P})| < 0, \quad j = 1, \dots, m. \quad (1)$$

Considering joint statistical testing is very important when the dependency across the individual tests is high. In order to understand intuitively this issue, we borrow an example from Romano et al. (2009). Consider 100 independent statistical tests each of them with a confidence level of $\alpha = 0.05$; the probability of rejecting at least one of these tests is extremely high, that is $1 - 0.95^{100} = 0.994$. Hence, the probability of committing an error of first type is very high, which calls for a procedure capable of controlling the probability of *false rejections* in the presence of a dependent multiple test structure.

In order to construct MHT (1), we exploit the duality between statistical tests and confidence sets therefore proceeding to the estimation of probability intervals for the statistics $\rho_j \in \mathbf{T}_j$, where \mathbf{T}_j is the domain of ρ_j , and testing that the critical value ρ belongs to the specific band. Hence, following Beran (1988), let the roots $R_{n,j}(\mathbf{X}, \rho_j)$ be the relevant function of the data, the confidence set for the statistic j is the set

$$C_{n,j} = \{\rho_j \in \mathbf{T}_j: R_{n,j}(\mathbf{X}, \rho_j) \leq c_{n,j}(\alpha, \mathbf{P})\}. \quad (2)$$

Notice that here for ease of presentation we refer to one sided intervals, whereas in the experimental section we effectively work with two sided intervals that can be achieved in a straightforward extension of this procedure.

Furthermore, we require that the joint confidence set $\mathbf{C} := \{C_{n,j}\}$ has *coverage probability* $1 - \alpha$ and has to be *balanced*, that is the confidence level for the interval $C_{n,j}$ remains the same $\forall j$. The first constraint forces the MHT to put in place a mechanism for controlling the joint probability of committing at least one error of first type, the so called *family-wise error rate* (FWER). The second constraint, balancing, is a very important property of the test: if lacking balance then the joint test would determine tighter confidence bands for worse models and wider intervals for better ones. The aforementioned procedure is achieved through *pre-pivoting*. In fact, indicating with $H_{n,j}(\cdot)$ the cumulative distribution function of $R_{n,j}$ and with $H_n(\cdot)$ the left continuous distribution of $\max\{H_{n,j}, \forall j\}$, the right boundary of the confidence set $C_{n,j}$ is then

$$c_{n,j} = H_{n,j}^{-1} [H_n^{-1}(1 - \alpha)]. \quad (3)$$

In case we wish to construct a double sided confidence set with joint probability $1 - \bar{\alpha}$, we simply define $\alpha = \bar{\alpha}/2$ and compute (3) as the right boundaries and, to determine the left hedge, we consider the left sided version of $c_{n,j}$, this time defining $H_n := \min\{H_{n,j}, \forall j\}$. The solution of Beran (1988) to (3) is the “plug-in” estimate

$$\hat{c}_{n,j}(\alpha, \hat{\mathbf{P}}) = \hat{H}_{n,j}^{-1} [\hat{H}_n^{-1}(1 - \alpha)], \quad (4)$$

calculated with bootstrapping, see Beran (1988), Beran (1990), Beran (2003).

The extension to the procedure of Beran consists of targeting the k -FWER instead of the mere FWER. The generalized FWER expands the capability of targeting multiple false discoveries and at the same time provides a control variable that can be used in multiple runs of the testing procedure. By adjusting the k -FWER, it is possible to detect weak departures from the null hypothesis, in the case whereby a certain hypothesis is accepted at a slightly lower k . Formally, we define the

$$k\text{-FWER} := \mathbf{P}\{\text{reject at least } k \text{ true models}\}. \quad (5)$$

Setting $k\text{-FWER} = \alpha$ means controlling for the joint probability of *at least* k false discoveries, thereby introducing a target probability of committing joint errors of Type I. The construction of multiple hypothesis tests controlling the k -FWER has followed different procedures in Lehmann and Romano (2005) and Romano and Wolf (2007) while in Romano and Wolf (2010) the authors achieve the generalization of (2), thereby introducing balancing in the MHT procedure while controlling the confidence level of at least k false discoveries. The construction of balanced right sided confidence sets with $k\text{-FWER} = \alpha$ is achieved by setting

$$H_n := k\text{-max}\{H_{n,j}, \forall j\},$$

with $k\text{-max}\{y_1 < y_2 < \dots < y_s\} := y_{s-k+1}$, $k \leq s$.

The procedure just introduced provides a double benefit to multiple testing: firstly, the extension to controlling the k -FWER of balanced multiple hypothesis tests raises the tolerance to false rejections, therefore it makes the acceptance threshold more strict and increases the significance of the null hypothesis that are accepted; secondly, the parameter k draws attention on individual test statistics that are not excessively far away from the null but close to the rejection region that with small variation of the k -FWER may be discarded or not.

3.2 The Step-Down Algorithm

Given the set-up of the balanced MHT with control of the k -FWER, it is possible to improve the performance of the test by means of adopting a *step-down method*. Step-down methods implicitly estimate the dependency structure of the individual tests achieving an improvement in the power of the MHT. The algorithm (Romano and Wolf, 2005, 2007, 2010) can be described as follows.

Let MHT be the set of m (right hand sided) simultaneous hypotheses

$$\mathcal{H}_j: R_{n,j} \leq c_{n,K,j}(\alpha, k) \quad (6)$$

where we now make explicit the set of indexes $K = \{1, \dots, m\}$ and the dependency of c on k . The sets A_s and B_s are, respectively, the sets of accepted and rejected hypotheses at the step s . At the start of the procedure, set $A_0 \equiv K$ and the counter $s := 0$

ALGORITHM A: Generic step-down method for control of the k -FWER

- If $R_{n,j} \leq \hat{c}_{n,A_0,j}(\alpha, k)$, $\forall j \in A_0$, then accept all the hypothesis and stop; otherwise, reject any \mathcal{H}_j for which $R_{n,j} > \hat{c}_{n,A_0,j}(\alpha, k)$ and include j in B_1 ; set $A_1 := A_0 \setminus B_1$ and increase the step s by 1;
- **while** $|B_s| \geq k$
 reject any \mathcal{H}_j for which $R_{n,j} > \hat{d}_{n,A_s,j}(\alpha, k)$ and include j in B_{s+1} , where

$$\hat{d}_{n,A_s,j}(\alpha, k) := \max_{\substack{I \subseteq B_s \\ |I|=k-1}} \{\hat{c}_{n,D,j}(\alpha, k): D = A_s \cup I\};$$

set $A_{s+1} := A_s \setminus B_{s+1}$ and increase the step s by 1;

end

The algorithm **A** is capable of increasing the statistical power, that is the probability of rejecting a false null hypothesis, because at each iteration the subset of the lowest p -value statistics is excluded, tightening confidence bands in the subsequent iteration and hence strengthening the ability to pick true discoveries. However, accounting for *at least* $k > 1$ false discoveries involves the possibility that at the previous stage we have rejected true hypothesis, but hopefully *at most* $k - 1$. As a consequence, at step s we have to consider within the current MHT the event of having previously dismissed $k - 1$ true nulls, a fact that would affect the current critical values. Nevertheless, iterating through the set B_s to include the event “rejection of $k - 1$ true nulls” might turn out to be a formidable task due to a rapidly growing number of possible combinations of size $k - 1$ from the previously rejected hypotheses. For this reason, Romano and Wolf (2007), Romano and Wolf (2010) propose a streamlined algorithm, which simplifies the computational burden of algorithm **A**.

ALGORITHM B: Streamlined step-down method for control of the k -FWER

- If $R_{n,j} \leq \hat{c}_{n,A_0,j}(\alpha, k)$, $\forall j \in A_0$, then accept all the hypothesis and stop; otherwise, reject any \mathcal{H}_j for which $R_{n,j} > \hat{c}_{n,A_0,j}(\alpha, k)$ and include j in B_1 ; set $A_1 := A_0 \setminus B_1$ and increase the step s by 1;
- **while** $|B_s| \geq k$
for each $j \in B_s$ calculate the p -value $\hat{p}_{n,j} = 1 - H_{n,j}$ and sort them in descending order $\hat{p}_{n,r_1} \geq \dots \geq \hat{p}_{n,r_{|B_s|}}$, where $\{r_1, r_2, \dots, r_{|B_s|}\}$ is the appropriate permutation of the p -value indices that gives this ordering; then pick a user specified integer $N_{\max} \leq \binom{|B_s|}{k-1}$ and let M be the largest integer such that $\binom{M}{k-1} \leq N_{\max}$;
reject any \mathcal{H}_j for which $R_{n,j} > \tilde{d}_{n,A_s,j}(\alpha, k)$ and include j in B_{s+1} , where

$$\tilde{d}_{n,A_s,j}(\alpha, k) := \max_{\substack{|I|=k-1 \\ I \subseteq \{r_1, r_2, \dots, r_M\}}} \{\hat{c}_{n,D,j}(\alpha, k) : D = A_s \cup I\}$$

set $A_{s+1} := A_s \setminus B_{s+1}$ and increase the step s by 1;
end

The rationale of algorithm **B** is to reduce the computational burden due to the number of combinations generated by calculating critical values $\hat{d}_{n,A_s,j}$ by limiting the pool of rejected hypotheses to those that are least significant. The streamlined step-down method tries to reduce the computational effort, limiting the set to be explored to the hypotheses that are most likely to be rejected. As a consequence, the algorithm is as close as possible to the generic algorithm **A**. The step-down algorithm defines a search path to strengthen the power of the MHT, driven by the implicit dependency structure of the individual test. At each iteration the algorithms **A** and **B** minimise the Type **II** error probability, hence improving the statistical power. Notice that in the case of two sided confidence sets the previous algorithms have to be modified accounting for left critical values computed as minima across the search set and furthermore including left p -values in the operational method. In the empirical section we use the bootstrap MHT augmented with the step-down algorithm **B**.

4 Conditional Distribution Models

In this section we introduce the models that are included in the collection of market risk forecasting tools and whose performance form the objective of the MHT experiment in the empirical section. The output of the model we are interested in is the conditional probability density forecast delivered by the different techniques. Although it is sufficient modelling just the tail of \mathbf{P} to produce the inference that is sought, in certain cases we will need the full distribution to project the system forward.

The suite of models includes: heuristic models such as the historical simulation (HS), which is a rolling window histogram, a rolling window Normal model (G) and RiskMetrics (RM)⁴; a non-parametric model based on a kernel regression (KR); parametric models such as the autoregressive conditional heteroskedastic model (CH), the quantile auto-regression model (QR) and several parametric distribution assumptions such as normality, student-t, generalized error distribution (GED) and the generalized Pareto distribution (GPD).

Historical Simulation

The historical simulation (HS) model consists of a rolling window histogram of the return distribution. The implicit assumption is that a t -left neighborhood data sample histogram is a good local estimate of the conditional distribution \mathbf{P}_t . Although this might be acceptable as an estimate of $\mathbf{P}_t(X_{t+\varepsilon})$, $\varepsilon > 0$, a model-free approach seems inadequate when $\varepsilon \gg 0$. Assuming i.i.d., we can compute the distribution Δ lags forward with numerical convolution or MC integration. This model is the most widespread in the financial community, because of the ease of implementation and mainly because it allows one to aggregate easily the many varieties of financial exposures which would otherwise require the design of an all-inclusive market risk model.

Normal Hypothesis

The classical assumption of the Black-Scholes model is that log-returns are normal. A practical approach to the estimation of a conditional mean-variance model is the plugging in of a rolling window sample mean and variance into the normal function to construct a model of $\mathbf{P}_t(X_{t+\varepsilon})$. This model should be able to capture some momentum and volatility clustering.

RiskMetrics

The RM model, J.P.Morgan (1996), consists of an exponential smoothing of the squared returns, which is used for a $t + 1$ variance proxy. Formally,

$$h_t = \theta \cdot h_{t-1} + (1 - \theta) \cdot x_{t-1}^2.$$

It was originally designed as a simple alternative to the GARCH model on the observation that the lag polynomial is often close to the stability condition and the GARCH parameters for financial time series are not widely different across a large collection of data. It presents the disadvantage that it cannot be projected forward. As a working template, we use the normal hypothesis to compute projections of the conditional probability distribution, assuming innovation of the variance proxy at current h .

Kernel Regression

A robust and efficient (but biased) technique to estimate a conditional distribution is to exploit the kernel regression of Nadaraya-Watson, cfr. Nadaraya (1964), Watson (1964) and Bierens (1987) for several statistical results. Formally,

$$\widehat{\mathbf{E}}\{y|x\} = \frac{\sum_j y_j K_h(x - x_j)}{\sum_j K_h(x - x_j)}$$

A drawback of this estimator is that it shows high variability when conditioning on values of $x_{t-\Delta}$ that are far away from the center of gravity of the sample distribution, therefore producing instable estimates of the tails. On the other hand, an attractive feature of the KR estimator is that it can generate directly an estimate of the distribution conditional on any lag; in this case the projection exercise is a direct output of the estimation function.

Quantile Regression, CAViaR

The quantile regression model, Koenker and Bassett (1978), is a statistical model of empirical percentile. Basically, it is a parametric model of relations between the explanatory variables and the percentile of the target variable. In this work, we employ the specification in Engle and Manganelli (2004), which accounts for autoregressive features of the model quantile. The model has been designed pretty much for the estimation of an auto-regressive VaR, therefore the epithet of conditional autoregressive VaR, designated as CAViaR. Formally, letting $X_{t-1} = \{y_{t-1}, \dots, y_{t-i}, \dots\}$, a quantile auto-regression model is defined as

$$\begin{aligned} y_t &= f(X_{t-1}; \beta) + \varepsilon_t^\theta \\ &= f_t(\beta) + \varepsilon_t^\theta, \end{aligned}$$

with the auxiliary assumption that the θ^{th} -quantile of the ε_t^θ distribution is equal to 0. The model estimation is carried out with the minimization of the loss function

$$\min_{\beta} \frac{1}{T} \sum_t (\theta - \mathbf{1}_{\{y_t < f_t(\beta)\}}) (y_t - f_t(\beta))$$

which is minimal whenever $f_t \equiv \theta$.

In this work we employ four different CAViaR specifications: the adaptive; the symmetric; the asymmetric; and the indirect GARCH. The latter three models are specified as in Engle and Manganelli (2004),

whereas the adaptive CAViaR is defined as

$$f_t = f_{t-1} + [\mathbf{1}_{\{y_{t-1} < f_{t-1}\}} \cdot b_1 + \mathbf{1}_{\{y_{t-1} \geq f_{t-1}\}} \cdot b_2] \cdot (y_{t-1} - f_{t-1}).$$

For each model, we estimate a quantile regression for the 1st-5th percentiles, in addition to the 7.5% and 10% levels, in order to smooth out the borders of the distribution. Hence, we use those percentiles as a point-wise tail estimate. The extreme value distribution is assumed to have a linear to higher order polynomial decay, matching the all-time minimum, with polynomial degree ranging from 1 to 20. In this exercise we are modeling the tail of the conditional distribution function only. We project the distribution forward, simply multiplying the knot points, that is the estimated percentiles, by the square root of time.

GARCH-EVD Models

Financial time series exhibit volatility clustering features and fat tailed distributions. The generalized autoregressive conditional heteroskedastic models, Engle (1982), Bollerslev (1986), represent the most successful statistical device in mimicking the evolution of financial time series in the past thirty years. The GARCH models come in a variety of fashions. However, a GARCH(1,1) does not seem an unreasonable assumption for financial time series, cfr. Hansen and Lunde (2001). In the MHT experiment we account for modeling time series volatility clustering with conditional heteroskedastic models, and incorporate asymmetry with exponential or threshold GARCH, cfr. Nelson (1991) and Glosten et al. (1993). Specifically, we estimate symmetric GARCH models, Engle (1982), Bollerslev (1986)

$$h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1}$$

and models capable of producing asymmetric distributions such as the TARCH(1,1) of Glosten et al. (1993)

$$h_t = \alpha_0 + (\alpha_1 + \gamma \mathbf{1}_{\{\varepsilon_{t-1} < 0\}}) \varepsilon_{t-1}^2 + \beta_1 h_{t-1}$$

and the EGARCH(1,1) of Nelson (1991)

$$\log h_t = \alpha_0 + \alpha_1 (|v_{t-1}| - \mathbf{E}|v_{t-1}|) + \beta_1 \log h_{t-1},$$

with $\varepsilon_t = v_t \sqrt{h_t}$.

The stochastic driver v_t is such that $\mathbf{E}v_t = 0$, $\mathbf{E}v_t^2 = 1$ and $v_t \sim G(\theta)$, where $G(\theta)$ is a parametric distribution of type normal or student-t. However, it is common knowledge that financial time series exhibit fat tailed distributions, Longin (1996). Further improvement is achieved by augmenting the model with a piecewise distribution for $G(\theta)$. Thereby, we adapt extreme-value distributions to each tail of the residuals with a quasi-maximum likelihood estimation (QMLE) of the model classes introduced above, while a conditional normal or GED is estimated for the mid percentiles. In order to parametrize a GPD for each tail, we exploit the idea in McNeil and Frey (2000) and maximize the Kolmogorov-Smirnov statistic of the empirical distribution of the exceedances. This approach is different from that taken by Bao et al. (2006), where the threshold is picked at a conventional level. A robust alternative semi-parametric estimation technique has been proposed in Mancini and Trojani (2011). The described econometric setup is able to capture the time dependency described by volatility clustering, the asymmetric effect and the thick tails phenomenon.

5 Backtesting and Multiple Hypothesis Testing

In this study we employ the generalised bootstrap MHT to design a suite of statistical tests to investigate the forecasting performance of the collection of market risk models presented above. We build upon a simulation approach, the stationary bootstrap Politis and Romano (1994b), to generate the statistic distributions. In this Section we present the sample measures and the devised tests that are employed in the experimental exercise to evaluate and compare the forecasting ability of the suite of models under analysis.

The target risk measures are the VaR and the ExS. Formally,

$$\rho = \text{VaR}_t(\Delta, \alpha) := \inf \{x \in \mathbb{R} : \mathbf{P}_t(X_{t+\Delta} \leq x) \geq \alpha\}. \quad (7)$$

Having defined (7), we introduce some simplifying notations. Let $\mathbf{P}_t(\cdot) = \mathbf{P}(\cdot | \mathcal{F}_t)$, $\mathbf{P}_t(\cdot | D) = \mathbf{P}(\cdot | D \cap \mathcal{F}_t)$,

$E = \{X_{t+\Delta} < \text{VaR}_t(\Delta, \alpha)\}$ and $\bar{E} = \{X_{t+\Delta} \leq \text{VaR}_t(\Delta, \alpha)\}$. The ExS can then be defined as

$$\rho = \text{ExS}_t(\Delta, \alpha) := \frac{\mathbf{E}_{\mathbf{P}_t} [X_{t+\Delta} | E] + [\mathbf{P}_t(\bar{E}) - \mathbf{P}_t(E)] \cdot \text{VaR}_{\Delta, \alpha}}{\mathbf{P}_t(\bar{E})}. \quad (8)$$

The expression (8) complies with market practice, e.g. Kerkhof and Melenberg (2003), whereby we have to account for the possibility that the $\text{VaR}_t(\Delta, \alpha)$ might have a finite probability (\mathbf{P} not surjective) or that an interval on X might have 0 probability (\mathbf{P} not injective). Although (7) is of practical use in calculation, this assumption is impractical and quite useless⁵, having a practical financial application only in limited cases⁶. Hence, we drop definitions (7), (8) and assume the cumulative distribution function to be bijective. Let $F_{t, \Delta}(x) := \mathbf{P}_t(X_{t+\Delta} \leq x)$ and define

$$\text{VaR}_t(\Delta, \alpha) = F_{t, \Delta}^{-1}(\alpha) \quad (9)$$

and

$$\text{ExS}_t(\Delta, \alpha) = \frac{1}{\alpha} \mathbf{E}_{\mathbf{P}_t} [X_{t+\Delta} \mathbf{1}_E]. \quad (10)$$

Definitions (9) and (10) are the ones that we refer to in the rest of this work. The corresponding sample measures of the quantities defined in (9) and (10) are either employed in the construction of simultaneous confidence set or enter a loss function between the realized values and their ex-ante expectations, cfr. White (2000). Mainly, the type of tests we set up concern the model forecasting ability of VaR and ExS over 1-day and 10-day time horizons, under low and high volatility conditions and on an average set up. We test the VaR model predictive performance, cfr. Bao et al. (2006), Kerkhof et al. (2009), using the sample measures of the *empirical coverage probability* (ECP)

$$\tilde{\rho} := \frac{1}{P} \sum_{t=R}^{T-\Delta} \mathbf{1}_A. \quad (11)$$

As the empirical probability coverage is a scale independent global measure, in order to construct sample measures for testing the model forecasting ability targeting the expected-shortfall, we need to introduce some standardisation of the data that could possibly render the shortfall forecast time independent. We can indeed achieve that by transforming the data with the Δ -step ahead conditional distribution $\mathbf{P}_t(x_{t+\Delta})$ to construct the *probability transform*:

$$y_{t+\Delta} = \int_{-\infty}^{x_{t+\Delta}} d\mathbf{P}_{t, \Delta}(s), \quad (12)$$

obtaining the random variable $Y \stackrel{\text{iid}}{\sim} \mathbf{U}[0, 1]$, cfr. Rosenblatt (1952), Diebold et al. (1998). Exploiting the sequentially independence property of Y , the latter work approaches the problem of assessing probability density forecasting performance by measuring the divergence of the transformed data from the uniform distribution. Building on this approach, several formal methods of testing density forecasts and applications to financial risk measurement have been designed, based on the likelihood-ratio test, cfr. Berkowitz (2001), the Kullback-Leibler information criterion, cfr. Bao et al. (2007) and the δ -functional method, cfr. Van der Vaart (1998) and Kerkhof and Melenberg (2003). In this article instead, we apply the probability transform and then derive the sample measure

$$\tilde{\rho} := \frac{\sum_{t=R}^{T-\Delta} y_{t+\Delta} \mathbf{1}_{E_{t+\Delta}^*}}{\sum_{t=R}^{T-\Delta} \mathbf{1}_{E_{t+\Delta}^*}} \quad (13)$$

whereas E^* is the event in which the observed y has breached the probability transformed $\text{VaR}_t(\Delta, \alpha)$. The approach undertaken here in constructing MHT of ExS involves estimating simultaneous confidence sets for the measure (13) and testing the theoretical ExS of a uniform distribution at the sought confidence level. To our knowledge, this is the first time that such a test for the expected-shortfall is devised. For reference, the target critical values for the statistics undergoing the testing procedures, are summarized in the following scheme:

α	1%	5%
$\text{VaR}_t(\Delta, \alpha)$	0.01	0.05
$\text{ExS}_t(\Delta, \alpha)$	0.005	0.025

In the generalised bootstrap MHT framework, backtesting the predictive ability of VaR and ExS for a collection of models is a relatively simple exercise. In the absolute model performance test, the method

of balanced confidence set for testing focuses on comparing the theoretical statistic with empirical critical values, rather than the empirical statistic to theoretical critical values. The step-down algorithm further refines the set of models, strengthening the power of true discoveries by rejecting those models that are most unlikely to produce statistically significant forecasts. Another point of view is offered with the relative performance test exercise, whereby we aim to produce a ranking of the model forecasting ability. With this test, we compare the divergence between the distances of the sample measures from their theoretical value for each benchmark model with respect to the rest of the pool, mapping all the possible double comparisons from the pool. Although it is possible to construct balanced confidence sets with control for the k -FWER for this type of test, we believe that the huge amount of multiple comparisons would result in an overflow of information. The testing design strategy here is to estimate the marginal probability of pairwise model superiority to produce a model ranking indicator. In relation to the testing dependencies we rely on the ability of the bootstrap MHT to take into account this feature.

More specifically, with the absolute model performance test, we derive estimates of balanced joint confidence intervals for the target statistic with a predetermined probability and an overall k -FWER, which is specified to control the probability of committing *at least* k false rejections across the suite of models. Thereafter, we leverage upon the duality between statistical tests and confidence intervals to infer conclusions about the significance of the result delivered by each model. For each confidence set, we check if the target critical value is included in the interval controlling the joint error of Type I, and iterate in a step-down fashion to improve the MHT statistical power. Formally, the absolute model forecasting ability test is defined as

$$\mathcal{H}_j: a_{n,K,j}(\alpha, k) \leq R_{n,j}(\mathbf{X}, \rho) \leq b_{n,K,j}(\alpha, k). \quad (14)$$

In the definition of problem (14), we point out the double sided nature of the multiple test and the route taken in testing using the confidence set approach. Practically, the output of the simulation process is the critical value vectors \mathbf{a} and \mathbf{b} , which incorporate the full dependency structure of the experiment and the k -FWER and balancing constraint. The step-down algorithm produces a further refinement of the procedure, as described previously. In the empirical section, we allow for lower confidence on longer time horizons, because of the increased variability of the statistics, while we keep the number of false rejections across the tests in the range of 10 – 13%.

With the relative performance test, we want to draw inference about each model’s predictive ability within the context of all models in the collection. We are targeting the superiority of each model in delivering the target risk measure, with the ultimate aim of answering the question: “which model is the best model?”. It follows intuitively that a model, which cannot deliver statistically significant performance results has little chance of providing superior performance relative to the rest of the pool. But, among sound models, which model should we choose in optimal statistical terms? This the question we tackle by developing relative model predictive ability bootstrap MHT. The relative forecasting ability has been first investigated by the seminal work of White (2000), who designs the Reality Check (RC), a joint statistical inference procedure that extends the methods of Diebold and Mariano (1995) and West (1996), and which has been in turn extended in several directions, cfr. Hansen (2001) and Corradi and Swanson (2006). In recent years, several works on risk model backtesting, see for instance Gonzales-Rivera et al. (2003), Bao et al. (2006), Kerkhof et al. (2009), and density forecast, cfr. Bao et al. (2007) have used the RC framework to cope with joint testing of model forecasting ability. This paper is different in that with the relative model performance exercise our test departs substantially from the Reality Check approach of White (2000), Hansen (2001). In fact, the bootstrap MHT we deliver is structurally different from the RC, which hinges on the statistic $\mathcal{W} := \max_i \mathbb{E}[\mathcal{L}_i - \mathcal{L}_0]$ where \mathcal{L}_i is the loss function of the model i predictive ability and \mathcal{L}_0 is the loss function of the reference model 0. The RC exploits asymptotic results with simulated second moments for the target statistics or simulated statistic distributions to infer realized p -values. Nonetheless, we observe a structural problem with this approach. The statistic \mathcal{W} usually generates distributions centered at the very far right side of the abscissa, entailing that in order for the researcher to reject the null hypothesis $\mathcal{W} \leq 0$, the realized model performance should be large. The standardization procedure proposed by Hansen (2001) attenuates, but does not eliminate, this phenomenon. Besides, in the presence of strictly competing models, the RC is most likely to accept the null every time, even in the standardized version. We conjecture that the form of the statistic is too stringent and, unlike the RC, we construct a different target statistic which measures the gain/loss in terms of distance from the critical value of the performance measures produced by two competing models, formally $R_{n,j}^i(\mathbf{X}, \rho) := |\rho_j - \rho^*| - |\rho_i - \rho^*|$. We propose a novel device in relative model comparison with a multiple testing approach, constructing a matrix, which we call the multiple comparison map, or MCM. An MCM is a double entry table, which compares each model predictive ability versus every other model in the pool. The key statistic is the probability estimate of the event that the absolute performance of

model i is better than the absolute performance of model j , formally $\mathbb{P}\{|\rho_j - \rho^*| - |\rho_i - \rho^*| \geq 0\}$, $\forall i, j$. The latter quantity is the p -value⁷ of the test statistic whose null hypothesis states that model i is better than model j , formally,

$$\mathcal{H}_j: R_{n,j}^i(\mathbf{X}, \rho) \geq 0 \quad (15)$$

that is the distance of the performance measure ρ_j of model j from the critical result ρ^* is greater than the same result for model i . Therefore, at row i , column j of the MCM we can read off the probability that model i has better forecasting ability than model j . Notice that the companion elements above and below the diagonal do not necessarily sum to one, because both the entries include the event that the measure is zero, which is not a zero probability event; in practice, the bootstrapping mechanism is likely to produce discrete distributions. In this implementation, the MCM produces the p -value of the test (15), which represents the estimated likelihood of paired model superiority. Finally, in order to increase the readability of the results, we provide a table which summarises the large amount of pairwise relative model forecasting ability tests. The synthetic table presented in the experimental section exhibits the indicator which is a count of the comparison tests giving the number of times a given model achieves a p -value greater than 0.5, thus producing a ranking of market risk model forecasting ability.

The MCM incorporates the robustness of the generalised bootstrap MHT approach, packaged into a readable display format for measuring the model superiority. It shifts the focus towards the relative performance of the suite of models running a thorough scan of each pair-wise comparison. The MCM provides a valuable tool for answering uniquely the question concerning the best model in the pool.

6 Experimental Section

6.1 Data Description and Modeling Approach

The data set consists of a large sample of a well diversified equity stock index, that is the Dow Jones Industrials Average index, ranging from December 31st 1970 to April 22nd 2013. We work with log-returns of the index daily close level series. This large sample allows us to consider at least two comparable volatility peaks, around October 1987 and October 2008 as well as a high number of volatility waves. To perform the backtesting experiment, we split the data sample into in-the-sample and out-of-sample segments, assuming size $T = R + P$, where R indicates the size of the in-sample data used for model estimation and P indicates the size of the sample used for prediction in the out-of-sample segment. The full sample size is $T = 10,674$. The working assumption here is that there exists stable transition probability distributions, albeit unknown. We subtract the sample average from the return sub-sample ending on December, 31st 1998, assuming thereon a zero off-set constant. We draw on a large sub-sample for first estimation and set $R = 6,572$, that is we start the out-of-sample exercise on January, 1st 1997 and use the same parameters for the parametric models throughout 260 observations, after which the model is estimated again. As a consequence we split the out-of-sample exercise into 16 blocks which are re-sampled 2,000 times with the stationary bootstrap of Politis and Romano (1994b). We choose that sample size so that we observe sensible smoothing of the statistic distributions. The optimal bootstrap block-length is estimated on the growing sample base with the Patton et al. (2009) algorithm. We deliberately discard the rolling-window approach for parametric models like GARCH-EVD and CAViaR because this practice increases rather than shrinks the forecast variability. For instance, the autoregressive coefficient of the symmetric GARCH equation exhibits wide variations if resulting from a two-year rolling sample monthly estimate as opposed to the procedure employed in the experiments consisting of a yearly estimate on a growing sample base. For reference, the mentioned rolling-window approach for a symmetric GARCH model would produce an average autoregressive coefficient of 0.869 with a standard deviation of 0.073 and a spike at 0.346, whereas the growing sample approach delivered an average coefficient of 0.920 with a standard deviation of 0.002. We believe that is the main reason for the poor performance of the GARCH models in, e.g., Kerkhof et al. (2009) and Bao et al. (2006). Hence, in this study we do not include rolling window versions of the GARCH or CAViaR models, relying on the results of similar models employed by other authors for reference. The rolling window approach applied to models that are designed to produce conditional and stationary distributions of the data generating process, is more likely to hurt their performances by the increased model uncertainty introduced with the parameter variability, rather than improving their local forecasts.

In this empirical study, we generate the sample statistics with resampling from subsamples of 260 days. There are two reasons for this: firstly to avoid the phenomenon of *location bias*, Corradi and Swanson (2007), that is the bias in resampling for recursive problems, whereby earlier observations are used more frequently than temporally subsequent observations when forming test statistics; and secondly

to construct artificial samples with classified volatility, in order to investigate the model performance in different volatility environments. We construct model predictive ability measures on 1-day and 10-day projections of the target risk measures, that are the empirical loss function, the empirical coverage probability and the VaR and ExS of the probability transformed sample distributions, whereby the model functional is constructed out of forecast distributions. As described earlier, the out-of-sample data is divided into 16 blocks of 1 calendar year.⁸ We investigate the full out-of-sample performance of the models. Furthermore, we back-test the performance in low / high volatility scenarios each corresponding to four blocks labelled as L/V and H/V, representing extreme sample years. The blocks are not necessarily time-contiguous.

Where necessary, the model 1-step and 10-step distributions are constructed via Monte Carlo integration. In order to consistently reduce the computational time, the GARCH-EVD distributions are constructed on a grid for the conditioning variable entailing an array of forecast distributions, which is used at run time by truncating over the prescribed grid the dependency on the current value. The historical simulation is projected forward via Monte Carlo integration. The RiskMetrics distributional assumption is Gaussian with h_t variance. The Kernel regression estimate is constructed in a similar manner as the GARCH-EVD distributions, that is on a grid for the conditioning variable that is determined on the historical sample as well. In order to reduce the computational time, the Kernel regression is also kept fixed until the subsequent estimation. The rolling window models are recalculated daily at time $t-1$. The CAViaR equation requires some inventiveness to be employed. As they stand, the quantile regressions cannot be projected forward or input in the probability transform, because they have naturally been designed to be free of any distributional assumption. This model is appealing both for the short term memory quantile feature as well as for the absence of an explicit probability assumption. Nevertheless, we need a conditional distribution to feed the Rosenblatt functional and construct the ExS backtesting procedure. Therefore, we proceed by estimating several quantile auto-regressions to construct a linear approximation of the tail of interest. We need to expand on the inner side in order to avoid polarization on the quantile of interest, that is the 5th in this exercise. Meanwhile, on the outer side, we need a tail assumption to work with. We start joining the all time minimum with the first percentile with a straight line and then with a rational function of degree 5, 10 and 20. To be sure that the operation model we are designing produces reasonable results, we ought to prove that the percentile order is what is expected to be. In this case, we rely on the careful choice of the pivot points, on the constraint preventing the autoregressive quantiles to cross each other, were that to happen and, of course, on the empirical evidence. The quantiles are carried forward in time simply multiplying by the square root of time.

Furthermore, we are also interested in the significance of conditioning in the presence of model misspecification. We include in the model a fully unconditional distribution assumption, based on a dual tail GPD distribution with constrained normal or GED distribution for the mid quantiles, estimated through MLE on the base sample. We also consider for the 10-step ahead forecast the unconditional distribution of the GARCH-EVD models. This distribution is constructed taking the expectation with respect to the conditioning variable, formally

$$\mathbb{P}(X) = \mathbb{E}^Y [\mathbb{P}(X|Y)].$$

The rationale in testing these models relies on the possibility that the forecast 10 steps ahead is possibly distorted by the conditioning, firstly because of the speed of the mean reversion of the volatility, which should be ruled out by the preliminary tests on the model parameters significance, but mostly because of possibly a misspecification of the model that might include unexpected innovations that impact rapidly and significantly the model projections.

Table 1 provides a summary of the models and the acronyms that are used in the next section.

6.2 Empirical Evidence

In this section, we summarize the empirical evidence obtained from our testing. To conserve space, we do not present all tables related to the analysis, although this section discuss the outcomes in full. The complete set of tables is available from the authors upon request.

With reference to the full sample of the target market risk measures and then the VaR numbers for the L/V and H/V blocks, Tables 2 to 5 present the balanced k -FWER confidence sets and the generalised bootstrap MHT results. The tables show the balanced confidence set estimates before the application of the step-down procedure, while the shaded cells correspond to those models rejected at the termination

of the aforementioned algorithm. The tables also show the bootstrap mean of the target statistic for each model, which represents the expected model performance. The critical values are applied according to the following scheme:

Measure	Horizon	Confidence	k
ECP	1d	99%	3
ECP	10d	95%	4
ExS	1d	99%	4
ExS	10d	95%	5

Further, in relation to the relative performance comparison, Tables 6 to 9 exhibit the MCM tables, that is the p -values of the pairwise tests for the market risk forecasting model performance for the short term horizon and full sample. As noted earlier, to conserve on space we have not presented the MCMs for the other test settings; however, these are available from the authors upon request. Moreover, in order to facilitate the reading of the model comparison tests, we also provide a collation of the results where we present a synthetic number which produces a model forecasting performance ranking. The index in Table 10 is a count of the number of times that each model produces a probability greater than 0.5 in each pairwise model comparison test.

In Table 2 we exhibit the full sample results for the VaR forecasting experiment. Although the MHT for the empirical coverage probability (ECP) of the VaR(1d, 5%) shows that the HS and the Gaussian models are significant good predictors, in forecasting VaR(1d, 1%) the class of heteroskedastic models augmented with EVD deliver superior results, as well as in the former experiment. The KR can surprisingly capture the 1d first percentile, while being rejected in estimating the fifth percentile: this model generally shows quite erratic performance. The Nadaraya-Watson kernel is quite sensitive to tail data and so is especially erratic in the tails; this model could possibly improve its performance slightly if iterating the estimation daily or using a weighted version of the kernel.⁹ The CAViaR exhibits the same underestimation effect (higher empirical coverage) that is visible in the designers' work Engle and Manganelli (2004), probably due to finite sample effect. Increasing the sample size reduces the bias effect allowing improved performance. The Student-t type models seem to suffer on the 5th percentile exercise. The unconditional models DT_n and DT_ged are systematically rejected in the 1d horizon. In the 10d forecasting exercise the higher variability produces more widespread acceptable model performance, despite having relaxed the Type I error and the generalised FWER. The EVD models seem to slightly overestimate the quantile over the 10d projection; this might be connected to the necessity for improving the likelihood optimisation. The good performance of the CH*t_avg model in the VaR(5%,10d) case, whereby this model is usually affected by critical performance of the higher percentile yet nevertheless performs well in the longer horizon exercise, seems to suggest that the joint estimation of the GARCH filter and the tail model may add to predictive power. In fact, this is the only fat-tailed GARCH model which has been estimated with full MLE. The unconditional GARCH-EVD models are accepted in both experiments and also the DT_* models are significant in the 5th quantile experiment. This performance raises the question concerning "how far" the conditional distribution is from the stationary one over the projection horizon. The RM model produces significant forecasts, except in the short time short tail experiment.

The generalised bootstrap MHT for the VaR in the Rosenblatt space perfectly confirms the results delivered by the empirical coverage probability tests, with the exception of the HS for the VaR(1d, 1%) and the QR2 class for the VaR(1d, 5%), most likely because these models are at the boundary of the acceptance set. In fact an increase of the number of tolerated false rejections induces the acceptance of the null for these tests that were rejected previously. We infer that these models deliver results whose critical values are at the boundary of the acceptance set.

Turning to the ExS experiment, we notice the large number of rejections over the 1d horizon. Contrary to the common sense intuition, the number of models that pass the test is greater in the smaller tail. The non-Gaussian heteroskedastic models with fat tail innovations provide statistically significant predictors for the ExS(1d, 1%), whereas only the symmetric EVD and CH3t are significant at the fifth percentile. In general, the GARCH-EVD are good predictors on shorter horizon, whereas they tend to show slightly biased forecast on the 10d horizon, though still significant. The tail adjustment in the QR model delivers significant results in several cases for the first percentile exercise. The historical simulation, proves to be an acceptable choice on short term horizon and for ExS(10d, 5%), while it fails on the far tail at the long horizon. The 10d horizon exercise shows again more wide spread significant results in the longer tail forecast, whereas in the small tail experiment only a few models outside the CH class can deliver significant results. We also produce but do not report VaR testing under the probability transform, which mostly return the same results as the one delivered by the empirical coverage probability measure. We also exclude unconditional models from the ExS forecasting model suite, which basically produce the

same conclusions as the VaR experiment.

In the stress test experiment, we evaluate the predictive ability of the model battery in a controlled volatility environment targeting the low volatility (L/V), in Table 4 we show the results for the VaR forecasts, and high volatility (H/V) subsamples, shown in Table 5. The general result is that Gaussian models tend to outperform with low volatility, while the GARCH-EVD class, mimicking more probability distribution characteristics, outperform almost systematically. In the low volatility scenario the RM can also deliver significant results.

Finally, in order to enrich the analysis thus far, we construct the MCM for the suite of risk models we have been testing with the generalised bootstrap MHT. This tool is able to garner information concerning the “best” model in the collection. In this exercise, we do not exclude the models that have been rejected at the absolute performance result level, in order to highlight the consistency of this approach with respect to the relative testing procedure. On the short-term horizon forecast, the GARCH-EVD models provide the top performances for all the target risk measures. The implicit GARCH CaViaR with tenth degrees rational tail is most likely the best performer for the ExS(1d, 1%) and ExS(1d, 5%), where few other models of this class exhibiting top performance over the 10d horizon, like some elements of QR2_*. An interesting result is the loss of performance ability of the CH1x model over the longer horizon at the fifth percentile, which coupled with the noticeable results of the CH*t_plus model for the same target, suggest either the necessity of fine tuning the estimation function or for a further expansion of the probability model. The HS model is a mid-rank performer, with few exceptions on one side in the case of the short term ExS forecasts and on the other side in the case of the empirical coverage probability in the first percentile and 10d horizon. Other results that stand out but do not show a pattern are the top performances shown by the rolling Gaussian, the unconditional dual fat tail, the QR1 and the RM model in the case of VaR(10d, 5%).

7 Conclusions

In this article, we extend the exercise carried out in Bao et al. (2006) by means of testing the model performance of a suite of models in forecasting 1-day and 10-day VaR and ExS with a generalised multiple hypothesis testing (MHT) methodology. We present the bootstrap MHT framework, which is a multiple hypothesis testing approach built on a non-parametric simulation device, capable of producing estimates of the statistic distribution of interest and delivering robust inferences that do not rely on analytic or asymptotic results, requiring only weakly stationary times series. Among its characteristics, the MHT is able to deliver statistical tests that take into account multiple test dependencies and that are sound to finite sample effects and non-pivotal statistics. From this perspective, we believe this approach to be particularly appealing to practitioners, due to the fact that it allows one to design robust multiple hypothesis tests, which easily can accommodate decision-making problems that concern competing risk-models.

We apply the MHT approach to produce absolute and relative model forecasting ability performance testing. In the absolute testing exercise, we estimate simultaneous balanced confidence sets controlling for the generalized family-wise error rate. Balancing multiple confidence sets is a highly desirable property of a MHT, in order to construct homogeneous intervals, which target the same confidence probability at the individual hypothesis level. Furthermore, controlling for the k -FWER is of capital importance in MHT because the probability of false discoveries in multiple hypothesis test reaches the boundary very rapidly. The k -FWER extends naturally the concept of confidence level for the scalar statistic. As a final layer, we use a step-down algorithm, which maximises the power of the test. We use this approach to test jointly a set of alternatives. Finally, in the second part of our suite of statistical tests, we exploit the bootstrap MHT to construct a novel relative comparison testing device, the multiple comparison map (MCM), that offers a complementary perspective on the target measure across the collection of alternatives, measuring the p -value of each pairwise individual hypothesis that the reference model provides superior market risk forecasts. The MCM is a novel tool intended to investigate model predictive ability at a more granular level. With the relative model forecasting performance test, we do not apply the procedure for controlling the generalised family-wise error rate, in order to produce a full ranking of the reference model suite. A further layer of robustness can be added by previously filtering the model set via the control for false discoveries in relation to the multiple pairwise model comparisons for each reference model. However, we rely on the capability of the bootstrap engine to take into account the test dependencies. The general objective of producing a ranking in relation to the model’s ability in forecasting the target risk measures is pursued by synthesising the multiple model comparison p -values into an index measure of the model

superiority.

The target risk measures are the VaR and ExS forecast on 1-day and 10-day time horizons. In the latter experiments we present a new simple test for the expected-shortfall, building on the probability transform. In the context of model forecasting ability, the bootstrap MHT framework is attractive, because it can cope with possibly non-nested misspecified models and parameter estimation risk. Concerning this latter issue, we adopt the working hypothesis that the parameters or the non-parametric estimates are set to their p -limits, due mainly to restrictions to the currently available computational power. We plan to expand this feature in future experiments and include estimation risk in the full simulation.

The empirical results that we obtain diverge significantly from the evidence collected in Bao et al. (2006). We have used more stable parameters for the GARCH models and optimized the estimates of the thresholds for the GPD. Furthermore, the generalised MHT approach we have employed, allows testing for the significance of model results in absolute terms, while the MCM determines a comparison of model performance that is more informative than the reality check p -value, which in the reference article fails to detect any model superiority over the benchmark RiskMetrics. We have compared the forecasting ability of several models with respect to different performance measures. In general, models which account for volatility mean reversion and fat-tailedness result in the best performance on the shorter horizon, whereas the heteroskedastic models with a Gaussian specification seem to perform slightly better on the 10d horizon. Most of the time the CH1x_* (see Table 1) is the best model. This result is in contrast with previous findings and shows that GARCH rolling window parameter estimates introduce high model uncertainty, inevitably compromising their forecasting performance. The Historical Simulation also performs reasonably, resulting in mid-ranking performance. Risk Metrics, another popular model among practitioners, does not exhibit a noticeable performance. Conditioning sometimes hurts the performance on a wide time horizon. The 10d horizon experiment shows a wider variability and in the case of the 5% tail a smaller number of rejected models. We also conduct a model stress test in order to investigate the model performance in low/high volatility scenarios, whereby we observe the Gaussian models performing better in a low volatility regime.

Further extensions or improvements that might be brought into the experiment include the following: the parameter estimation risk, that would entail the estimation of parametric and non-parametric models at the resampling level; the inclusion of jump-diffusion models, possibly capable of capturing non-smooth surprises on wider forecasting horizons; the exploration of the soundness of the stable transition probability assumption, with the introduction of switching regime models. We leave this to further research activity.

8 References

- Artzner, P., P. Delbaen, J.M. Eber, and D. Heath, 1999, Coherent measures of risk, *Mathematical Finance* 9.
- Bao, Y., T. Lee, and B. Saltoğlu, 2006, Evaluating predictive performance of value-at-risk models in emerging markets: A reality check, *Journal of Forecasting* 25.
- Bao, Y., T. Lee, and B. Saltoğlu, 2007, Comparing density forecast models, *Journal of Forecasting* 26.
- Beran, R., 1988, Balanced simultaneous confidence sets, *Journal of the American Statistical Association* 83.
- Beran, R., 1990, Refining bootstrap simultaneous confidence sets, *Journal of the American Statistical Association* 85.
- Beran, R., 2003, The impact of bootstrap on statistical algorithms and theory, *Statistical Science* 18.
- Berkowitz, J., 2001, Testing density forecasts with applications to risk management, *Journal of Business and Economic Statistics* .
- Bierens, H.J., 1987, Kernel estimators of regression functions, in *Advances in Econometrics: Fifth World Congress*, volume 1 (Palgrave, London).
- BIS, 2011, Messages from the academic literature on risk measurement for the trading book, Technical Report 19, Bank for International Settlements.

- Bollerslev, T., 1986, Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics* 3.
- Christoffersen, P., 1998, Evaluating interval forecasts, *International Economic Review* 39.
- Corradi, V., and N.R. Swanson, 2006, Bootstrap conditional distribution test in the presence of dynamic misspecification, *Journal of Econometrics* 133.
- Corradi, V., and N.R. Swanson, 2007, Nonparametric bootstrap procedures for predictive inference based on recursive estimation schemes, *International Economic Review* .
- Diebold, F.X., and R.S. Mariano, 1995, Comparing predictive accuracy, *Journal of Business and Economic Statistics* .
- Diebold, F.X., A.G. Todd, and A.S. Tay, 1998, Evaluating density forecasts with applications to financial risk management, *International Economic Review* .
- Efron, B., 1979, Bootstrap methods: Another look at the jackknife, *The Annals of Statistics* 7.
- Engle, R., and S. Manganelli, 2004, Caviar: Conditional autoregressive value-at-risk by regression quantile, *Journal of Business & Economic Statistics* 22.
- Engle, R.F., 1982, Autoregressive conditional heteroskedasticity with estimates of the variance of the united kingdom inflation, *Econometrica* 50.
- Glosten, L.R., R. Jagannathan, and D.E. Runkle, 1993, On the relation between the expected value and the volatility of the nominal excess return on stocks, *The Journal of Finance* 48.
- Gonzales-Rivera, G., L. Tae-Hwy, and M. Santosh, 2003, Forecasting volatility: A reality check based on option pricing, utility functions, value-at-risk and predictive likelihood, *Working Paper* .
- Hansen, P.R., 2001, An unbiased and powerful test for superior predictive ability, *Working Paper* .
- Hansen, P.R., and A. Lunde, 2001, A comparison of volatility models: Does anything beat a garch(1,1)?, *Working Paper* .
- Horowitz, J.L., 2000, The bootstrap, *Working Paper* .
- J.P.Morgan, 1996, RiskmetricsTM, Technical Report 4, J.P.Morgan, Reuters.
- Kerkhof, J., and B. Melenberg, 2003, Backtesting for risk-based regulatory capital, *Working Paper* .
- Kerkhof, J., B. Melenberg, and H. Schumacher, 2009, Model risk and capital reserves, *Working Paper* .
- Koenker, B., and G. Bassett, 1978, Regression quantiles, *Econometrica* 46.
- Kupiec, P., 1995, Techniques for verifying the accuracy of risk management models, *Journal of Derivatives* 3.
- Künsch, H.R., 1989, The jackknife and the bootstrap for general stationary observations, *The Annals of Statistics* 17.
- Lahiri, S.N., 1999, Theoretical comparisons of block bootstrap methods, *The Annals of Statistics* 27.
- Lehmann, E.L., and J. Romano, 2005, Generalization of the familywise error rate, *The Annals of Statistics* 33.
- Liu, R.Y., and K. Singh, 1992, Moving blocks jackknife and bootstrap capture weak dependence, in *Exploring the Limits of Bootstrap* (John Wiley & Sons).
- Longin, F.M., 1996, The asymptotic distribution of extreme stock market returns, *The Journal of Business* 69.
- Mancini, L., and F. Trojani, 2011, Robust value at risk prediction, *The Journal of Financial Econometrics* 9.
- McNeil, A.J., and R. Frey, 2000, Estimation of tail-related risk measures for heteroskedastic financial time series: an extreme value approach, *Working Paper* .
- Nadaraya, E.A., 1964, On estimating regression, *Theory of Probability and its Applications* 9.

- Nelson, D. B., 1991, Conditional heteroskedasticity in asset returns: a new approach, *Econometrica* 59.
- Nordman, D.J., 2009, A note on the stationary bootstrap's variance, *The Annals of Statistics* 37.
- Patton, A., D.N. Politis, and H. White, 2009, Correction to: Automatic block-length selection for dependent bootstrap, *Econometric Reviews* 28.
- Politis, D.N., and J.P. Romano, 1994a, Limit theorems for weakly dependent hilbert space valued random variables with applications to the stationary bootstrap, *Statistica Sinica* 4.
- Politis, D.N., and J.P. Romano, 1994b, The stationary bootstrap, *Journal of the American Statistical Association* 89.
- Politis, D.N., and H. White, 2004, Automatic block-length selection for dependent bootstrap, *Econometric Reviews* 23.
- Romano, J., M.S. Azeem, and M. Wolf, 2009, Hypothesis testing in econometrics, *Working Paper Series, Institute for Empirical Research, University of Zurich* 444.
- Romano, J., and M. Wolf, 2005, Exact and approximate stepdown methods for multiple hypothesis testing, *Journal of the American Statistical Association* 100.
- Romano, J., and M. Wolf, 2007, Control of generalized error rates in multiple testing, *The Annals of Statistics* 35.
- Romano, J., and M. Wolf, 2010, Balanced control of generalized error rates, *The Annals of Statistics* 38.
- Rosenblatt, M., 1952, Remarks on a multivariate transformation, *The Annals of Mathematical Statistics* 23.
- Van der Vaart, A.W., 1998, Functional delta method, in *Asymptotic Statistics*, chapter 20 (Cambridge University Press).
- Watson, G.S., 1964, Smooth regression analysis, *Sankhyā: The Indian Journal of Statistics* 26.
- West, K.D., 1996, Asymptotic inference about predictive ability, *Econometrica* 64.
- White, H., 2000, A reality check for data snooping, *Econometrica* 68.

Notes

¹We keep referring to VaR in this work although we always mean the target quantile of the empirical returns. Nevertheless, originally, the VaR is defined as a monetary measure that better refers to a consolidated portfolio of asset values rather than returns. However, it is theoretically easy to switch from one measure to another.

²The statistical power of a test is defined as the probability of rejecting false null hypothesis, that is one minus the probability of committing Type **II** error. The power measures the ability of the test in rejecting false hypotheses.

³In statistics, a root is a generic term to indicate a function of the sample data. We use indifferently as a synonym for statistic, or as a more elaborate function.

⁴The RiskMetrics model is presented as a heuristic model because the model parameter θ is fixed a-priori and we assume conditional normality to project the system forward.

⁵Moreover, as we will see later on, we have to assume an invertible distribution function for the application of the Rosenblatt transform.

⁶The main example of this sort is a credit risk model with constant recovery.

⁷To be precise, here we refer to the ex-ante probability of the null hypothesis, that is before observing any statistic. This p -value is the probability of committing a type **I** error in each pairwise comparison, where the null hypothesis corresponds to the superior performance of the reference model. The confidence level has been set to 0.5, because in each comparison we pick the model with the highest probability of not being rejected as the superior performer.

⁸The last block is less than 1 year but the resampled data is let run for a full year.

⁹We have run some experiments with the weighted version of this model which did not seem to stabilize the tails.

Name	Model
CH1g	Gaussian GARCH(1,1)
CH1t	Student-t GARCH(1,1)
CH1x_n	GARCH(1,1) with $\varepsilon \sim$ dual GPD tailed and normal mid-quantile
CH1x_ged	GARCH(1,1) with $\varepsilon \sim$ dual GPD tailed and GED mid-quantile
CH2g	Gaussian TARCH(1,1)
CH2t	Student-t TARCH(1,1)
CH2x_n	TARCH(1,1) with $\varepsilon \sim$ dual GPD tailed and normal mid-quantile
CH2x_ged	TARCH(1,1) with $\varepsilon \sim$ dual GPD tailed and GED mid-quantile
CH3g	Gaussian EGARCH(1,1)
CH3t	Student-t EGARCH(1,1)
CH3x_n	EGARCH(1,1) with $\varepsilon \sim$ dual GPD tailed and normal mid-quantile
CH3x_ged	EGARCH(1,1) with $\varepsilon \sim$ dual GPD tailed and GED mid-quantile
CH1g_avg	Unconditional Gaussian GARCH(1,1)
CH1t_avg	Unconditional Student-t GARCH(1,1)
CH1x_n_avg	Unconditional GARCH(1,1) with $\varepsilon \sim$ dual GPD tailed and normal mid-quantile
CH1x_ged_avg	Unconditional GARCH(1,1) with $\varepsilon \sim$ dual GPD tailed and GED mid-quantile
CH2g_avg	Unconditional Gaussian TARCH(1,1)
CH2t_avg	Unconditional Student-t TARCH(1,1)
CH2x_n_avg	Unconditional TARCH(1,1) with $\varepsilon \sim$ dual GPD tailed and normal mid-quantile
CH2x_ged_avg	Unconditional TARCH(1,1) with $\varepsilon \sim$ dual GPD tailed and GED mid-quantile
CH3g_avg	Unconditional Gaussian EGARCH(1,1)
CH3t_avg	Unconditional Student-t EGARCH(1,1)
CH3x_n_avg	Unconditional EGARCH(1,1) with $\varepsilon \sim$ dual GPD tailed and normal mid-quantile
CH3x_ged_avg	Unconditional EGARCH(1,1) with $\varepsilon \sim$ dual GPD tailed and GED mid-quantile
DT_n	Unconditional dual GPD tailed and normal mid-quantile
DT_ged	Unconditional dual GPD tailed and GED mid-quantile
G0	2 years rolling window Gaussian with 0 mean
Gm	2 years rolling window Gaussian
HS	2 years rolling window Histogram
KR	Kernel Regression
RM	RiskMetrics
QR1	Adaptive CAViaR
QR1_005	Adaptive CAViaR with 5 degree rational tail
QR1_010	Adaptive CAViaR with 10 degree rational tail
QR1_020	Adaptive CAViaR with 20 degree rational tail
QR2	Symmetric CAViaR
QR2_005	Symmetric CAViaR with 5 degree rational tail
QR2_010	Symmetric CAViaR with 10 degree rational tail
QR2_020	Symmetric CAViaR with 20 degree rational tail
QR3	Asymmetric CAViaR
QR3_005	Asymmetric CAViaR with 5 degree rational tail
QR3_010	Asymmetric CAViaR with 10 degree rational tail
QR3_020	Asymmetric CAViaR with 20 degree rational tail
QR4	GARCH Indirect CAViaR
QR4_005	GARCH Indirect CAViaR with 5 degree rational tail
QR4_010	GARCH Indirect CAViaR with 10 degree rational tail
QR4_020	GARCH Indirect CAViaR with 20 degree rational tail

Table 1: Name and Model Type.

	ExS 1%		ExS 5%		VaR 1%		VaR 5%	
	1d	10d	1d	10d	1d	10d	1d	10d
CH1g	10	25	26	42	6	27	22	23
CH1t	18	1	9	1	16	21	6	1
CH1x_n	33	35	32	10	22	30	21	8
CH1x_ged	32	34	33	9	22	30	21	8
CH2g	9	32	19	38	4	34	19	22
CH2t	24	0	13	0	19	15	7	0
CH2x_n	30	28	29	12	18	26	12	10
CH2x_ged	29	26	30	11	18	26	12	10
CH3g	6	31	10	44	3	33	13	28
CH3t	25	15	34	2	20	22	16	3
CH3x_n	20	36	24	26	15	32	9	21
CH3x_ged	19	37	23	27	15	32	9	21
CH1g_avg	-	13	-	4	-	2	-	2
CH1t_avg	-	40	-	34	-	24	-	29
CH1x_n_avg	-	22	-	17	-	12	-	14
CH1x_ged_avg	-	23	-	18	-	12	-	15
CH2g_avg	-	19	-	6	-	8	-	4
CH2t_avg	-	41	-	37	-	28	-	26
CH2x_n_avg	-	31	-	24	-	17	-	16
CH2x_ged_avg	-	32	-	25	-	17	-	16
CH3g_avg	-	14	-	7	-	9	-	6
CH3t_avg	-	33	-	35	-	24	-	29
CH3x_n_avg	-	21	-	22	-	14	-	18
CH3x_ged_avg	-	20	-	21	-	14	-	18
DT_n	12	8	4	15	11	4	2	32
DT_ged	11	7	5	14	11	4	2	32
G0	2	6	6	16	1	6	18	34
Gm	1	2	2	5	0	0	15	19
HS	31	12	31	20	13	7	14	25
KR	0	10	0	3	12	2	0	5
RM	8	11	14	34	7	18	17	33
QR1	5	3	20	43	9	20	6	34
QR1_005	16	16	17	39	-	-	-	-
QR1_010	28	36	15	36	-	-	-	-
QR1_020	21	45	11	31	-	-	-	-
QR2	4	4	28	46	8	19	10	27
QR2_005	15	17	27	44	-	-	-	-
QR2_010	26	37	25	41	-	-	-	-
QR2_020	23	44	22	39	-	-	-	-
QR3	3	5	21	32	5	10	4	16
QR3_005	14	18	18	30	-	-	-	-
QR3_010	27	42	16	29	-	-	-	-
QR3_020	22	43	12	28	-	-	-	-
QR4	7	9	8	23	3	5	3	11
QR4_005	17	25	7	19	-	-	-	-
QR4_010	34	46	3	13	-	-	-	-
QR4_020	13	26	1	8	-	-	-	-
High rank	QR4_010	QR4_010	CH3t	QR2	CH1x_n	CH2g	CH1_g	G0
	CH1x_n	QR1_020	CH1x_ged	QR2_005	CH1x_ged	CH3g	CH1_n	QR1
	CH1x_ged	QR2_020	CH1x_n	CH3g	CH3t	CH3x_ged	CH1_ged	RM
	G0	Gm	Gm	CH3t	QR4	CH1g_avg	DT_n	CH1g_avg
	Gm	CH1t	QR4_020	CH1t	G0	KR	DT_ged	CH1t
Low rank	KR	CH2t	KR	CH2t	Gm	Gm	KR	CH2t
Max rank	34	46	34	46	22	34	22	34

Table 10: The table contains the model forecasting performance ranking index. The number is a counter of the number of times the corresponding reference model produces a p -value greater than 0.5 in each pairwise relative forecasting ability test.