Received: 30 November 2022 | Revised: 21 September 2023 | Accepted: 28 September 2023

DOI: 10.1002/asmb.2828

RESEARCH ARTICLE

WILEY

Assessing model risk in financial and energy markets using dynamic conditional VaRs

Angelica Gianfreda[1,2] | **Giacomo Scandolo**[3]

[1]DEMB, University of Modena and Reggio Emilia, Modena, Italy

[2]Energy Markets Group, London Business School, London, UK

[3]DISEI, University of Florence, Florence, Italy

Correspondence
Giacomo Scandolo, DISEI, University of Florence, Florence, Italy.
Email: giacomo.scandolo@unifi.it

Abstract

It has been recognized that model risk has an important effect on any risk measurement procedures, particularly when dealing with complex markets and in the presence of a wide range of implemented models. We consider a normalized measure of model risk for the forecast of daily Value-at-Risk, combined with a model selection and an averaging procedure. This allows us to restrict the set of plausible models on a daily basis, making the initial choice of competing models less crucial and then yielding a more reliable assessment of model risk. Using AR-GARCH-type models with different distributions for the innovations, we assess the dynamics of model risk for different financial assets (a stock, an equity index, an exchange rate) and commodities (electricity, crude oil and natural gas) over 15 years.

KEYWORDS

commodity risk, forecasting, model uncertainty, risk management

1 | INTRODUCTION AND BACKGROUND

Big data, artificial intelligence and machine learning are dramatically fostering the number of models being developed and deployed. When these models are integrated into business decision-making situations, institutions are becoming exposed to greater model risk and subsequent potential losses. Therefore, the importance of model risk management has increased rapidly. And, it refers to all phases, from models design and their implementation to the final quantification of capital requirements. Therefore, the proposed approach aims at providing a way to compare and assess model risk when a large number of models can be considered.

In the financial literature, it is well-known that the choice of the underlying probabilistic model for the risk factors can have a significant impact on risk forecasts.[1] Indeed, it has been observed that the range of possible risk values, computed using risk metrics such as the Value-at-Risk (VaR) or the Expected Shortfall (ES), can be surprisingly wide even imposing apparently stringent constraints on the distribution of risk factors; in this framework see for instance Reference 2 in the univariate case, and Reference 3 in the multivariate one.

The hazard of producing a poor risk assessment due to the choice of an unsuited model is usually called *model risk*. It is common to distinguish between two aspects of model risk, that is the *estimation risk* and the *misspecification risk*. The former one refers to the uncertainty arising from parameters estimation (or calibration), once a parametric family of distributions has been chosen. The literature on this aspect is well developed; see References 4-7, among many others.

Whereas, the latter one refers to the choice of the parametric family itself[1] and its subsequent quantification and management is more difficult and it has been less investigated[2]. Then, our paper aims at filling this gap.

To deal with the misspecification risk, the Worst-Case and the Model Averaging approaches have emerged. For both of them, the starting point is represented by a finite set of plausible *models*, that is, parametric families of distributions for the relevant risk factors. In order to obtain a completely specified distribution, each family is fitted to the data, hence providing a *fitted model*.

On one hand, in the Worst-Case approach, the maximum and minimum values for the risk measure ρ (denoted $\rho_{\max}$ and $\rho_{\min}$) are computed over all competing fitted models. The most conservative risk value, that is, $\rho_{\max}$, is then taken as a "robust" forecast. Moreover, the amount of model risk can be quantified as some "distance" between $\rho_{\max}$, $\rho_{\min}$, and, possibly, the risk value ρ^* obtained under a fitted *reference* model or simply obtained by some standard procedures (as the historical simulation); see References 8,10 and 11 for different proposals in this regard. Within this approach, the considered models are usually disjoint (as parametric families) or at least non-nested.[3]

On the other hand, the Model Averaging approach requires the specification of a weight for each model, and then uses the set of weights to average out the risk values obtained under different fitted models. Model risk can then be assessed through some (weighted) dispersion measures of the risk values. Within this approach, the number of models can be substantial and nested specifications are often considered, particularly in multiple regression analyses. This approach may come in a Bayesian form in which prior weights are updated to posterior weights with respect to data, and in a simpler frequentist form in which weights are directly built on the basis of some fitting ability criterion, generally based on the maximized likelihood. The Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC) are common choices for this purpose, as they reward fitting performance, while penalizing for over-parametrization. See Reference 12 for a comparison of Bayesian and frequentist model averaging.

In the present paper, we aim at empirically assessing misspecification risk when forecasting daily VaR (at 1% and 5% levels), for some selected financial and energy assets over years 2001–2015. In particular, we adopt a mixed approach, between the Worst-Case and the (frequentist) Model Averaging approach. Through a dynamic daily procedure, we quantify for each asset the so-called *Relative Measure of Model Risk* (henceforth RMMR), introduced in Reference 10. The definition of this measure relies on the maximum, minimum and *reference* risk values among a class of competing fitted models.

Given that it has been argued that the Worst-Case approach relies on the class of plausible models in a critical way, we consider here a mix of the Worst-Case and the (frequentist) Model Averaging approaches. Specifically, for each model and asset we compute daily the Information Criteria (henceforth IC, for AIC or BIC) based on maximized likelihood, and then we use it for different purposes: first, we select daily the *best model*, that is, the one having the lowest[4] IC, to be used to compute the reference risk value; second, we build weights of models, based on the IC, and use them to derive an average forecast, again to be used as the reference risk value; third, and more importantly, we use the weights in order to trim the original set of plausible models, discarding the worst performing ones. The goal of this last step is to make the assessment of model risk less reliant on the initial choice of competing models. We investigate model risk, assessed via RMMR, with different combinations of the set of models and of the choice of the reference risk value. Additionally, we inspect two different weights constructions, one based on AIC and the other one based instead on BIC. Moreover, we follow three alternative strategies for the choice of the reference risk value: in the first one, a particular reference model is a priori selected and kept for the entire sample; in the second one, the best fitting model is chosen on a daily basis, where the fitting ability is measured by IC; in the third one, the (possibly trimmed) weighted average of risk values is considered on a daily basis. We emphasize that the measure of model risk that we use produces a pure number, independent from the reference currency, hence allowing for immediate comparisons across different assets, countries and markets.

[1]For example, if we assume that stock returns follow an unconditional Student-t distribution, characterized by three parameters (location, scale and degrees of freedom), then the estimation risk arises from the variability inherent in the estimation process of the parameters. Indeed, the outcome of this process depends on the data set, on the estimation window and possibly on the employed inference method (e.g., maximum likelihood vs. method of moments). Instead, the misspecification risk occurs at a more fundamental level and it is due to the choice of the parametric distribution itself: where in our example, it is set to be the Student-t instead of some alternative ones, such as the Normal or Skew Student-t distributions.

[2]For completeness, it is worth to mention that there are additional sources of model risk (such as the *identification* and *granularity risk*), which are however not considered here. Further details can be found in References 8 and 9.

[3]A model is *nested* if, viewed as a family of distributions, is contained in another model. For instance, the Student-t model is nested in the Skew Student-t one.

[4]We note that such Criteria are *decreasing* in the maximized likelihood and *increasing* in the "complexity level" of the model.

Recent papers close to our contribution are References 11,13,14 and 15 provide an empirical dynamic assessment of model risk for a range of US financial assets, proposing a measure of model risk defined as the ratio of the maximum over the minimum VaR. Contrary to what we do here, they do not consider a *reference estimate* and implicitly give all models the same level of plausibility, thus making the initial choice of the competing models more crucial than in our paper; we view this as a drawback of their approach. More similarly to us, Reference 13 propose a methodology to rule out models not passing some standard back-tests; then, they compute the mean absolute deviation of the VaR estimates under all remaining models; and, finally, divide it by the mean estimate, in order to obtain a possible quantification of model risk. Still, in their approach, no reference model is involved and all models receive equal weights. Moving to model risk in energy markets,[14] quantify parameter uncertainty in complex stochastic models where the sensitivities of a derivative value, corresponding to specific pricing models, are used to quantify model risk in the economic evaluation of power plants. And, a similar approach is taken by Reference 15, who assess the impact of parameter uncertainty on the adaptation of investments to climate change. However, these latter two studies have a focus on valuation instead of risk forecasting and do not compare estimation models as we do.

As far as the model design is concerned, we rely on the literature on risk forecasting which makes a widespread use of GARCH-type models, since they are able to capture well-known features of financial time series, as non-normality and volatility clustering. Moreover, their flexibility allows to model additional features, such as conditional skewness, conditional excess kurtosis, and leverage effects. Therefore, following References 16-18, and 13 among many others, we consider a set of competing AR-GARCH-type models, coupled with different parametric families of distributions for the standardized innovations.

Given that the detection, quantification, and management of model risk are becoming crucial tasks, particularly in energy and commodity markets where modelling is often complex, we provide empirical examples based on real data for financial assets and commodities, studying the evolutions of RMMR from 2001 to 2015 for the Deutsche Bank stock prices, the USD/EUR exchange rate, the German equity index, Brent crude oil prices, ICE UK natural gas prices, and the German day-ahead auction electricity prices; hence, enlarging the set of energy commodities investigated in earlier studies.[5]

The paper is structured as follows: the definition of the Relative Measure of Model Risk, the construction of the weights, and the description of the considered GARCH models are reported in Section 2. Data description and the preliminary analysis are presented in Section 3, whereas empirical and simulated results are provided in Section 4. Finally, Section 5 concludes.

2 | METHODOLOGY

2.1 | General setting

Let S denote the value of a financial variable, for example, a price or an index level. On day t, the value that the variable S will take on the next day (that is S_{t+1}) is not known, but there is usually considerable interest in forecasting some of its (probabilistic) characteristics. In particular, for fixed $\alpha \in (0,1)$, for example, $\alpha = 1\%$ or 5%, we are interested in forecasting the *daily* Value-at-Risk at level α, that is the quantity VaR implicitly defined by[6]

$$P(S_{t+1} \leq S_t - \mathrm{VaR}) = \alpha.$$

Here, P is the probability on the underlying measurable space $(\Omega, \mathcal{F})$, where all random variables are defined. Equivalently, if F is the cumulative distribution function (cdf) of the price/index *change variable* $Y_{t+1} = S_{t+1} - S_t$, that is, $F(x) = P(Y_{t+1} \leq x)$, then we can write

$$\mathrm{VaR} = -F^{-1}(\alpha). \tag{1}$$

In other words, VaR is (minus) the α-quantile of Y_{t+1}. Strictly speaking, all above probabilities are conditional on the information set available up to day t, $\mathcal{I}_t$, but we do not explicitly include this aspect in the notation, for ease of exposition.

[5]See References 19,20 and 17 among those considering energy brent, crude oil, heating oil, propane, and gasoline prices.

[6]In the present paper, all distributions of S_{t+1} are absolutely continuous; in the more general case, the definition $\mathrm{VaR} = \sup\{x : P(S_{t+1} \leq S_t - x) \geq \alpha\}$ may be employed.

In the presence of competing distributions for Y_{t+1}, each one yields a different VaR forecast through (1). In this paper, a *model* is a parametric family $M = \{F(\,\cdot\,;\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ of univariate cdfs, depending on a vector of parameters $\boldsymbol{\theta}$ lying in a parameter set $\Theta \subset \mathbb{R}^p$, $p \geq 1$. Each cdf is a plausible distribution for Y_{t+1}, conditional on the information up to time t. As such, the specification of F may contain, apart from the parameters, past observations of the variable S or related variables, as better described in what follows.

We consider a finite set $\mathcal{M} = \{M_1, \ldots, M_K\}$ of competing models for Y_{t+1} (in the present paper $K = 9$). For each $k = 1, \ldots, K$, the model $M_k = \left\{ F_k\left(\,\cdot\,;\boldsymbol{\theta}^{(k)}\right) : \boldsymbol{\theta}^{(k)} \in \Theta^{(k)} \right\}$, with $\Theta^{(k)} \subset \mathbb{R}^{p_k}$, is characterized by a different functional form of the cdfs (F_k), different parameters ($\boldsymbol{\theta}^{(k)}$) and, possibly, a different number of parameters (p_k). Nested models are allowed (i.e. $M_k \subset M_l$). Each model M_k is fitted to data,[7] resulting in a fitted vector of parameters $\widehat{\boldsymbol{\theta}}^{(k)}$ and an associated *fitted model* $\widehat{F}_k(\,\cdot\,) = F_k\left(\,\cdot\,;\widehat{\boldsymbol{\theta}}^{(k)}\right)$. Finally, such a fitted model is used, via (1), in order to obtain a risk forecast, that we denote VaR_k.

As proposed in Reference 10, the *Relative Measure of Model Risk* (henceforth RMMR) for using a particular *reference forecast* VaR^* for VaR at level α, in the presence of the competing models in $\mathcal{M}$, is defined by

$$\mathrm{RMMR} = \frac{\max_{\mathcal{M}} \mathrm{VaR} - \mathrm{VaR}^*}{\max_{\mathcal{M}} \mathrm{VaR} - \min_{\mathcal{M}} \mathrm{VaR}} \tag{2}$$

where $\max_{\mathcal{M}} \mathrm{VaR} = \max_k \mathrm{VaR}_k$ and $\min_{\mathcal{M}} \mathrm{VaR} = \min_k \mathrm{VaR}_k$ are respectively the highest and lowest risk values obtained from the family of K fitted models.

The higher is RMMR, the lower is VaR^* with respect to the other risk forecasts. In other words, a high level of RMMR provides a signal that the reference forecast VaR^* may be too optimistic and may significantly underestimate the actual riskiness of the asset/index, a much undesirable situation. On the contrary, a low value of RMMR means that the reference forecast is overestimating risk with respect to all other competing models. This is of course not desirable either, but less so than risk underestimation.

It is natural to set the reference risk forecast either as the risk value under a given model (i.e., $\mathrm{VaR}^* = \mathrm{VaR}_{k^*}$ for fixed k^*) or as a weighted average under all models (i.e., $\mathrm{VaR}^* = \sum_k w_k \mathrm{VaR}_k$ for given non-negative weights summing to 1). We note that in both cases VaR^* necessarily lies between $\min_{\mathcal{M}} \mathrm{VaR}$ and $\max_{\mathcal{M}} \mathrm{VaR}$, so that RMMR is in the unit interval $[0, 1]$. However, it is also possible to set VaR^* by other means, for instance via a model not belonging to the set $\mathcal{M}$, or some other forecasting procedure, not necessarily linked to the specified competing models (e.g., historical simulation). In this case, VaR^* does not need to lie between $\min_{\mathcal{M}} \mathrm{VaR}$ and $\max_{\mathcal{M}} \mathrm{VaR}$, and this means that RMMR may turn negative or higher than 1. We do not view this as a drawback: indeed, when $\mathrm{VaR}^* < \min_{\mathcal{M}} \mathrm{VaR}$ it means that the reference forecast VaR^* is lower than the lowest plausible forecast, and VaR is so underestimated that the model risk is higher than 1. For instance, if RMMR = 3, then VaR^* equals $\max_{\mathcal{M}} \mathrm{VaR} - 3 \cdot (\max_{\mathcal{M}} \mathrm{VaR} - \min_{\mathcal{M}} \mathrm{VaR})$, or, in plain words, VaR^* is lower than the highest plausible VaR by 3 times the range of plausible VaR.

It is immediately seen that RMMR is insensitive to scaling, meaning that if we change the currency or we rescale the indices by a fixed constant, the resulting RMMR does not change.[8] This is a desirable feature of this measure of model risk, not shared by other proposals made in the literature; see Reference 10 for a discussion on this important point and further properties of RMMR.

The RMMR value clearly depends on both $\mathcal{M}$ and VaR^*: indeed, excluding or adding competing models, or changing the reference forecast surely affects the quantity computed in (2). In order to curb this sensitivity, one can consider, as we do in the present paper, two procedures in which the set of competing models and the reference forecast are not fixed, but they may change from day to day. These changes are solely driven by data and are based on empirical evaluation, thus making the assessment of model risk less subjective and more intrinsic. We start with a set $\mathcal{M}$ of K competing models, and we consider $\mathcal{M}$ to be wide enough for all our purposes. On a daily basis, we consider the VaR forecast under the best fitting model as the reference one ($M^* \in \mathcal{M}$); note that M^* may (and does) change as we roll over the data window. In addition, the weights w_k reflecting the fitting ability of each model are computed and used for two purposes: first, they are used to discard the worst fitting models, thus reducing the set of competing models to a set of *plausible* ones, $\mathcal{P} \subset \mathcal{M}$; second, they are used to compute the weighted average of risk forecasts under the plausible models, thus giving more influence to best fitting models. This weighted average, updated on a daily basis, is then used in place of VaR^* in the definition of RMMR. In the rest of this section, we are going to describe the above procedure in detail.

[7] At this stage, it is irrelevant which inference method is used.

[8] This is true even if VaR is replaced by another positively homogeneous risk measure, such as Expected Shortfall.

2.2 | Model selection and averaging

As described before, each single model M is specified by means of a p-dimensional vector of parameters $\boldsymbol{\theta}$, which may include coefficients of regression of the conditional mean and variance, skewness and shape parameters.[9] Such parameters are estimated via Maximum Likelihood on past data, resulting in a risk forecast computed using (1) under the fitted model $\widehat{F}$. As a by-product of the estimation process, we compute the maximized value of the Likelihood function, $\widehat{L}$. This quantity or, equivalently, the maximized Log-Likelihood, $\widehat{\ell} = \log \widehat{L}$, provides an important tool for ranking all considered models according to their *plausibility*: the higher is $\widehat{\ell}$, the better is the fit of the model to the data.

However, it has been argued that $\widehat{\ell}$, when used for comparisons, should be corrected for the number of parameters used to specify the model. This leads to two celebrated measures of fitting ability: the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), introduced in References 21 and 22, respectively. They are defined as

$$\mathrm{AIC} = -2\widehat{\ell} + 2p \qquad \mathrm{BIC} = -2\widehat{\ell} + \log(n)\, p \tag{3}$$

where p is the number of parameters in a model and n is the length of the dataset. Roughly speaking, the lower is AIC or BIC, the better a model adapts to the given dataset. Both Information Criteria penalize for over-parametrization, but, everything else being equal, the BIC penalization is stronger (indeed $\log n > 2$ as long as $n \geq 8$), hence more desirable for our purposes. The two criteria differ under other aspects: provided the "true" model is in the class of competing models, the BIC is shown to detect it with probability tending to 1 as n increases, a property not shared by the AIC. In turn, the AIC proves superior to the BIC under the opposite assumption that the true model is not among those considered. We refer to References 23 and 12 for extensive discussion on AIC, BIC and other Information Criteria, not considered in the present paper. In Section 4, we perform a preliminary comparison between results obtained through AIC and BIC, showing the superiority of the latter criterion, at least for the goals of this paper.

At every given date and for any given asset/index, we declare as the "best" model the one providing the lowest value of the Information Criterion IC (either AIC or BIC). Formally, the (*daily*) *best* model is

$$M_{\mathrm{best}} = \arg.\min \{\mathrm{IC}_k \ : \ M_k \in \mathcal{M}\} \tag{4}$$

where IC_k is the Information Criterion computed for the model M_k. We can then use the estimate of VaR under the fitted model resulting from M_{best}, that is $\mathrm{VaR}_{\mathrm{best}}$, as the reference model in (2), and denote $\mathrm{RMMR}_{\mathrm{best}}$ the resulting measure.

Going further, we use the IC values to build weights $w_k \in [0, 1]$, one for each model M_k, that sum up to one. Such weights should be decreasing functions of IC values, in such a way that models with good fitting ability have more influence: below, we present two possible constructions. Once the weights have been set, we can compute the weighted average of all risk forecasts as

$$\mathrm{VaR}_{\mathrm{avg}} = \sum_{k=1}^{K} w_k \mathrm{VaR}_k \tag{5}$$

where VaR_k is the VaR forecast obtained under model M_k. This definition is a simple instance of a *combination of forecasts*, a procedure which may substantially improve the accuracy of any of the individual forecasts; see Reference 24 for a review. By using $\mathrm{VaR}_{\mathrm{avg}}$ in place of VaR^* in (2), we obtain a quantity that we denote $\mathrm{RMMR}_{\mathrm{avg}}$.

2.3 | Building weights

As proposed in Reference 25, a simple way to build weights consists firstly in computing the so called *IC differences* defined as

$$\Delta_k = \mathrm{IC}_k - \min_{j=1,\dots,K} \mathrm{IC}_j \geq 0, \qquad k = 1, \dots, K$$

[9]For instance, an AR(5)-GARCH(1,1) model with Student-t innovations is specified in terms of 10 parameters: six coefficients for the conditional mean equation expressed as an Auto-Regressive process with five lags, three coefficients for the conditional variance equation following a GARCH model, and finally one parameter for the degrees of freedom of the (standardized) Student-t distribution of innovations.

so that $\Delta_k = 0$ for the best model (i.e., the one with lowest IC); and, secondly, in defining the so-called *Akaike weights* as

$$w_k = \frac{\exp(-\Delta_k/2)}{\sum_{j=1}^{K} \exp(-\Delta_j/2)}. \tag{6}$$

By their definition, these weights are all positive and sum up to unity. The motivation for this construction is discussed in Reference 23.[10]

However, Akaike weights tend to take extreme values close to 0 or 1, as we document later. Hence, we propose an alternative construction. First, we normalize all Δ_k to the interval $[0, 1]$ through

$$a_k = \frac{\max_j \ \Delta_j - \Delta_k}{\max_j \ \Delta_j}$$

in such a way that a low IC (i.e., $\Delta_k \simeq 0$) corresponds to $a_k \simeq 1$ and a high IC (i.e., $\Delta_k \simeq \max_j \Delta_j$) corresponds to $a_k \simeq 0$. Then, we define the new weights as

$$w_k = \frac{a_k^2}{\sum_{j=1}^{K} a_j^2}. \tag{7}$$

Again, it is immediately seen that these weights are non-negative and that their sum is one. Note that in both (6) and (7), the weight w is decreasing in IC so that "good" models (i.e., models with low IC) receive high weight. The Akaike weights have a theoretical appeal, while the ones in (7) seem quite ad-hoc. However, contrary to Akaike weights, the ones we propose in (7) show a homogeneous empirical pattern, observed over almost all days and assets/indices (see Table 5 for further insights).

Since we are going to trim the set of models by excluding those with small weight, relying on (6) sometimes results in an excessive shrinking of $\mathcal{M}$. For the purpose of assessing model risk, the goal of the selection process should be aimed at discarding the few models that perform particularly bad, rather than picking the few ones that perform particularly well. For this reason, we believe that the weights we propose in (7) represent a better choice than Akaike ones. As we show later, this insight is confirmed by the empirical results.

2.4 | Identifying the set of plausible models

The weights can be used for building a subset of plausible models on a daily basis. Indeed, by sorting them in descending order, we can keep just the first ordered models until the cumulative weight passes a given threshold. In this way, the initial set $\mathcal{M}$ of competing models can be shrunk by discarding the less plausible ones. Then, the maximum and minimum in (2) can be computed on this reduced set of models. This use of weights in order to produce a sort of "confidence interval" of models is discussed in Reference 23. Explicitly, let us assume that on a certain day the weights are sorted in descending order as[11]

$$w_{\pi(1)} > w_{\pi(2)} > \ \dots \ > w_{\pi(K)}$$

where π is a suitable permutation of $\{1, \dots, K\}$. Let k' be the lowest index greater or equal to 2 such that

$$\sum_{k=1}^{k'} w_{\pi(k)} \geq 0.95.$$

[10] Basically, for two models having the same number of parameters, the ratio of their weights reduces to the ratio of their Likelihoods. This is true both for AIC and for BIC (in the latter case provided the employed datasets have equal length).

[11] In practice, the possibility of ties is remote.

Then, the set

$$\mathcal{P} = \{M_{\pi(k)} \ : \ 1 \leq k \leq k'\} \subset \mathcal{M} \tag{8}$$

can be interpreted as a sort of 95% confidence interval for the set of models, and we call it the set of *plausible models* (on a given date).

The construction of $\mathcal{P}$ obviously depends on the threshold level; in principle, the higher is the threshold, the larger is the set of plausible models. The choice of 0.95 is suggested in Reference 23, since it is the typical level employed for building confidence intervals for estimated parameters. Later, we provide some empirical evidence on the robustness of $\mathcal{P}$ with respect to the choice of the threshold.

Various alternative approaches for constructing a set of plausible models have been proposed in the statistical literature. For instance, Reference 23 propose a cut-off applied directly on the Information Criterion or on the maximized Log-Likelihood, while Reference 26 discard models having a weight below 5% of the maximum weight. A quite different approach has been presented in Reference 27, where a so called *model confidence set* is produced through a sequential procedure, involving at any step, an equivalence test (for detecting whether or not all models are *equally good*) and an elimination rule (for discarding *bad* models). The procedure is run until all models are deemed equally good. This approach is proved to possess good asymptotical properties, but we do not consider it in the present paper.

Once the subset $\mathcal{P} \subset \mathcal{M}$ has been built, then it is natural to restrict the average of VaR in (5) only to the plausible models, hence arriving to a new definition of weighted average VaR as follows[12]

$$\mathrm{VaR}_{\mathrm{avg}} = \frac{\sum_{M_k \in \mathcal{P}} w_k \mathrm{VaR}_k}{\sum_{M_k \in \mathcal{P}} w_k}. \tag{9}$$

Even though this new definition of $\mathrm{VaR}_{\mathrm{avg}}$ is very close to the one in (5), this choice guarantees that $\mathrm{RMMR}^{\mathrm{mod}}_{\mathrm{avg}}$, defined below, lies in $[0, 1]$. Moreover, the max/min VaR in (2) can be naturally computed over $\mathcal{P}$ instead of over the larger $\mathcal{M}$. Therefore, the *modified* RMMR is[13]

$$\mathrm{RMMR}^{\mathrm{mod}} = \frac{\max_{\mathcal{P}} \ \mathrm{VaR} - \mathrm{VaR}^*}{\max_{\mathcal{P}} \ \mathrm{VaR} - \min_{\mathcal{P}} \ \mathrm{VaR}}. \tag{10}$$

As before, in this definition VaR^* may be: the VaR forecast under a fitted reference model M^*, kept fixed for the entire dataset; or, $\mathrm{VaR}_{\mathrm{best}}$ as defined above (notice that the best model necessarily lies in $\mathcal{P}$); or, $\mathrm{VaR}_{\mathrm{avg}}$ as defined in (9). The resulting measures are denoted respectively $\mathrm{RMMR}^{\mathrm{mod}}_{M^*}$, $\mathrm{RMMR}^{\mathrm{mod}}_{\mathrm{best}}$ and $\mathrm{RMMR}^{\mathrm{mod}}_{\mathrm{avg}}$.

As discussed in Reference 10, reducing the set of models does not necessarily yield to a decrease in model risk. In other words, for a given choice of VaR^* we may (and we actually do often) observe the inequality $\mathrm{RMMR}^{\mathrm{mod}} > \mathrm{RMMR}$. However, the use of weights to rule out badly fitting models should make the initial choice of the competing models less crucial for the RMMR value.

For any fixed model M^*, RMMR_{M^*} quantifies the (relative) model risk that we are prone to by fixing the reference model m^* and considering all alternative models as perfectly plausible. Instead, $\mathrm{RMMR}^{\mathrm{mod}}_{M^*}$ rules out the alternative models which provide a poor fitting of the data on a specific day, deemed implausible. Then, $\mathrm{RMMR}^{\mathrm{mod}}_{M^*}$ seems to provide a more sensible assessment of the actual model risk than RMMR_{M^*}. It must be noticed that when M^* turns itself to be implausible (i.e., $M^* \notin \mathcal{P}$), VaR_{M^*} may[14] not lie in the range $[\min_{\mathcal{P}} \mathrm{VaR}, \max_{\mathcal{P}} \mathrm{VaR}]$. As a consequence, $\mathrm{RMMR}^{\mathrm{mod}}_{M^*}$ does not need to be constrained to the interval $[0, 1]$; however, as we already commented, measures of model risk make perfect sense also in this case. Instead, it can be easily seen that $\mathrm{RMMR}_{\mathrm{best}}$, $\mathrm{RMMR}^{\mathrm{mod}}_{\mathrm{best}}$, $\mathrm{RMMR}_{\mathrm{avg}}$ and $\mathrm{RMMR}^{\mathrm{mod}}_{\mathrm{avg}}$ all lie in $[0, 1]$ by construction.

[12] A simpler approach, proposed by Reference 28 and going under the name of *thick modelling*, assigns equal weights to all plausible models.

[13] Note that when building $\mathcal{P}$ we have forced $k' \geq 2$, that is, we select at least 2 models, in order to prevent the denominator in (10) to vanish.

[14] Indeed, we observe this possibility in all our series.

TABLE 1 Descriptive statistics for logarithmic returns and price changes computed over the full sample.

Asset	Mean	Std. dev.	Min	Max	Skewness	Kurtosis
Log-returns						
DB	−0.000313	0.0243	−0.1807	0.2230	0.21	11.65
$/€	0.000038	0.0062	−0.0384	0.0461	0.09	5.55
DAX	0.000131	0.0153	−0.0887	0.1079	−0.02	7.62
OIL	0.000135	0.0212	−0.1696	0.1368	−0.10	7.01
GAS	0.000053	0.0364	−0.2627	0.4776	3.03	32.43
Price changes						
ELE	0.002372	11.26	−191.22	200.80	0.79	96.57

3 | DATA AND COMPETING MODELS

In the present paper we consider six variables: Deutsche Bank stock prices (henceforth: DB), USD/EUR exchange rate levels ($/€), German equity index level (DAX); Brent crude oil prices (OIL), ICE UK natural gas prices (GAS), the day–ahead auction prices for electricity observed on the European Energy Exchange for delivery in the German/Austrian zones (ELE). For each of the listed variables, let S_t be their level on a given day t. As described in the previous sections, we are interested in assessing model risk when forecasting VaR at level 1% or 5% for the variable $Y_{t+1} = S_{t+1} - S_t$.

For all variables, but electricity prices, we follow common practice and we specify competing models for the log-return $R_{t+1} = \log(S_{t+1}/S_t)$ between t and $t + 1$, so that

$$Y_{t+1} = S_t(\exp(R_{t+1}) - 1).$$

Since S_t is known at day t (and positive) and $h(x) = e^x - 1$ is strictly increasing, by known properties of VaR we have (at any level α)

$$\mathrm{VaR}(Y_{t+1}) = -S_t(\exp(-\mathrm{VaR}(R_{t+1})) - 1) \tag{11}$$

which directly links the VaR of the log-returns to the VaR of Y.

Instead, for electricity it is not possible to resort to log-returns, since prices may become negative: in this case, we directly model the price difference Y_{t+1}. The considered market zones (German/Austrian) are indeed characterized by a high penetration of renewable energy sources which has increased the complexity of the electricity price dynamics, given that wind (and solar to less extent) is highly variable and partially predictable. This results in nil or even negative prices, the frequency of which has increased over the years and across several markets.

All time series have been collected from Datastream, from 01/01/2001 to 31/12/2015, and are quoted on a basis of 5 days per week (weekends not included), for a total of 3914 observations. Descriptive statistics of log-returns and price differences (for electricity only) are presented in Table 1. In all considered series, there is clear evidence of asymmetry and fat tails, especially for natural gas.

Among *stylized facts* for financial asset returns, we also find confirmation that the variance series, measured by squared returns or squared price changes, displays positive correlation with its own past, hence giving support to a time–varying volatility dynamics. In addition, the unconditional distributions of returns and price changes are definitely not normal.[15] Furthermore, the inspection of empirical autocorrelation and partial autocorrelation functions, together with the time–structure and non-normality of the data, suggest us to select autoregressive processes for the conditional mean and a GARCH process for the conditional variance. For all series X_t, where either $X_t = R_t$ (log-returns, for the first five assets/indices) or $X_t = Y_t$ (price differences for electricity), we consider $K = 9$ competing models.

[15]Having more than 2000 but less than 5000 observations, we implemented the Shapiro–Francia test as in Reference 29, observing that the index values for all assets are far from being nil (for the null hypothesis of normality) even if the magnitude of these indices is decreasing over time, clearly indicating the departure from normality for all considered series.

TABLE 2 Number of parameters (p) in each model.

Model	GED	GHYP	JSU	NIG	NORM	SGED	SNORM	SSTD	STD
p	10	12	11	11	9	11	10	11	10

Specifically, each model is formulated as an AR(5)–GARCH(1,1), coupled with nine different parametric families of distributions for the (standardized) innovations. Explicitly, for a given variable, we posit

$$X_t = \mu_t + \sigma_t Z_t, \tag{12}$$

where

$$\mu_t = \overline{\mu} + \sum_{i=1}^{5} \phi_i X_{t-i} \tag{13}$$

is the conditional mean following an AR(5) process, and

$$\sigma_t^2 = \omega + \alpha(X_{t-1} - \mu_{t-1})^2 + \beta\sigma_{t-1}^2 \tag{14}$$

is the conditional variance following a GARCH(1,1) model; the parameters ω, α and β must satisfy known constraints. The innovation series (Z_t) are assumed IID with nine possible parametric standardized (i.e., mean 0, variance 1, whenever these moments are defined) distributions, which are[16]

1. the *Normal* distribution (NORM), with no additional parameters, other than location and scale;
2. the *Student-t* distribution (STD) with a tail parameter (degrees of freedom) in addition to location and scale;
3. the *Generalized Error Distribution* (GED), which generalizes the Normal distribution, with an additional tail parameter;
4. the skewed generalizations of the previous three distributions, defined as in Reference 32, which present an additional skew parameter: they are the *Skew Normal* (SNORM), the *Skew Student-t* (SSTD), and the *Skew GED* (SGED);
5. the *Johnson's* S_U (JSU) distributions, obtained as parametric non-linear transforms of a standard Normal, with two additional parameters;
6. the *Normal Inverse Gaussian* (NIG) which is a normal variance-mean mixture with the Inverse Gaussian as the mixing distribution, with two additional parameters;
7. the *Generalized Hyperbolic family* (GHYP) which generalizes the NIG distribution, adding an extra-parameter that gives further control on the tails.

Since within each family we are considering the standard representative, we are actually considering four types of distributions:

1. NORM, which has no additional parameters and therefore cannot account for asymmetries or fat tails;
2. STD and GED, which have a tail parameter controlling tail thickness, but no parameter for the skewness (hence, they are symmetric);
3. SNORM, which has a skew parameter, but not a tail parameter;
4. SSTD, SGED, JSU, NIG, and GHYP, which have both skew and tail parameters and with GHYP having actually two tail parameters.

These distributions display different complexity levels, from very simple (but still common) choices like NORM, to very flexible families, such as the JSU and GHYP. Therefore, we think that this set is large and diversified enough for our purposes of assessing model risk. We stress that the complete specification of each of the nine models includes both the parameters of the AR-GARCH part (9 in total: 6 for the AR part, 3 for the GARCH part) and the additional parameters of the (standard) distribution of the innovations. The total number of parameters in each model is summarized in Table 2.

[16]The exact densities of these distributions are described in Reference 30 and implemented in the `rugarch` R-package (see Reference 31).

4 | EMPIRICAL RESULTS

4.1 | The procedure

For any given asset/index, we estimate nine AR(5)–GARCH(1,1) models under the stated different distributional assumptions.

After one of the above distributions is chosen, the parameters of both the AR-GARCH process and of the innovations distribution are determined by Maximum Likelihood (ML)[17] on a daily basis, by adopting a rolling-window approach with a window size of $n = 256$ past observations (one year of data). Explicitly, for any series and for two considered VaR orders ($\alpha = 1\%$ or 5%) the step-by-step procedure can be summarised as follows:[18]

1. We start on day $t = 261$ considering an estimation window made of the last 256 observations, that is $\mathcal{W}_t = (X_{t-255}, \dots, X_t)$. Note that when $t = 261$, the window is $\mathcal{W}_{261} = (X_6, \dots, X_{261})$ as the first five observations have to be excluded for the estimation of the AR(5) part of the model.
2. For any of the nine choices of the innovations distributions, we jointly estimate, by ML over the estimation window $\mathcal{W}_t$, the parameters of the AR(5)-GARCH(1,1) model specified as in Equations (12), (13) and (14), together with the additional distributional parameters.
3. For each given model, we compute the VaR for X_{t+1}:

$$\mathrm{VaR}_\alpha(X_{t+1}) = -\widehat{\mu}_{t+1} + \widehat{\sigma}_{t+1}\mathrm{VaR}_\alpha(Z_{t+1}) \tag{15}$$

   where $\widehat{\mu}_{t+1}$ and $\widehat{\sigma}_{t+1}$ are obtained via (13) and (14), employing the estimated parameters, and $\mathrm{VaR}_\alpha(Z_{t+1})$ depends on the estimated parameters for the innovations distribution.[19] For electricity, this is already the final VaR forecast of the price difference; otherwise, we need to apply the transform (11). In both cases, for each asset/index and both $\alpha = 1\%$ and 5%, we end up with the (nine) forecasted VaR_k of the variable Y_{t+1}, with $k = 1, \dots, 9$.
4. We use the maximized log-likelihood $\hat{\ell}_k$ of each model M_k, retrieved at step 2, to compute BIC_k (or AIC_k) values, using (3) with $n = 256$ and p_k being the number of parameters in model M_k summarized in Table 2.
5. We select the (daily) best model, that is, the one with lowest BIC/AIC; recall we denote $\mathrm{VaR}_{\mathrm{best}}$ the corresponding VaR estimate.
6. We build the weights as in (7) and use them firstly to select the set of plausible models as in (8), and secondly to compute the $\mathrm{VaR}_{\mathrm{avg}}$ as in (9).
7. We evaluate various (modified) RMMR values as in (10), taking as reference VaR either VaR_{M^*}, for a fixed model $M^* \in \{M_1, \dots, M_9\}$, or $\mathrm{VaR}_{\mathrm{best}}$ or $\mathrm{VaR}_{\mathrm{avg}}$.
8. We increment the time t by one day and go back to step 2. Hence, at the second iteration, $t = 262$ and $\mathcal{W}_{262} = (X_7, \dots, X_{262})$ and so on.

Recalling that we have 3913 observations for our series X_t, the procedure outlined above yields various series for the RMMRs of length 3653 (i.e., $3913 - 261 + 1$) which allow us to inspect their dynamics through days and even over years.

To report some information about the estimation on the whole sample, we observe that the unconditional mean $\overline{\mu}$ is generally non-significant, as expected from the descriptive statistics of returns and price changes. Secondly, heteroskedasticity is an important feature to be included. On the contrary, evidence for skewness and kurtosis turns from significant to non-significant values, according to the considered distributions.

In the remaining of this section, we present the main empirical findings of our study: the details of the best/worst models, a comparison of weights built under the two schemes, and the analysis of the dynamics of the various Relative Measures of Model Risk.

[17] It is worth to recall that Reference 13, instead, first use Quasi Maximum Likelihood (i.e., they estimate the parameters of the conditional mean and variance as if innovations were normal), then estimate the remaining parameters on inferred innovations.

[18] All computations were executed using the R software and the `rugarch` package on a PC with an Intel(R) Core(TM) i7 processor. Computational times for each round of the procedure (i.e., steps 1 to 7) are in the order of few seconds.

[19] The Equation (15) relies on the known identity $\mathrm{VaR}(aX + b) = a\mathrm{VaR}(X) - b$ $(a > 0)$ applied to (12), noting that μ_{t+1} and σ_{t+1} are known at day t.

4.2 | The best and worst models

Concerning the fitting performance of employed models, Table 3 shows the number of days in which any of the nine distributions provided the lowest AIC/BIC values among all competing models for any asset/index. In Table 4, we indicate for any asset which model has proved to be the *best* or *worst* one over the entire period, that is, it has provided the lowest AIC/BIC (or the highest Maximized Likelihood, MaxL) for the highest or lowest number of days.

Looking at both tables, sometimes we observe marked differences between the rankings provided by the three criteria, even though the number of parameters p in the models seem quite close each other (ranging from 9 to 12). Indeed, MaxL just looks at the fitting ability, regardless of the complexity of the model, AIC imposes a soft penalty for over-parametrization, and BIC magnifies the AIC penalty with a factor close to 3 (since $\log 256 \approx 5.5$.) This is well exemplified by the rankings for Electricity (see Table 4): the best model according to MaxL is GHYP with $p = 12$ parameters; this number is reduced to 11 by AIC, which selects JSU as the best model, and further reduced to 10 by BIC, for which the best model is instead STD.

Looking at Table 3, we observe that for every asset there is no model clearly outperforming all other models. In particular, for each assets and criteria, the frequency with which the *best overall* model provides the lowest AIC/BIC remains well below 40%.

Looking at Table 4, we can interestingly observe that GHYP, although having a complex structure able to capture many features of financial variables, does not rank best according to MaxL, except for Gas and Electricity. When over-parametrization is taken into account (AIC/BIC), it even turns into the worst overall model. Not surprisingly the NORM model always ranks worst for MaxL; however, when parameters are considered via BIC, it remains the worst model only for Gas. Additionally, JSU and the relatively simple STD perform well for the three financial assets and Oil. More generally, we observe that the best/worst models for Oil tend to be more in line with those of the three financial assets than with those of the other two energy commodities.

TABLE 3 Number of days in which any given model provides the lowest AIC (and BIC, in parentheses) among all nine competing models and across assets.

	DB	$/€	DAX	OIL	GAS	ELE
GED	455 (594)	545 (697)	433 (574)	348 (456)	824 (1140)	372 (591)
GHYP	39 (14)	418 (197)	157 (90)	54 (9)	659 (501)	681 (504)
JSU	873 (639)	803 (742)	996 (762)	877 (636)	627 (559)	727 (633)
NIG	717 (577)	360 (324)	606 (501)	849 (645)	613 (521)	695 (542)
NORM	143 (359)	96 (180)	106 (272)	131 (345)	0 (4)	41 (124)
SGED	242 (152)	370 (241)	388 (282)	288 (193)	596 (497)	321 (230)
SNORM	113 (183)	555 (596)	246 (327)	180 (254)	10 (16)	49 (58)
SSTD	374 (198)	142 (109)	273 (191)	398 (267)	159 (131)	331 (278)
STD	692 (932)	359 (562)	441 (649)	523 (843)	160 (279)	431 (688)

TABLE 4 List of the *best* and the *worst overall models* according to maximized likelihood (MaxL), AIC and BIC.

	Best models			Worst models		
Assets	MaxL	AIC	BIC	MaxL	AIC	BIC
DB	JSU	JSU	STD	NORM	GHYP	GHYP
$/€	JSU	JSU	JSU	NORM	NORM	SSTD
DAX	JSU	JSU	JSU	NORM	NORM	GHYP
OIL	JSU	JSU	STD	NORM	GHYP	GHYP
GAS	GHYP	GED	GED	NORM	NORM	NORM
ELE	GHYP	JSU	STD	NORM	NORM	SNORM

TABLE 5 *Akaike weights* as in (6) versus weights as in (7) for nine different models specified for Gas on the 22 January 2002.

Model	Akaike weights	Weights as in (7)
GED	99.52	23.88
GHYP	0.00	0.00
JSU	0.00	13.24
NIG	0.08	16.49
NORM	0.00	0.00
SGED	0.00	12.39
SNORM	0.00	0.51
SSTD	0.36	17.90
STD	0.04	15.61

Note: Weights are in percentages.

These results clearly show the importance of accounting for the additional estimation effort when more parameters enter in the model. Furthermore, it is possible to understand even better the importance of the penalty term when comparing these findings with the ranking depicted by simply looking at the Maximized Likelihood, since this step affects the subsequent construction of weights and then the selection of models. Based on this evidence, we conclude that, in our analysis, the AIC measure is not able to properly penalize for over-parametrization: for this reason in what follows we just use BIC.

We have also inspected whether the best models, as selected via BIC, have a good predictive power as measured through a suitable scoring function; and, results are reported in Appendix B.

4.3 | Empirical analysis of weights

We have already highlighted that a potential drawback of Akaike weights as in (6) is that they tend to assume extreme values, close to zero or one. In order to provide empirical evidence of this aspect, we now make a comparison of the Akaike weights as in (6) with the weights we propose as formulated in (7), when BIC is used as Information Criterion.

In Table 5 we report the two sets of weights computed the 22 January 2002 for Gas.[20] We observe that eight models out of nine get almost vanishing Akaike weights, and just one model totals more than 99.5%; this behavior raises obvious doubts on the use of such weights in model trimming and averaging procedures. On the contrary, according to our alternative construction, just three models receive negligible weights, while the remaining six models are characterized by quite uniform weights, the highest one being just 24%.

Following the practice of assuming a threshold equal to 0.95, it is possible to identify the set of plausible models, which in the case of Akaike weights reduces to[21] $\mathcal{P} = \{\text{GED}, \text{SSTD}\}$. On the contrary, according to our alternative weights, we have

$$\mathcal{P} = \{\text{GED}, \text{JSU}, \text{NIG}, \text{SGED}, \text{SSTD}, \text{STD}\}.$$

We repeated the analysis on all other assets and days, and the empirical results on our series show that most of the time at least six models out of nine are deemed plausible according to our alternative weights. On the contrary, most of the time just two out of nine models are deemed plausible using the Akaike weights. Even taking a threshold higher than 0.95 does not seem to offer a practical solution when working with Akaike weights. Indeed, we find that a threshold guaranteeing at least five plausible models on most days would be (much) higher than 99.99%.

When working with our proposed weights, we have also considered two other thresholds (0.90 and 0.99), building the corresponding plausible sets $\mathcal{P}$. For each asset, it turns out that by moving the threshold from 0.95 to 0.90 we discard at

[20] A similar pattern is observed across all other assets/indices and dates in our dataset.

[21] And SSTD is included just because we enforce $\mathcal{P}$ to be made by at least two models.

TABLE 6 Standard deviations and maximum values (in percentages) of the nine weights series for DAX.

	Akaike weights		Weights as in (7)	
	Std. dev.	**Max**	**Std. dev.**	**Max**
GED	32.48	99.99	8.42	50.41
GHYP	0.29	16.34	3.86	24.22
JSU	33.17	99.99	9.61	78.71
NIG	27.33	99.99	8.61	64.31
NORM	22.40	99.99	7.47	35.91
SGED	23.47	99.99	8.22	63.91
SNORM	24.96	99.99	9.62	74.60
SSTD	17.56	99.99	6.93	48.65
STD	31.16	99.99	7.75	61.05

most a single model from $\mathcal{P}$ in nearly 90% of the dates. Likewise, by moving the threshold from 0.95 to 0.99 most of the days we add at most a single model to $\mathcal{P}$. We interpret these findings as an indication of robustness of our procedure with respect to the adopted threshold (0.95).[22]

To provide further comparison between the two weighting schemes, Table 6 presents the standard deviation and the maximum value on the entire dataset for all nine weights for DAX; results for other assets are similar hence omitted, but available on request.

For each given distribution, we see that the Akaike weights have higher standard deviation compared to the alternative ones. Furthermore, the maximum value of each Akaike weight over the entire sample is almost[23] 100% for eight out of nine models, while it is always less than 80% for the weights proposed in (7). As a consequence of these two considerations, we believe that the weights we propose represent a better choice for the analysis of model risk. Therefore, in what follows we will compute them according to our construction as in (7), using BIC as previously motivated.

In order to provide a more complete picture, we provide in Table 7 the average, the standard deviation, the minimum and the maximum values of the weights (as defined in (7)) computed over the entire period across assets and models. As expected and identified in Table 4, for each asset the best overall model is characterized by a high mean weight (though not necessarily the highest). Next, we can observe that the variability of weights does not seem very high, particularly for the DB stock and OIL. The minimum value that any weight can reach is (almost) 0, while the maximum value never exceeds 83%, as observed for the exchange rate.

4.4 | Historical dynamics of RMMR

Using the procedure outlined at the beginning of this section and employing weights computed as in (7) using BIC, we are able to produce several series for the Relative Measure of Model Risk (RMMR), for all studied assets/indexes. Specifically, our results refer to the following series, for VaR both at 1% and at 5%):

1. the *base* RMMR for a fixed reference model M^*, denoted RMMR_{M^*} and computed as in (2);
2. the *modified* RMMR in terms of the plausible set $\mathcal{P}$ (with threshold 0.95), computed as in (10) for
   - a reference model M^* fixed throughout the sample period, that is, with $\text{VaR}^* = \text{VaR}_{M^*}$; we denote the resulting measure $\text{RMMR}^{\text{mod}}_{M^*}$;
   - the daily best model, that is, with $\text{VaR}^* = \text{VaR}_{\text{best}}$ as defined in (4); we denote the resulting measure $\text{RMMR}^{\text{mod}}_{\text{best}}$;

[22] We also analyse the impact of moving the threshold from 0.90 to 0.99 directly on the RMMR values and, most of the days, we find the (relative) difference quite small.

[23] Precisely, it is higher than 99.9999% in all cases.

TABLE 7 Descriptive statistics (in percentages) of the weights as in (7) for any given asset and model.

		DB	$/€	DAX	OIL	GAS	ELE
GED	Mean	13.18	15.13	14.63	12.60	18.17	14.19
	Std. dev.	5.71	10.96	8.42	4.63	7.76	10.44
	Min	0.05	0.00	0.00	0.00	0.00	0.00
	Max	54.27	66.99	50.41	46.43	60.86	65.01
GHYP	Mean	0.35	5.27	1.84	0.30	2.55	10.84
	Std. dev.	0.99	7.31	3.86	0.87	2.57	9.33
	Min	0.00	0.00	0.00	0.00	0.00	0.00
	Max	6.96	32.27	24.23	6.29	16.50	59.98
JSU	Mean	14.61	13.10	14.22	14.59	15.58	15.08
	Std. dev.	6.43	14.05	9.61	5.51	6.20	9.62
	Min	0.00	0.00	0.00	0.00	0.00	0.00
	Max	53.22	82.83	78.71	35.74	58.34	79.64
NIG	Mean	14.50	7.84	11.45	14.62	14.77	14.48
	Std. dev.	6.51	10.14	8.61	5.93	6.61	9.62
	Min	0.00	0.00	0.00	0.00	0.00	0.00
	Max	46.80	81.11	64.31	33.34	60.24	81.17
NORM	Mean	9.09	7.01	7.80	9.49	1.19	4.30
	Std. dev.	6.26	7.31	7.47	6.62	3.23	7.21
	Min	0.00	0.00	0.00	0.00	0.00	0.00
	Max	39.00	43.55	35.91	45.29	33.78	62.61
SGED	Mean	11.79	13.65	13.80	11.49	16.93	12.41
	Std. dev.	5.18	10.48	8.22	4.24	6.40	8.75
	Min	0.00	0.00	0.00	0.00	0.00	0.00
	Max	52.90	63.33	63.91	43.75	54.46	62.51
SNORM	Mean	11.76	16.51	13.01	12.16	3.50	4.93
	Std. dev.	5.35	13.67	9.62	4.45	5.13	7.10
	Min	0.00	0.00	0.00	0.17	0.00	0.00
	Max	51.10	76.28	74.60	37.93	49.25	51.29
SSTD	Mean	11.73	9.76	11.01	11.99	13.07	11.55
	Std. dev.	5.29	7.43	6.93	5.04	5.13	8.09
	Min	0.00	0.00	0.00	0.00	0.00	0.00
	Max	34.97	45.99	48.65	35.33	51.02	67.21
STD	Mean	12.96	11.73	12.23	12.77	14.24	12.23
	Std. dev.	5.55	8.69	7.75	5.12	5.18	8.45
	Min	0.00	0.00	0.00	0.00	0.06	0.00
	Max	38.93	44.30	61.05	34.26	48.21	56.16

TABLE 8 Some descriptive statistics of the base RMMR (for $VaR_{1\%}$), for some choices of the reference models when considering the set of all nine competing models, $\mathcal{M}$.

Reference		DB	$/€	DAX	OIL	GAS	ELE
Normal	Mean	0.63	0.80	0.77	0.67	0.47	0.61
	Std. dev.	0.26	0.26	0.24	0.25	0.31	0.35
Best model	Mean	0.34	0.54	0.23	0.34	0.52	0.65
	Std. dev.	0.24	0.30	0.22	0.22	0.28	0.30
Worst model	Mean	0.99	0.43	0.98	0.99	0.96	0.65
	Std. dev.	0.04	0.22	0.06	0.03	0.12	0.30

– the (daily) weighted average of estimates, that is, with $VaR^* = VaR_{avg}$ as defined in (5); we denote the resulting measure $RMMR^{mod}_{avg}$.

In Table 8 we present the average values and standard deviations of the base RMMR for $VaR_{1\%}$, computed using the fixed model NORM, and the asset-specific overall best and worst models as reference models (with respect to BIC as summarized in Table 4).

We see that the average RMMR is lower when using the overall best model than when considering the overall worst model, for four out of six assets, the two exceptions being $/€ and electricity. It is important to observe that, in principle, a better performance of a model, as measured through its fitting ability *does not* necessarily mean a lower model risk. Indeed, recalling the definition of RMMR, it is possible to observe low model risk when the VaR forecast produced under the reference model is high in comparison to the forecast under the other models, which may not be the case even when the best fitting model is taken as the reference one. For Electricity, it is interesting to note that all three considered reference models are characterized by a comparable amount of model risk. On the contrary, we see that for some assets (notably, DB, DAX, and OIL) the average model risk for the best and worst models are substantially different. Turning to the standard deviation of RMMRs, we see that it is quite high and stable across assets for the best and the NORM model, and generally lower for the worst models.

Table 9 shows the descriptive statistics related to various modified RMMR (for $VaR_{1\%}$) as described before. First we consider $RMMR^{mod}_{M^*}$ for three fixed models M^* throughout the sample, that is, Normal and asset-specific overall best and worst models for BIC (as in Table 4). Then, we consider $RMMR^{mod}_{best}$ (daily best model) and $RMMR^{mod}_{avg}$ (weighted average).

We can observe that for all assets (but electricity), the average RMMR for the overall best model is much lower than the corresponding value for the worst model, with the latter being consistently greater than 1. Next, we see that the average RMMR for the daily best model is the lowest one and, as expected, the RMMR for the average estimate is always close to 0.50. Since the RMMR for the best and worst models are not constrained to the interval $[0, 1]$, we see a sharp increase of their volatility, particularly for the worst model.[24] On the contrary, the RMMR for the average estimate is stable around its mean value. Finally, by looking at the maximum value, we can notice that the RMMR for overall worst models can take, for all assets, values which are above 5 (even 16 in the $/€ case). Surprisingly high values, well beyond 1, are observed also when considering the overall best models: this means that just fixing a model at the onset, regardless of how this model is chosen, may carry a substantial amount of model risk. We notice that even when considering the daily best model, the RMMR value can sometime reach its upper bound 1.

To show the dynamics of these measures, Figure 1 shows the daily values of the *base* RMMR for $/€ throughout 2002, using JSU (overall best) as reference model. We can observe that in the first part of the year, the RMMR varies between values close to 0 and 1, whereas they tend to remain below 0.5 in the second half of the year. However, in general the series is volatile and this may mask the overall dynamics. A similar behaviour is observed for the RMMR series associated to other assets and reference models. Therefore, in order to catch the long-run patterns and make easier comparisons across models and assets, we continue our analysis by plotting rolling means computed on a 256-day basis.

To appreciate the dependence of the RMMR on the selected level of the Value-at-Risk (1% vs. 5%), we present in Figure 2 the dynamics of the two series of rolling means for Gas, using NORM as the reference model. We see that the difference is substantial and that model risk is higher for the lower value of α (1%). This is in line with the well-received

[24]This happens because the worst model seldom happens to be plausible.

TABLE 9 Some descriptive statistics of the modified RMMR (for $VaR_{1\%}$) for some reference models, when considering the restricted set of plausible models, $\mathcal{P}$.

Reference		DB	$/€	DAX	OIL	GAS	ELE
Normal	Mean	0.98	1.05	1.05	1.04	1.31	0.64
	Std. dev.	0.51	0.65	0.43	0.52	0.44	0.91
	Max	7.66	15.92	7.44	5.95	5.55	5.04
Overall best model	Mean	0.51	0.50	0.65	0.52	0.64	0.77
	Std. dev.	0.35	0.49	0.36	0.32	0.33	0.44
	Max	2.99	9.62	6.85	2.28	3.44	8.25
Overall worst model	Mean	1.60	1.37	1.38	1.59	1.31	0.73
	Std. dev.	0.50	0.67	0.44	0.46	0.44	0.50
	Max	7.75	16.16	7.33	5.99	5.55	9.41
Daily best model	Mean	0.28	0.27	0.24	0.31	0.20	0.29
	Std. dev.	0.29	0.33	0.29	0.30	0.28	0.34
	Max	1.00	1.00	1.00	1.00	1.00	1.00
Weighted average	Mean	0.46	0.44	0.46	0.48	0.49	0.44
	Std. dev.	0.11	0.14	0.12	0.11	0.13	0.14
	Max	0.81	0.91	0.92	0.87	0.88	0.95

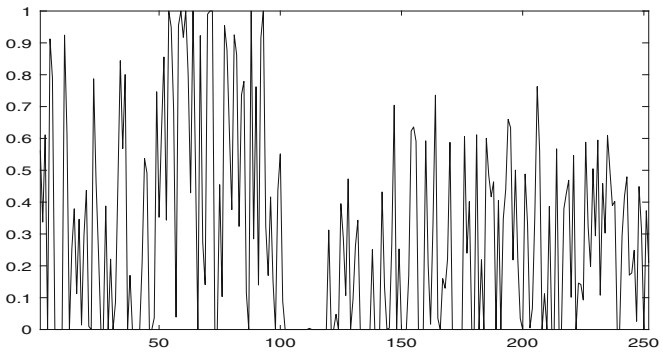

FIGURE 1 Dynamics of base RMMR for $VaR_{1\%}$ for $/€ over 2002, with JSU as reference model.

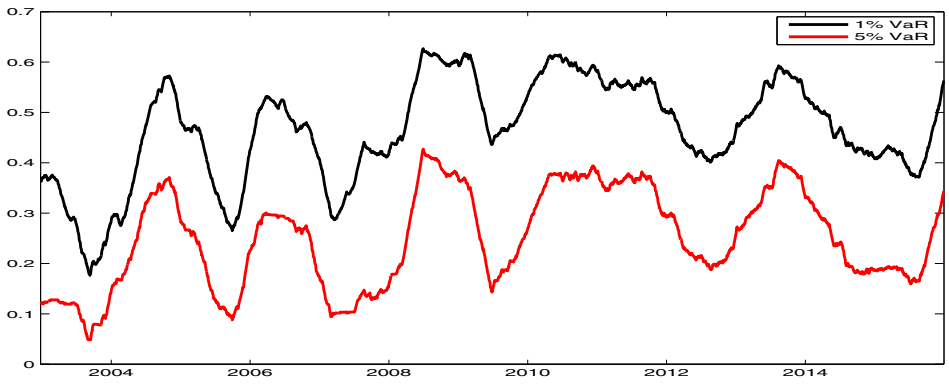

FIGURE 2 Dynamics of the (256-day rolling means) base $RMMR_{M^*}$ for Gas, using NORM as reference model and comparing $VaR_{1\%}$ versus $VaR_{5\%}$.

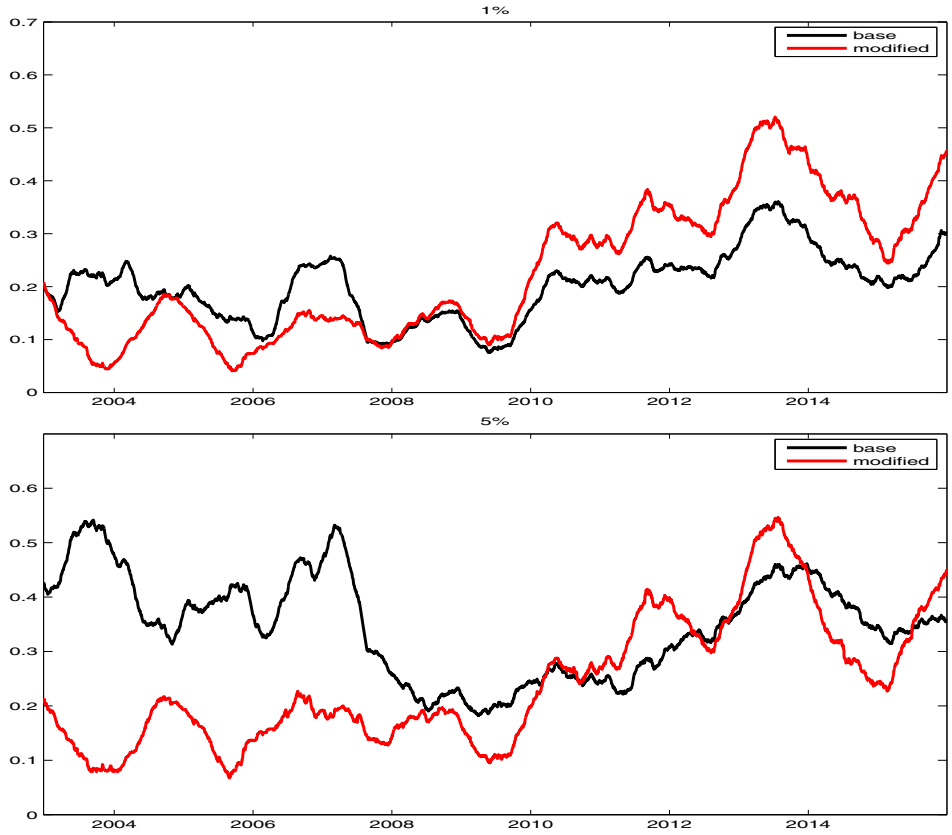

FIGURE 3 Dynamics of the (256-day rolling means) base and modified RMMR for Gas, using GED as reference model. Top panel: $\text{VaR}_{1\%}$; bottom panel: $\text{VaR}_{5\%}$.

fact that the Normal model tends to underestimate financial risks at low VaR levels (e.g., $\alpha = 1\%$), while it tends to overestimate risks at high levels (e.g., $\alpha = 5\%$).

In Figure 3 we instead compare for Gas the dynamics of the base versus the modified RMMR (i.e., RMMR_{M^*} vs. $\text{RMMR}^{\text{mod}}_{M^*}$) using the overall best model (GED) as the reference one, considering both $\text{VaR}_{1\%}$ and $\text{VaR}_{5\%}$. Although displaying roughly similar dynamics, the two quantities are sometimes quite apart. We can also notice that reducing the set to just plausible models (i.e., moving from RMMR to RMMR^{mod}) does not always yield a decrease in model risk. It seems that the modified RMMR increased after 2008, when the financial crisis started, and a possible explanation can be found in the progressive "financialization" of energy markets; see for instance.[33]

The analysis for other assets is reported in the Appendix A, whereas the main findings are summarized as follows: (i) the overall worst model provides by far the highest amount of RMMR for DB, DAX, OIL and GAS; (ii) the overall best and worst models provide comparable model risk for the \$/€ and electricity; (iii) most of the time, using the daily best as a reference model yields the lowest amount of model risk; (iv) using a Normal distribution clearly exposes to model risk; (v) selecting each day the currently best fitting model exposes us to less model risk than using an *a priori* fixed one, even though it is the overall best fitting model.

Finally, it may be rightly argued that a low model risk value in a period of high VaR may be as problematic as a high model risk in a period of low VaR. Therefore, in Figure 4 we compare for DAX the base RMMR for $\text{VaR}_{1\%}$ under the SSTD reference model, the normalized[25] $\text{VaR}_{1\%}$ estimates (under the same SSTD distribution), and the product between the RMMR and the normalized VaR forecast, over all days in 2003. We can see that around March-April this product is roughly at the same level as in August-September, even though in the former period the model risk is clearly lower (but VaR is higher). Therefore, we think that the product RMMR×VaR (or a similar quantity) is a meaningful indicator, providing precious additional insights.

[25]The VaR estimates are normalized by dividing them by their initial value (beginning of 2003), in order to ease the comparison.

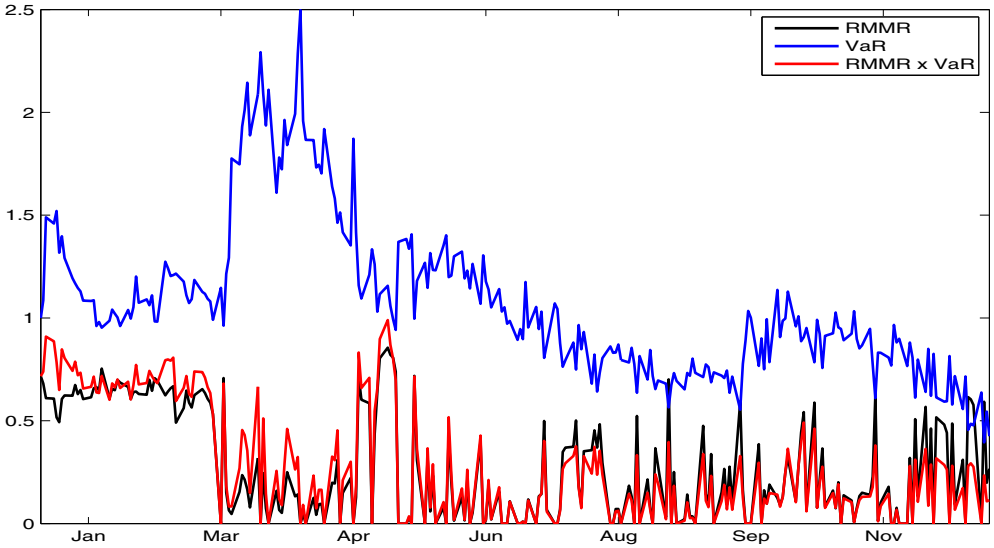

FIGURE 4 Dynamics during year 2003 of the base RMMR ($VaR_{1\%}$) for DAX, using SSTD as reference model, of (normalized) $VaR_{1\%}$ estimates under the SSTD model, and of the product between the RMMR and the normalized VaR estimate over year 2003.

4.5 | Simulated series

In order to check how our approach works in a controlled setting, we applied it to two simulated series.[26] In particular, we simulate two series of returns using an AR-GARCH specification with a fixed set of parameters[27] and employ two different distributions for the innovations: the NORM for the first series that we call "Series 1", and the JSU for the second series that we call "Series 2".[28]

As far as the first series simulated from a Normal is concerned, we include it in the initial set of competing models hence consider all nine distributions, as listed in Section 3. The aim of this part of the simulation is to inspect what may happen if the "true" data generating process is among the competing ones. Then, on a daily basis, we fit all nine models to the simulated series, we compute weights and identify the set of plausible models. Finally, we quantify the relative measure of model risk, considering as reference models: the Normal (hence, the "true model"), the overall best, worst[29] and the average VaR. Some numerical results are summarized in Tables 10 and 11. First and more importantly, the NORM model turns out to be "plausible" 92% of times, whereas when real data were considered the frequency for NORM was at most 50% and often much lower. Next, looking at Table 10, we see that the best overall model (according to BIC) is the Skew Normal distribution (SNORM), which nests the NORM model and is deemed plausible with a frequency of 96%. We also found that the mean weight for NORM and SNORM are 11% and 25%, respectively, which sum up to 36%: all these means are much higher than those observed for the real series of the six assets considered. Finally, looking at Table 11, we see that the values of model risk for different reference models are in line with those found for the six assets (reported in Table 9). Therefore, we can state that the proposed trimming procedure does not tend to discard the "true" data generating process, even if it is nested in a larger model (SNORM in this current case).

As far as the second series is concerned, since it has been simulated from a JSU, this has been excluded from the set of competing models. This part of the simulation aims at inspecting what may happen if the "true" data generating process is not among the competing ones. We chose the JSU as it is not included in, nor it includes, any other competing distributions and so it can be considered as quite "far" from the other models. Then, on a daily basis, we fit the remaining 8 models (thus excluding JSU) to the simulated series and we compute weights, plausible models and measures of model risk, considering the same reference models as for "Series 1": the overall best and worst (respectively, GED and NORM as

[26] We thank two referees for suggesting this extension of our investigation.

[27] Precisely, those obtained from fitting an AR(5)-GARCH(1,1) to gas series over the full sample.

[28] We used the software R and the package `rugarch` to run all simulations.

[29] According to BIC and being respectively, SNORM and GED; as it can be seen in Table 10.

TABLE 10 List of the *best* and *worst overall models* according to the maximized likelihood (MaxL) and BIC for the two series, simulated from NORM in the first case and from JSU in the second one.

	Best models		Worst models	
	MaxL	**BIC**	**MaxL**	**BIC**
Series 1	SNORM	SNORM	STD	GED
Series 2	SGED	GED	NORM	NORM

TABLE 11 Some descriptive statistics of the modified RMMR (for $VaR_{1\%}$) for some reference models for the two simulated series.

Reference		**Series 1**	**Series 2**
Normal	Mean	1.06	0.67
	Std. dev.	0.47	0.48
	Max	18.39	5.39
Overall best model	Mean	0.55	0.55
	Std. dev.	0.51	0.54
	Max	16.77	5.35
Overall worst model	Mean	1.03	0.67
	Std. dev.	0.45	0.48
	Max	17.06	5.39
Weighted average	Mean	0.49	0.41
	Std. dev.	0.15	0.11
	Max	0.89	0.87

shown in Table 10) and the average VaR. These numerical results are summarized in Tables 10 and 11. We can notice that the gap between the mean model risk under the worst and the best overall models is just 0.12, a value which is sensibly lower than the corresponding gap found for all other real series, except for electricity. This finding shows that when the set of competing models does not contain any model which is sufficiently close to the "true" one, there is no clear winner in the race for the best model. This means that all models perform nearly at the same level, both for what concerns the fitting ability and the model risk associated with their use.

5 | FINAL REMARKS AND CONCLUSIONS

In this paper, we provide an empirical assessment of model risk for both financial assets and energy commodities, implementing a modification of a measure of model risk first proposed in Reference 10. Under the well-established setting of GARCH models, but relaxing the assumption of normality, replaced by a wide range of alternative distributions for innovations, we are able to quantify model risk over a long period.

In our empirical analysis, we adopt a general and commonly used modelling framework (i.e., GARCH), kept fixed for simplicity. However, the procedure can be applied to a wide range of model designs, where the modelling is instead adapted to the industry business, the peculiarity of the market or the specificities of the assets considered.

We propose the construction of new weights different from, and more uniform than, those introduced by Akaike. Discarding the worst models, we restrict the set of "possible" models to a set of "plausible" ones on a daily basis. In this way we can obtain a more intrinsic and robust assessment of model risk by deriving a pure number that allows comparisons across markets, models and assets.

Moreover, by averaging out the forecasts, using the weights we build, we never significantly over- or under-estimate risk. Our empirical results emphasize that the distributional assumptions made in price modelling can produce a relevant discrepancy in risk forecasts and then trigger substantial model risk.

In general, we find that better models tend to produce less model risk. Although not completely surprising, this pattern is quite evident across different assets and levels of VaR. We also discover that, again not surprisingly, fixing a model at the onset as the reference one, even if this model is the best performing over the entire period, generally carries much more model risk than relying on models chosen day-by-day.

We also provide empirical evidence that the amount of model risk associated to a given model crucially depends on the estimated risk measure. In particular, under normality we observe more model risk at the more extreme levels for VaR. Instead, the model risk associated to the daily best distribution is more stable across VaR levels.

Finally, from a practical point of view, the proposed approach is timely and extremely relevant since the diffusion of artificial intelligence, machine learning and big data are expanding the set of models at a fast pace. Hence, it can provide support to the decision-making processes reducing the exposition to model risk and its potential losses.

We elaborate a bit further on this point. From a practical point of view, the procedure delineated in this paper can be used to answer the following question: *given the analysis performed on the behaviour of the RMMR measure for a set of competing models, which one should be used today to predict VaR for tomorrow?* Indeed, let us first recall that the RMMR value describes the relative position of a forecast with respect to the risk values provided by a set of competing but all "plausible" models. As such, values of RMMR that are above (respectively below) 0.5 signal a possible underestimation (overestimation) of risk. Both scenarios are undesirable, but underestimation can have more serious consequences, and financial regulations acknowledge this fact. Therefore, the whole series of RMMR values, computed daily for each competing model/forecast, can be used to "rank" the models. In practice, one can consider a "scoring function" S such that $S(0.5) = 0$, S is strictly decreasing for $x \leq 0.5$ and strictly increasing for $x \geq 0.5$, and $S(x) > S(1-x)$ holds for $x > 0.5$. Note that the last requirement translates the fact that, given the same departure from 0.5, underestimation ($x > 0.5$) has to be more penalized than overestimation. An example of such a function is $S(x) = (x-0.5)^2$ for $x \leq 0.5$ and $S(x) = k(x-0.5)^2$ for $x \geq 0.5$, with a fixed $k > 1$. On daily basis and for each model, S is evaluated at the RMMR value. Then, the average score for each model can be computed on a chosen window and the model having the lowest average score can be considered as the best one as far as model risk is concerned.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from Refinitiv. Restrictions apply to the availability of these data, which were used under license for this study. Data are available from the author(s) with the permission of Refinitiv.

ORCID

Angelica Gianfreda https://orcid.org/0000-0002-4350-874X

REFERENCES

1. Beder TS. VaR: seductive but dangerous. *Financ Anal J*. 1995;51:12-24.
2. De Schepper A, Heijnen B. How to estimate the value-at-risk under incomplete information. *J Comput Appl Math*. 2010;233(9): 2213-2226.
3. Bernard C, Vanduffel S. A new approach to assessing model risk in high dimensions. *J Bank Financ*. 2015;58:166-178.
4. Jorion P. Risk^2: measuring the risk in value-at-risk. *Financ Anal J*. 1996;52(6):47-56.
5. Christoffersen P, Goncalves S. Estimation risk in financial risk management. *J Risk*. 2005;7:1-28.
6. Gourieroux C, Zakoïan J-M. Estimation-adjusted VaR. *Economet Theor*. 2013;29(4):735-770.
7. Bignozzi V, Tsanakas A. Parameter uncertainty and residual estimation risk. *J Risk Insur*. 2016;83(4):949-978.
8. Kerkhof J, Melenberg B, Schumacher H. Model risk and capital reserves. *J Bank Financ*. 2010;34(1):267-279.
9. Boucher CM, Daníelsson J, Kouontchou PS, Maillet BB. Risk models-at-risk. *J Bank Financ*. 2014;44:72-92.
10. Barrieu P, Scandolo G. Assessing financial model risk. *Eur J Oper Res*. 2015;242(2):546-556.
11. Daníelsson J, James KR, Valenzuela M, Zer I. Model risk of risk models. *J Financ Stab*. 2016;23:79-91.
12. Claeskens G, Hjort NL. *Model Selection and Model Averaging*. Cambridge University Press; 2008.
13. Kellner R, Rösch D. Quantifying market risk with value-at-risk or expected shortfall? Consequences for capital requirements and model risk. *J Econ Dyn Control*. 2016;68:45-63.

14. Bannör K, Kiesel R, Nazarova A, Scherer M. Parametric model risk and power plant valuation. *Energy Econ*. 2016;59:423-434.
15. Truong C, Trück S, Mathew S. Managing risks from climate impacted hazards – the value of investment flexibility under uncertainty. *Eur J Oper Res*. 2018;269(1):132-145.
16. Mittnik S, Paolella MS. Conditional density and value-at-risk prediction of Asian currency exchange rates. *J Forecast*. 2000;19(4): 313-333.
17. Fan Y, Zhang Y-J, Tsai H-T, Wei Y-M. Estimating value-at-risk of crude oil price and its spillover effect using the GED-GARCH approach. *Energy Econ*. 2008;30(6):3156-3171.
18. Grigoletto M, Lisi F. Looking for skewness in financial time series. *Econometr J*. 2009;12(2):310-323.
19. Giot P, Laurent S. Market risk in commodity markets: a VaR approach. *Energy Econ*. 2003;25(5):435-457.
20. Hung J-C, Lee M-C, Liu H-C. Estimation of value-at-risk for energy commodities via fat-tailed GARCH models. *Energy Econ*. 2008;30(3):1173-1191.
21. Akaike H. Information theory and an extension of the maximum likelihood principle. Paper presented at: 2nd International Symposium on Information Theory, Akademiai Kiado Budapest. 1973.
22. Schwarz G. Estimating the dimension of a model. *Ann Stat*. 1978;6(2):461-464.
23. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer-Verlag; 2002.
24. Winkler RL, Makridakis S. The combination of forecasts. *J R Stat Soc: Ser A*. 1983;146(2):150-157.
25. Akaike H. Information measures and model selection. *Bull Int Stat Inst*. 1983;44:277-291.
26. Miljkovic T, Grün B. Using model averaging to determine suitable risk measure estimates. *North Am Actuar J*. 2021;25(4): 562-579.
27. Hansen PR, Lunde A, Nason JM. The model confidence set. *Econometrica*. 2011;79(2):453-497.
28. Granger CW, Jeon Y. Thick modeling. *Econ Model*. 2004;21(2):323-343.
29. Shapiro SS, Francia RS. An approximate analysis of variance test for normality. *J Am Stat Assoc*. 1972;67(337):215-216.
30. Rigby R, Stasinopoulos D. Generalized additive models for location, scale and shape. *J R Stat Soc: Ser C*. 2005;54:507-554.
31. Ghalanos A. Rugarch: univariate GARCH models. *R Package Version*. 2022;1:4-9.
32. Fernandez C, Steel MFJ. On Bayesian modeling of fat tails and skewness. *J Am Stat Assoc*. 1998;93(441):359-371.
33. Henderson BJ, Pearson ND, Wang L. New evidence on the financialization of commodity markets. *Rev Financ Stud*. 2015;28: 1285-1311.
34. Gianfreda A, Bunn D. A stochastic latent moment model for electricity price formation. *Oper Res*. 2018;66(5):1189-1203.

How to cite this article: Gianfreda A, Scandolo G. Assessing model risk in financial and energy markets using dynamic conditional VaRs. *Appl Stochastic Models Bus Ind*. 2023;1-26. doi: 10.1002/asmb.2828

APPENDIX A. FURTHER HISTORICAL DYNAMICS OF RMMR

Considering Value-at-Risk at level 1%, Figures A1 and A2 show, for each asset/index, the dynamics of the modified RMMR using three possible reference distributions: the overall best and worst models (with respect to BIC as in Table 4) and the daily best model, as defined on a daily basis by (4).

First, we notice that for DB, DAX, OIL and GAS the overall worst model provides by far the highest amount of RMMR, consistently above 1. On the contrary, for the \$/€ and Electricity series the overall best and worst models provide comparable model risk, with the RMMR value for the former often close to or even above 1. Most of the time, using the daily best as a reference model yields the lowest amount of model risk. We observe that there is no clear trend in none of the RMMR series, except maybe for Gas, where the dynamics for RMMR associated to overall worst and best models display a weak upward trend after 2008.

Figure A3 provides a comparison of the modified RMMRs across all six assets/indices, for both Value-at-Risk at level 1% and 5%. Three different choices for the reference distribution are made: the fixed NORM distribution, the daily best model for each asset and, finally, the overall best model for each asset with respect to BIC, as in Table 4.

Looking at the graphs in the top row, we see that using a normal distribution clearly exposes to model risk, with average values close to 1, particularly when $\alpha = 1\%$. Looking at the two graphs on the middle row, the model risk carried by using the daily best model is, as expected, quite moderate and stable, even if it sometimes displays some peaks, particularly for energy commodities. The two bottom graphs show that the model risk associated to the overall best model can be

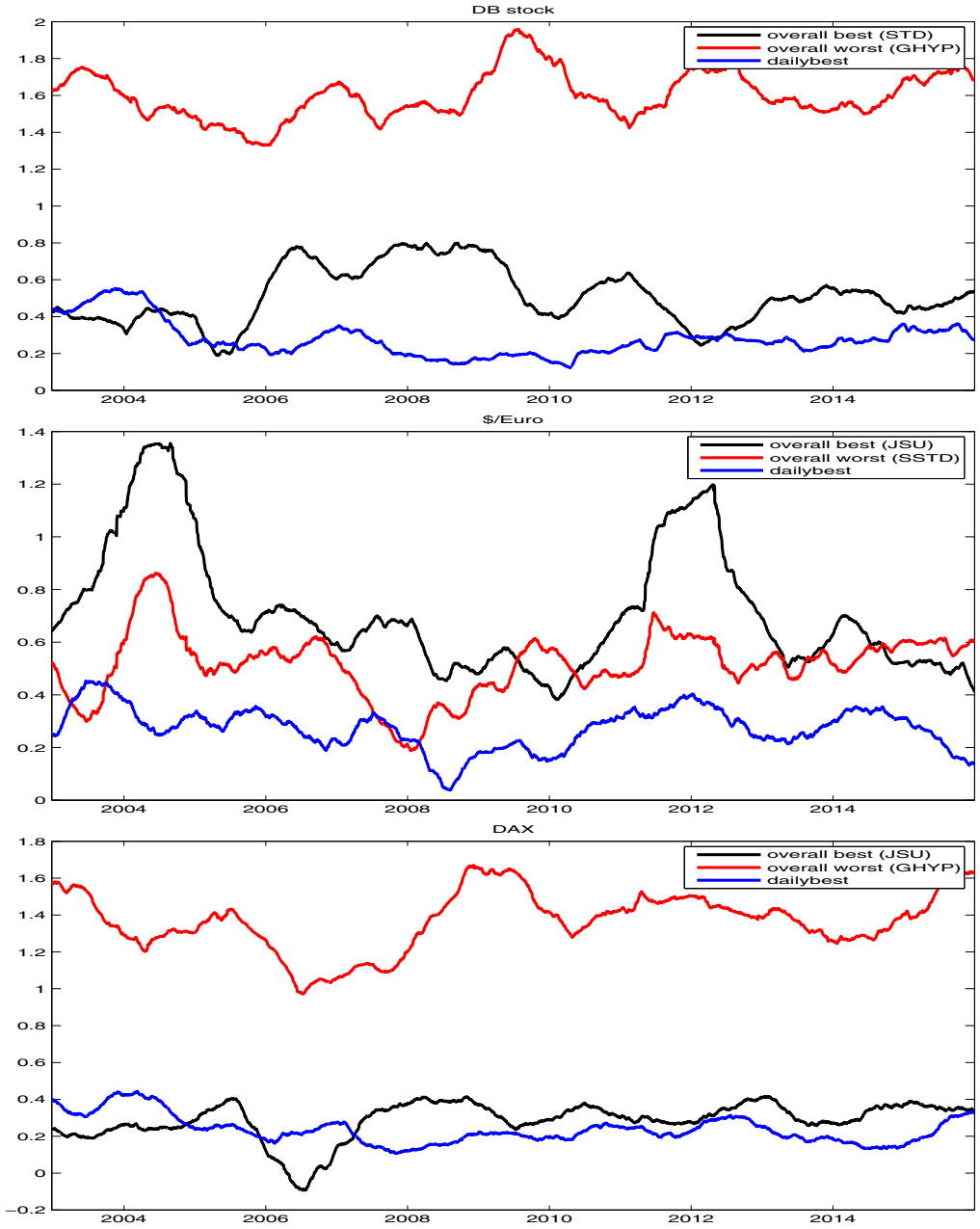

FIGURE A1 Dynamics of the (256-day rolling means) modified RMMR for $\text{VaR}_{1\%}$, using as reference models the overall best, the overall worst, and the daily best models. Top panel: DB; middle panel: \$/€; bottom panel: DAX.

substantial, particularly for Electricity, and when $\alpha = 5\%$. We can also notice that the model risk at $\alpha = 5\%$ for \$/€ turns from very high values (around 1.7) to very low and even negative values.

Looking specifically at Electricity in Figure A4, we provide a graphical comparison of the modified RMMRs (with VaR at level 1% and 5%) using the overall best model (STD), the daily best model and the weighted average forecast. We can see that the overall best model produces the highest level of model risk, whereas the daily best model results in the lowest level. This is not surprising given that the nature of these electricity prices changed dramatically with the progressively increasing renewable generation, which acted as shape-shifters of the price densities, see Reference 34. In the middle, the daily weighted average forecast case is characterized by a quite stable level of model risk, which mildly fluctuates around 0.5: this comes at no surprise, given the definition of $\text{RMMR}^{\text{mod}}_{\text{avg}}$. Similar results are recovered when looking at the other assets, in line with the summary statistics in Table 9. We can conclude that if we change the model, selecting each day the currently best fitting model, exposes us to less model risk than using an *a priori* fixed one, even

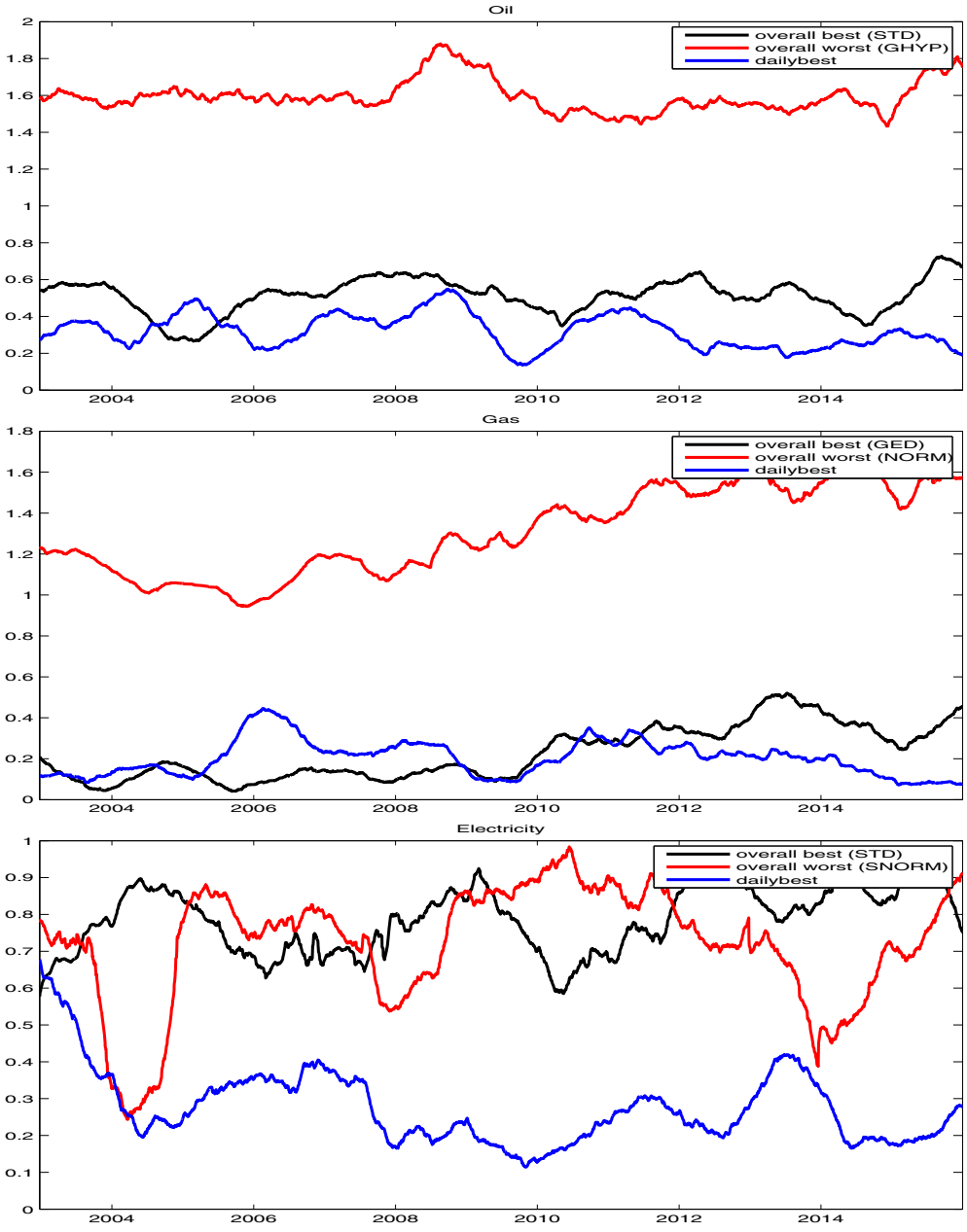

FIGURE A2 Dynamics of the (256-day rolling means) modified RMMR for $\text{VaR}_{1\%}$, using as reference models the overall best, the overall worst, and the daily best models. Top panel: Oil; middle panel: Gas; bottom panel: Electricity.

though it is the overall best fitting model. Instead, averaging out *plausible* estimates has a double effect: it smooths the dynamics of the RMMR and it additionally brings its level toward 0.5 and this means that we never significantly over- or under-estimate risk.

APPENDIX B. BEST MODELS VIA SCORING FUNCTIONS

We assessed the predictive ability of the nine models for each of the six assets using a suitable scoring function. Precisely, as our goal is quantile forecasting at level α, we employed the so-called *pinball loss function* defined as

$$S_\alpha(x, y) = (\mathrm{I}(x \geq y) - \alpha)(x - y),$$

FIGURE A3 Dynamics, across all six assets/indices, of the (256-day rolling means) modified RMMRs for $VaR_{1\%}$ (left) and $VaR_{5\%}$ (right), using as reference models: NORM (top row), daily best model (middle row), overall best model (bottom row).

where I is the indicator function. We computed, for each model and asset, the average scoring over the entire dataset, that is,

$$\overline{S} = \frac{1}{n}\sum_{i=1}^{n} S_\alpha(x_i, y_i)$$

where n is the number of forecasts, x_i is the quantile forecast (under a given model) issued at day $i-1$ for the variable on day i and, finally, y_i is the series observation on day i. Good predictive ability is associated with low values of $\overline{S}$.

Numerical results for the level $\alpha = 1\%$ are collected in Table B1, while in Table B2 we report, for each asset, the model having the lowest average scoring for $\alpha = 1\%$ and $\alpha = 5\%$. For comparison, we display also the best model according

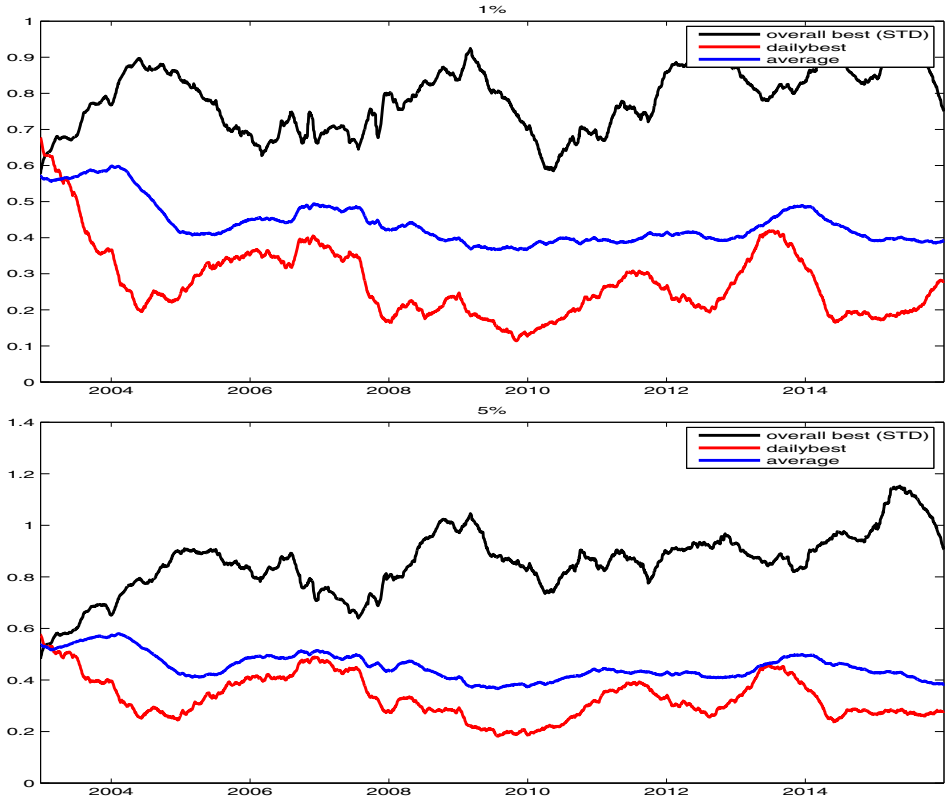

FIGURE A4 Dynamics of the modified RMMR (averaged over a 256-day rolling window) for electricity, using the overall best model (STD), the daily best model and the average model. Values for $VaR_{1\%}$ (top panel) and for $VaR_{5\%}$ (bottom panel) are reported.

TABLE B1 Average pinball for VaR at 1%.

	DAX	ELE	$/€	GAS	OIL	DB
GED	0.376	0.217	0.174	0.625	0.490	0.499
GHYP	0.421	0.240	0.191	0.791	0.537	0.562
JSU	0.300	0.213	0.142	0.736	0.420	0.460
NIG	0.331	0.219	0.151	0.734	0.398	0.432
NORM	0.463	0.235	0.205	0.697	0.631	0.606
SGED	0.358	0.207	0.170	0.653	0.488	0.496
SNORM	0.381	0.237	0.170	0.804	0.505	0.528
SSTD	0.344	0.231	0.158	0.750	0.444	0.458
STD	0.361	0.235	0.158	0.732	0.463	0.473

Note: Values are multiplied by 1000 (but for ELE) to ease comparisons.

TABLE B2 List of the models having the lowest average scoring for two values of α and best models according to BIC.

Assets	$\alpha = 1\%$	$\alpha = 5\%$	BIC
DB	NIG	SNORM	STD
$/€	JSU	SNORM	JSU
DAX	JSU	SGED	JSU
OIL	NIG	SNORM	STD
GAS	GED	SNORM	GED
ELE	SGED	SNORM	STD

to BIC, as already presented in Table 4. We can see that, as far as the DAX, \$/€, GAS, and OIL series are concerned, models that provide superior in-sample fitting ability (as measured by BIC) tend to perform well also out-of-sample, providing comparatively low mean values of the pinball loss, at least for $\alpha = 1\%$. Instead, the evidence for ELE and DB is mixed.