MOODY'S ANALYTICS

B&H RESEARCH

APRIL 2013

Validation of risk factor modelling in 1-year VaR capital assessments

Gavin Con FFA
Harry Hibbert
Sandy Sharp FIA
Craig Turnbull FIA
Moody's Analytics Research

Contact Us

Americas
+1.212.553.1658
clientservices@moodys.com

Europe
+44.20.7772.5454
clientservices.emea@moodys.com

Asia (Excluding Japan)
+85 2 2916 1121
clientservices.asia@moodys.co

Japan
+81 3 5408 4100
clientservices.japan@moodys.com

Overview

Principle-based capital assessment requires rigorous and robust validation of firms' internal models. This paper discusses the validation of the risk factor element of firms' internal models, particularly in the context of 1-year VaR capital models. Examples of back-testing, sensitivity testing and stress and scenario testing of equity, interest rate and credit models are developed to illustrate the challenges that a rigorous validation process must address. The inherent limitation of the quantity and quality of historical data in validating model performance deep in the tail of the probability distribution, and the consequent use of subjective assumptions that is required to fill that vacuum is a recurring theme in model validation. Differentiating between evidence-based science, expert judgment and Knightian uncertainty is important to insightful validation, but is challenging when deep in the tails of risk factor distributions.

Contents

1. Introduction

A principle-based approach to risk and capital assessment gives financial institutions the freedom to tailor their capital assessment methodology to best capture the specific risk profile of the firm. Rigorous validation of the risk methodology is a critical check that this freedom is being exercised responsibly. More generally, the validation process helps firms and their regulators to develop a better understanding of the limitations that are present in the implemented risk assessment methodology. This can provide important feedback into the ongoing development of the model and calibration methodology.

This report discusses the validation approaches that can be applied to the risk factor modelling element of firms' principle-based risk and capital assessments. The discussion focuses on 1-year 99.5% Value-at-Risk assessment, but much of it is relevant to risk factor modelling validation more generally, both in the shorter-term horizon modelling of banks and the longer-term horizon modelling used in the run-off approaches to principle-based statutory reserving that are currently prevalent in North America.

Risk factor modelling is arguably the most challenging and important quantitative element of a principle-based risk assessment. There is an inherent limitation on the volume of relevant historical data that can be used to make forward-looking statements about the severity of tail events. Expert judgement must fill the vacuum created by this deficiency. As a result, well-intended risk factor model and calibration processes produce a wide range of feasible conclusions about the severity of events that occur in a 99.5th percentile 1-year scenario, and the range of feasible assumptions can often imply materially different conclusions for the assessed level of capital requirement. This presents a significant challenge for firms and their regulators. It can also create an incentive to abuse the freedom provided by a principle-based system: The recent reporting[1] of practices at major banks provides a timely illustration of where a firm might systematically understate risk and capital requirements by making use of the in the ambiguity in the risk factor calibration method they select.

This paper discusses the range of validation methods that can be applied to risk factor modelling and calibration and the inherent challenges that need to be addressed along the way. We believe a deep understanding of these issues is vital to the robust validation of principle-based capital assessments.

2. The key elements of risk factor model validation

The importance of validation in principle-based risk assessment is well-recognised. The International Association of Insurance Supervisors' Common Framework views validation as a key component of any internal model approach to capital assessment [2]. In the UK the newly-formed Prudential Regulation Authority (PRA) has outlined its supervisory approach which states that internal models should be supported by adequate testing and justification of the model on an ongoing basis[3]. Similarly, it figures prominently in the Solvency II Internal Model framework[4] . There are three key elements that are generally recognised in risk factor validation. These are summarised in the table below.

[1] "Basel eyes set periods for banks' risk models", 17 February 2013, Financial Times.

[2] See IAIS' ICP17 Capital Adequacy (October 2010).

[3] See The Prudential Regulation Authority's approach to insurance supervision (April 2013)

[4] Article 122 (Validation Standards) of the Solvency II directive outlines the requirement to have an effective process for validating the internal model

THE KEY ELEMENTS OF RISK FACTOR MODEL VALIDATION

1.Back-testing	**Testing the model against historical experience** Comparison of risk estimates produced by the model methodology against actual realised outcomes.
2. Sensitivity Testing	**Testing the robustness of the model assumptions** What is the feasible range of assumptions that could be made under a reasonable risk methodology? What is the impact of these variations in assumptions on the modelling conclusions?
3. Stress and Scenario Testing	**Assess the impact of a single event (stress test) or combination of events (scenario test) and compare with model's measures of risk** The sources of the stresses and scenarios can either be specific historical events (1927 crash, etc.) or stresses that incorporate views from economists and risk experts on forward-looking downside risk. scenarios.(disorganised Euro defaults, etc).

The main goal of this paper is to present practical examples of how risk factor model validation can be approached and to outline the challenges associated with this exercise. Validation examples are developed using 1-year real-world models and calibrations provided by the B&H Economic Scenario Generator (ESG). (For the avoidance of doubt, in this context we are *not* using the ESG for market-consistent valuation. An ESG can also be used to value liabilities and construct the market-consistent balance sheet and, of course, this must also be subject to validation, but this is not the focus of this paper.) Whilst the examples in this paper are focused on market risks, most of the techniques and discussion set out in this paper are equally applicable to the validation of non-market risk factors.

3. Back-testing

3.1 Back-testing percentile estimates

In back-testing, the risk factor model and calibration methodology is implemented at a number of historical dates. The risk estimates produced by the model at each of these historical dates are then compared with the actual outcomes that arose over the risk horizon of the model. Formal statistical tests can, at least in theory, be applied to the comparison of the model estimates and to the actual risk outcomes in order to validate that the model is adequately estimating the riskiness of the variable(s) in question. For example, in the context of 1-year 99.5% VaR, the model methodology could be applied to 1000 historical dates to produce 1000 estimates of the 99.5th percentile of the 1-year S&P 500 return. If the model is performing adequately, we would expect 5 of the 1000 observed S&P 500 returns to be more severe than the 99.5th percentile estimate produced by the model for the corresponding period.

This simple example above highlights some of the limitations that inevitably arise in back-testing. The statistical analysis of back-testing results requires each back-test to be independent of the others. This means that the risk projection horizons of each back-test must be non-overlapping. In the context of a 1-year VaR, the dates that the back-tests are applied must therefore be at least one year apart[5]. For most of the variables that are modelled as risk factors, there is likely to be at very most a few hundred years of data. The risk calibration methodology might also use variables that have only been observable for a few decades or less (e.g. option-implied equity volatilities or central bank inflation forecasts, etc.)[6]. More generally, there is the issue of whether the methodology that has been developed for today's environment would be applicable in the economic system of one or two hundred years ago.

With, say, 300 non-overlapping back-tests, the number of 'breaches' of the percentile estimate would have a binomial distribution with $n = 300$ and $p = 0.005$, assuming the model was indeed providing accurate estimates of the 99.5th percentile of the risk factor distribution. The probability distribution for the number of 'breaches' observed in this case is shown below in Figure 1.

[5] More over-lapping periods could be produced by 'converting' the one-year risk model into a risk model for a shorter projection horizon and applying the back-tests to that model. We believe this is generally unlikely to be useful approach, especially in tail modelling using non-normal distributions. Whilst it **might** be reasonable to apply, say, a linear variance assumption in converting from monthly to annual volatilities, the correspondence between the monthly and annual tail percentiles will be more complex.

[6] To elaborate on the back-testing approach, the process does not involve merely taking the most up-to-date "1-in-200" stress and applying this stress at historic dates and comparing against the actual outcome. The approach should apply, to the full extent possible, the internal model's risk factor model methodology at the historic date. The methodology underlying the internal model is applied at historic dates and a distribution produced. This means that the distribution and stresses produced are relevant to the economic conditions at the historic date. To take a simple example, in periods of higher interest rates, the absolute size of the yield curve stresses produced by the risk methodology may be higher than when interest rates are low. In such cases, this should be incorporated into the results produced by the risk model in the back-tests.

Figure 1 Probability distribution for 'breaches' of 99.5th percentile estimate; 300 back-tests

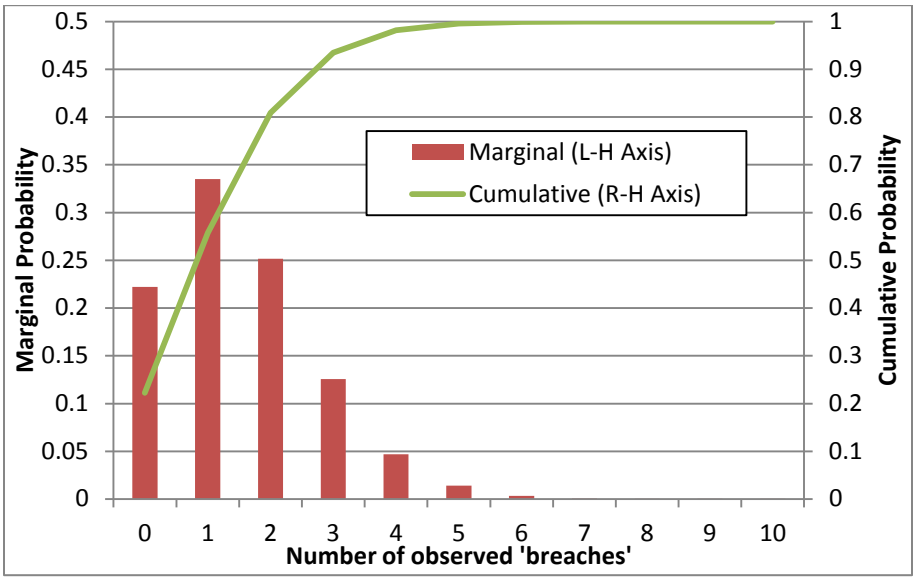


Figure 1 highlights the statistical limitations of back-testing with the limited volume of back-tests that will typically be available to a firm calculating 1-year VaR. Even with 300 years of back-testing and a perfectly-functioning risk model, there is a greater than 20% probability that no 'breaches' of the model's 99.5% percentile estimates will be observed. So this particular statistical test cannot really tell us anything about the likelihood of the model *over-estimating* risk. But it may be able to tell us something about the likelihood that the model is *under-estimating* risk—to reject with 95% confidence the null hypothesis that the model is producing 'breaches' with 0.5% probability, we would need to observe at least 5 instances where the observed annual return was more severe than the model's estimated 99.5th percentile (the probability of observing 5 or more instances is less than 2% under the null hypothesis). Whilst this is useful statistical information, such long-dated back-testing will often create the temptation to explain away such instances as historical anomalies that cannot occur again and that the risk methodology is not designed to capture.

There is a further technical complication with statistical analysis of back-testing – it only makes sense if the risk methodology is intended to be Point-in-Time (PIT). Most insurance firms' risk factor models are intended to be Through-the-Cycle (TTC). In this case, the model's risk estimates may often misestimate the actual severity of the risk over the risk projection horizon *by design*. This makes interpretation of back-testing results even more difficult – if, in the above example, we find that actual S&P returns were worse than the model's estimate of the 99.5th percentile in 30 of the 300 cases, it might be argued that in 29 of those 30 cases one-year market volatility exceeded the TTC volatility, and that this therefore is not evidence of failure of the risk model as it is not attempting to capture volatility in excess of the TTC level.

3.2 Back-testing distributional estimates

The above discussion focused on the back-testing of the risk model's performance in estimating a *specific percentile* of the variable in question. In the context of a VaR capital assessment, this is a natural focal point of the validation. However, the back-testing process could also attempt to validate the model's ability to estimate the *probability distribution* of the variable in question. Internal models are generally intended to produce whole probability distribution for capital, so this seems a reasonable approach to validation. It *may* also be able to provide more statistically powerful information than the percentile validation. For instance, we saw in section 3.1 that the binomial distribution hypothesis test could not tell us anything about the likelihood of the model over-estimating risk. If we compare the actual variable outcomes with the estimated probability distribution and find that none of the 300 outcomes were more severe than the estimated 70th percentile, this would provide strong evidence the model is significantly over-estimating risk.

This back-testing approach would again fully apply the risk model methodology at historical dates and compare model output with the actual realised outcomes for the variable in question. Now armed with the full probability distributions produced by the risk model in each back-test, more validation analysis is possible. At the simplest level, the binomial distribution hypothesis testing described in section 3.1 could be re-run for a wide array of percentiles of the distribution rather than only for the 99.5th percentile. This can provide additional statistical information about the performance of the risk model in different parts of the variable's

probability distribution, potentially highlighting where the model performs well and where it has material limitations. This approach can be taken irrespective of the form of probability distribution produced by the risk model for the variable in question.

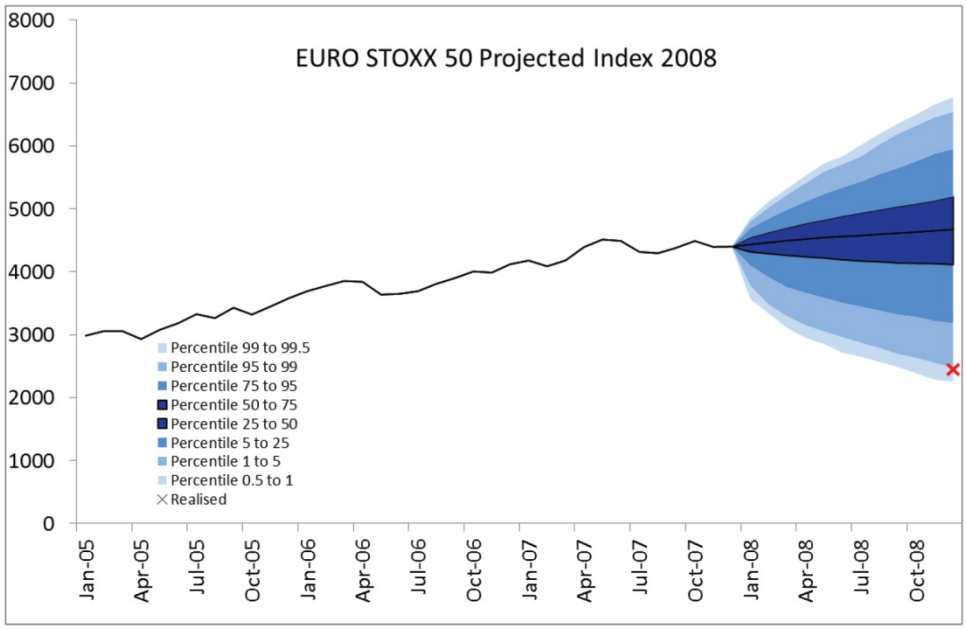
3.3 Back-testing examples: equity returns

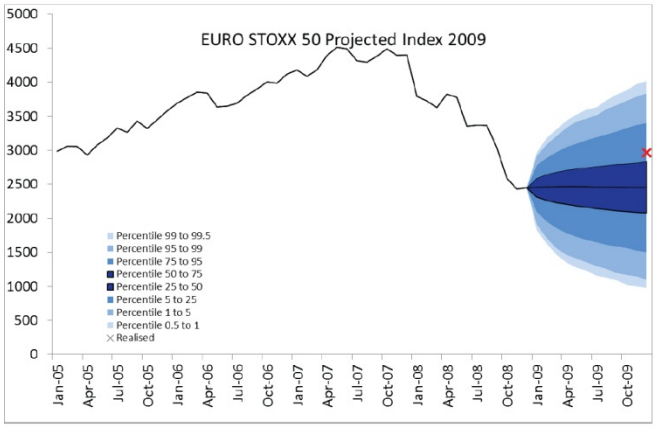
3.3.1 EURO STOXX 50 RETURNS

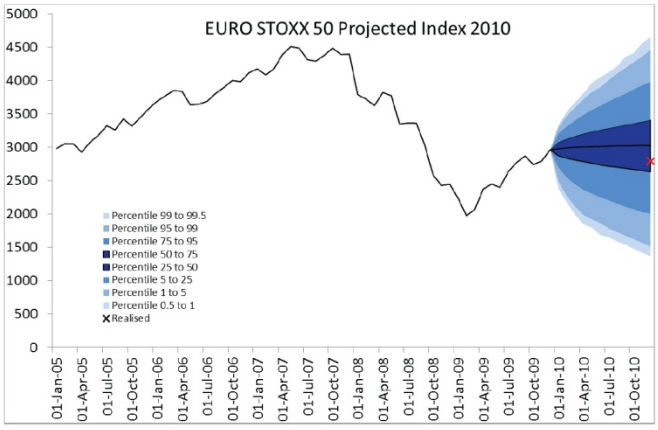
This example considers the risk arising from exposure to the EURO STOXX 50 equity index. The risk factor model for the 1-year total index return is the ESG's Stochastic Volatility Jump Diffusion (SVJD) model. Its standard 1-year VaR calibration approach is used. This model/calibration approach fits the model to distributional targets and moments, utilising option-implied volatilities to inform the real-world distribution of equity returns over the one year risk horizon. The SVJD model displays dynamic features that are exhibited in equity markets such as stochastic volatility and sudden large movements in asset prices. Note that this model choice is merely intended to be illustrative – the following approaches and discussion can be equally applied to any risk factor modelling approach that may have been adopted in a firm's internal model.

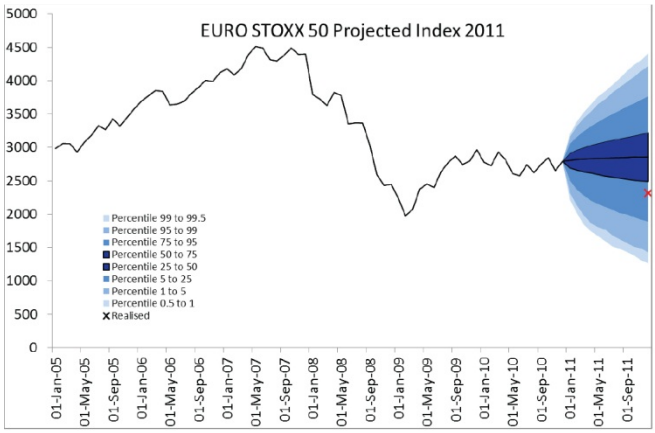
To illustrate the back-testing process, back-tests have been applied at five historic dates. These comprise calendar year-ends from 2007 to 2011. In each case, the model's standard 1-year real-world calibration at the calibration date is used, and 10,000 simulations for equity returns produced to obtain the model probability distribution for the EURO STOXX 50 equity index over the following calendar year.

Figure 2 Back-tests of 1-year EURO STOXX 50 return model over last 5 calendar years









These results are summarised in the table below:

SUMMARY OF RESULTS OF BACK-TESTS OF 1-YEAR EURO STOXX 50 RETURN MODEL

Year	Actual % change in EUROSTOXX	Map realised outcome to percentile on forecast distribution
2008	(44)%	0.9th %tile
2009	+21%	81st %tile
2010	(6)%	26th %tile
2011	(17)%	12th %tile
2012	+14%	66th %tile

Following the discussion of section 3.1, we should be sceptical of drawing any conclusions from the results of five back-tests and we would recommend extending this testing process as far back as data permits. (In this specific case, the use of option prices in the calibration process will limit the length of the historical back-test horizon). Nonetheless, the lack of any 'breaches' in the sample may give some support to the argument that the model is not obviously under-estimating risk (there is only a 2.5% probability that a 'correct' model would produce 1 or more or breaches from a sample of 5). From a more heuristic perspective, 2008 can be regarded as a particularly poor year for equity market performance by historical standards, and some comfort may be derived from the model's performance in estimating a 99.5th percentile for that year that exceeded the actual return.

3.3.2 GLOBAL COMPOSITE EQUITY PORTFOLIO

The above results considered the back-test results of a single equity index. These can clearly be repeated for any given index. It may also be of interest to consider the joint behaviour of a number of equity indices. This will be useful in validating the equity market tail dependencies in the model. To illustrate this, the back-testing was repeated for a portfolio which was invested 50% in EURO STOXX 50 and 50% in S&P500, with the USD currency exposure hedged in to EUR. The results are summarised below:

SUMMARY OF RESULTS OF BACK-TESTS OF 1-YEAR GLOBAL COMPOSITE EQUITY RETURN MODEL

Year	Actual % change in Portfolio	Map realised outcome to percentile on forecast distribution
2008	(41)%	0.6th %tile
2009	+22%	85st %tile
2010	+3%	43rd %tile
2011	(9)%	19th %tile
2012	+14%	69th %tile

In the global composite portfolio, we can see that the model suggests that 2008 was a 0.6th percentile tail event for the global equity portfolio (compared to a 0.9th percentile event for the EURO STOXX 50 index). Whilst this does not suggest that the model is fundamentally under-estimating equity risk or tail dependency, it highlights that tail dependency is an important feature to test for in the validation process.

Validation should consider the firm's own portfolio behaviour rather than the index behaviour analysed above. The firm's risk models could allow for basis risk using a 'tracking error' volatility assumption, and this assumption could also be tested in the back-testing process.

A further back-testing example using risk-free interest rates is provided in the Appendix.

3.4 Profit and loss attribution

The Profit and Loss (P&L) attribution shows the sources of profit and losses split by risk type and is therefore a form of validation tool. One application is to identify any material risks which were not captured in the internal model. The importance of this benefit should not be under-estimated. The Profit and Loss attribution is one of the few elements of the validation process which can identify unmodelled risks. Most aspects of validation focus on the risks already captured in the internal model. However, it is often the risks which weren't captured, and thus that management were not aware of, which lead to unexpectedly large losses.

The Profit and Loss attribution can be used to perform a back-test by comparing the actual observed changes in capital to the distributions underlying the internal model. More specifically, for each risk analysed in the attribution report, the observed capital movement can be mapped to a movement in the underlying risk variable and the percentile on the risk driver distribution inferred. Actuaries and risk managers within the business can then use this to verify that the observed movement is not out of place in terms of its percentile on the distribution. This is particularly valuable for a period when financial markets were distressed to assess whether extreme percentiles were reached. For example, the attribution report for 2008 will show the observed movement in capital attributed to credit risk and these can be compared against the credit risk capital requirement to help assess whether the size of the capital requirement appears appropriate, in light of the experience observed during the financial crisis.

In the earlier examples we considered the underlying risk factor (e.g. equity returns or interest rates). The Profit and Loss attribution report opens up the possibility of extending the back-testing approach to look at the distribution of capital rather than the underlying risk drivers. This would require the following two components:

1. Capability to evaluate the net asset position quickly in each of the scenarios (i.e. both assets and liabilities need to be re-valued quickly which typically involves the use of proxy functions).
2. The insurer has in place a profit and loss attribution which attributes the movement in capital over the year to underlying risk factor.

This is unlikely to be feasible for past years, because either (1) or (2) may not exist. However, going forward this is an approach which insurers can consider given the increased use of proxy functions and subsequent ability to produce a probability distribution for net assets.

4. Sensitivity testing

Sensitivity testing of an internal model involves assessing the extent to which the model output and the assessed capital requirement are sensitive to the model's underlying assumptions. One of the main objectives of sensitivity testing is to determine the key assumptions that have a material impact on the capital results, and how results can change if other reasonable assumptions are made. Critically, this requires the validation process to be able to identify what range of assumptions could reasonably be made. A deep understanding of the model and calibration approach is required to identify what are its most subjective or least reliable elements and how the assessed capital requirement is impacted by reasonable changes in those assumptions. This is an interesting and challenging area in the context of risk factor models: the differences between evidence-based science, expert judgement and Knightian uncertainty are often in the eye of the beholder, particularly when modelling extreme tail probabilities.

Natural areas of consideration for sensitivity testing of risk factor models include:

» **Assumptions directly related to the treatment of data.** Decisions around the choice of datasets to use in model calibration processes can involve significant judgement or subjectivity and this can be an important area for validation to address. For example:

- What is the impact of using a dataset with a different historical time period to the one used in the calibration?
- If a weighting scheme has been applied to the data, how sensitive is the model output to the different choices for the weighting scheme?
- Have particular historical 'outliers' been removed from the dataset? How sensitive are the model calibrations to this? (When calibrating a model for the purposes of estimating the likelihood of extreme events occurring, the exclusion of historical extreme events from the calibration data is always a matter of considerable subjectivity and / or judgement.)
- If the data set is limited, possibly because data for the variable is only available for recent periods, should the data be extended using data from, for instance, an alternative index or different economy?

- Frequency of the data used in the analysis e.g. whether to use daily, monthly, quarterly or annual data.

» **Areas where expert judgement has been applied to assist in deriving the parameters in the model.** This may relate to the application of judgement to choice of model and/or assumptions regarding the data. For example:

Are there forward-looking assumptions that have been derived from sources other than historic data? An example could be that the firm has assumed interest rates will, on average, rise over the next year by more than is implied by today's forward curve as they believe the yield curve is temporarily distorted. How sensitive are results to this assumption?

Correlation and tail dependency is an area where expert judgement often must be applied –more statistical data is required to make statements about *joint* behaviour than about marginal behaviour. It is therefore natural to particularly consider the sensitivity of the correlation setting method and its assumptions to understand the potential impact on capital requirements.

» **The choice of probability distribution for the risk factor modelling.** Although statistical tests should be applied to assess goodness-of-fit, the choice of model distribution will generally still involve a significant element of expert judgement—there is rarely sufficient volume of historical data to provide powerful statistical testing of alternative distribution assumptions, particularly in the tails. Again, it is important that the sensitivity testing around the choice of distribution should cover joint distributions as well as marginal distributions. The validation of probability distribution choices may go beyond purely statistical or quantitative analysis and also consider if fundamental properties of the risk variable in question are reflected in the distribution choices. For example, does the model choice allow arbitrage between government bonds of different maturities? Does this matter in the context of the risk assessment? The importance of having model choices that reflect these fundamental properties will largely depend on the complexity of the asset-liability profile. For example, a dynamic swap re-balancing strategy may produce highly unreliable results from an interest rate model that is not arbitrage-free.

It is also worth considering whether fundamental decisions about the over-arching calibration objectives should be subject to sensitivity testing. An example is whether a Point-In-Time approach (also referred to as a conditional calibration) is adopted or a Through-The-Cycle (also referred to as unconditional calibration) is adopted as the form of probability distribution that is intended for the risk model[7]. The point-in-time approaches utilises the market data available at the valuation date to calibrate the model and will produce a conditional stress test. The through-the-cycle approach produces an unconditional distribution, and leads to capital requirements based on an "average" level of risk. Part of the sensitivity analysis process may consider how sensitive results are to this choice of calibration approach.

4.1 Example: interest rate risk

To illustrate how sensitivity testing can be performed in practice, the following example considers the modelling of interest rate risk using the ESG's Extended 2-factor Black-Karasinski model. The example uses the standard ESG Point-in-Time calibration for 1-year VaR modelling. As discussed above, a robust sensitivity analysis requires a detailed understanding of the model and calibration method that has been implemented. So, before developing the sensitivity analysis, we first provide an overview of the process used to produce this calibration[8]. This yield curve model is a short rate model that specifies a stochastic process for the short-term interest rate as follows:

$$d\,[ln\,r(t)] = \alpha_1(\ln m(t) - \ln r(t))dt + \sigma_1 dZ_1(t)$$
$$d[\,ln\,m(t)] = \alpha_2(\mu(t) - \ln m(t))dt + \sigma_2 dZ_2(t)$$

In the Extended 2-factor Black-Karasinski, the starting risk-free yield curve is a direct input to the model (the function $\mu(t)$ is specified so that the model produces an exact fit to the starting yield curve). There are therefore four parameters (α_1, α_2, σ_1 and σ_2) that determine the one-year volatility produced by each point on the yield curve and the correlation between these points.

In the standard calibration method, these four parameters are calibrated to targets for the 1st percentile of the interest rate distribution (par yields) of maturities 1, 3, 10 and 30 years. These percentile targets are set assuming that the probability distribution of each par yield is log-normal with a mean specified by the current forward par-yield (i.e. no term premium in the yield curve) and with volatility derived from the current at-the-money market swaption implied volatility of appropriate tenor and one-year maturity. The method assumes that swaption implied volatility is an unbiased measure of one-year real world interest

[7] A detailed discussion of the two approaches is given in： http://www.barrhibb.com/research_and_insights/article/pitfalls_of_through-the-cycle

[8] A more detailed discussion of the calibration method can be found in the B&H methodology document: "1-year ahead interest rate tails: Initial analysis & discussion" (September 2010)

rate volatility, i.e. no volatility risk premia are assumed, and swaption implied volatility is equal to volatility for corresponding government bonds.

A sensitivity analysis is developed by identifying some of the key assumptions made in the model calibration process and then producing alternative calibrations that are based on other reasonable assumptions. Sensitivity testing should focus on areas where subjective judgement has been involved in the calibration process. Based on the above calibration process, natural areas of the calibration process to consider for sensitivity testing could include:

» **The choice of percentile that is targeted in the calibration process**. For example, does the calibration materially change if it targets the 0.5th percentile of the distribution rather than the 1st percentile?

» **The zero term premium assumption**. The calibration assumes that the term premium is zero, i.e. that interest rates are expected to follow the forward rate path. The firm may have an expectation that the yield curve will evolve differently. This could be based on a long-term view of the term premiums embedded in yield curves, or a short-term view based on the unusual current economic environment and the impact that is having on bond pricing. The model has a market price of risk parameter that can be used to produce expected paths for interest rates that differ from the forward implied path. For example, if the firm has an expectation that at the end of the year the long end of the yield curve will be 50bp above the forward implied curve, then the model can be calibrated such that the average curve reflects this assumption.

» **The assumption that the (Point-in-Time) volatility forecasts should be derived from swaption-implied volatility.** The calibration process could instead consider using a volatility derived from historical yield curve data. Alternative approaches to calibrating to historical data, and in particular the impact of using different lengths of historical data period, can be sensitivity tested. To derive the volatility assumption in the sensitivity test below, monthly EUR swap data (1, 3 and 10 year maturity) was used and the volatility was derived as the annualised standard deviation of the change in the log of the swap rate. An exponential weighted moving average (EWMA) estimator of standard deviation was used (this places more weight on more recent data, which is consistent with the Point-in-Time calibration objective).

» **The assumption that rates are lognormally distributed.** An important and obvious candidate for performing a sensitivity test is the choice of probability distribution. In the 2-factor Black-Karasinski model the short rate is lognormally distributed and longer-term yields are approximately lognormal. In this example the B&H ESG's Libor Market Model plus (LMM+) has been considered as an alternative choice of model. This is an extension to the standard Libor Market Model, where the forward rate distribution is shifted (displaced diffusion) and a stochastic variance process is incorporated[9]. The LMM+ model has more flexibility to control the distribution of rates and can capture fatter left-hand tails in interest rate distributions (including negative rates).

To illustrate the approach to sensitivity testing in the context of interest rate modelling, the example below considers a simplified asset-liability portfolio with a stream of fixed liability cash flows backed by an asset portfolio of government bonds which produces a stream of fixed level cash flows. For illustrative purposes, a duration gap of 2 years has been assumed - the duration of the assets is 8 years and that of the liabilities is 10 years. Both assets and liabilities are valued by reference to the EUR government bond curve10 and the initial surplus is zero.

If we focus solely on the capital requirement, defined as the 1-year VaR at the 99.5th percentile, then the results are given below (capital requirement is expressed as % of initial liability). The 99.5th percentile of the 5 and 10-year spot rate is also included to illustrate the how the tail of the yield curve distribution is impacted.

[9] Refer to the B&H documentation "LMMPlus: Model Definition" (December 2011) for a full description of the LMM+ model

[10] For this analysis the calibration data set only includes French and German government bonds

SUMMARY OF SENSITIVITY TESTS

	Sensitivity	Capital Requirement	99.5th percentile of 5-year spot rate	99.5th percentile of 10-year spot rate
Base	Base calibration	3.9%	22bp	71bp
1	Calibrate to 0.5th percentile instead of 1st percentile	4.2%	23bp	64bp
2	Change term premium assumption to deviate from the forward implied path. 50bp increase at the long end of the curve, relative to the forward implied curve.	3.0%	28bp	90bp
3	Historically derived volatility assumption, using EWMA, with mean data age of 1 year	3.6%	32bp	81bp
4	Historically derived volatility assumption, using EWMA, with mean data age of 3 years	3.2%	40bp	97bp
5	Historically derived volatility assumption, using EWMA, with mean data age of 5 years	2.8%	44bp	107bp
6	Interest rates modelled using LMM+	3.7%	0bp	48bp

The above results of the sensitivity analysis lead to the following conclusions:

» The decision regarding the inclusion of a term premium in the calibration approach is important. In sensitivity 2 the median of the distribution is higher relative to the base calibration, consistent with the view that the yield curve is expected to be higher than the forward implied path. Consequently, the left hand tail of the distribution is also impacted, leading to smaller downward interest rate stresses relative to the initial yield curve. In our example this reduces the interest rate capital requirement by almost a quarter.

» The analysis also highlights that the choice of data used to create the1-year volatility forecasts is critical (i.e. current swaption prices or historical data; if historical data then what historical period). At end-2012 when swaption implied volatilities are high relative to historic levels, their use in the calibration process gives a higher capital requirement. The base calibration has calibrated directly to swaption prices and has made no allowance for any risk premia that may be present in these option prices. The use of historical data produces significantly lower volatility estimates and hence a significantly smaller capital requirement. The results produced by the historical calibration method are highly sensitive to the choice of the data window. This is perhaps unsurprising given how much variation in volatility of yield curves has been observed in recent history. The calibration approaches considered in the sensitivity tests are all intended to still be in the spirit of the Point-in-Time calibration objective. A longer-term Through-the-Cycle calibration method could produce a quite different view of the interest rate tail risk.

» The choice of model and hence interest rate probability distribution is important for the rate behaviour produced in the tails. In the above analysis sensitivity 6 shows the results obtained for the LMM+ model. Consistent with the calibration approach adopted for the 2-factor Black-Karasinski (2FBK) model, conditionality is incorporated by calibrating to swaption implied volatility for 1-year options. A key feature of the LMM+ model is the inclusion of an absolute volatility component in the model. Relative to the lognormal model, this feature results in higher volatility and results in a fatter left hand tail, most noticeably at the short end of the yield curve where yields are currently lower. The yield curve distributions produced by the 2FBK and LMM+ models therefore differ significantly at shorter terms, and to a lesser degree at longer terms. Whilst the LMM+ model choice does produce a more extreme left-hand tail for long rates, the impact on increasing capital requirements that is produced by this is somewhat offset by the even more extreme tail produced by short rates: in the example there are excess asset cash flows at the short end that perform well in these tail scenarios, and the total capital requirement therefore falls when we move to this alternative yield curve model. This highlights that the type of yield curve risk exposure produced in this example is driven by the projected behaviour of the *slope* of the yield curve as much as the *level* of the yield curve. So in the context of this example the validation process should consider sensitivity testing the end-year probability distribution of the slope of the yield curve.

4.2 Tick-box exercise or useful validation tool? The so-what in sensitivity testing

A sensitivity analysis will generally show that quite different capital requirement conclusions can be reached when different yet feasible model assumptions are used. What can the firm do with this information? Is sensitivity testing just a "tick-box" exercise to meet regulatory validation requirements or does it provide a meaningful insight in to the chosen risk factor model and calibration approach? Sensitivity testing cannot provide many, if any, definitive answers on how to calibrate the model. Furthermore, sensitivity testing relies heavily on expert understanding of the model and calibration process – for sensitivity testing to provide useful insights it requires in-depth model and calibration process understanding in order to decide what sensitivities to perform.

However, we believe that well-targeted and rigorous sensitivity testing can highlight where a model's most critical and/or tenuous assumptions lie, and what their impact on capital requirements can be. The inherent subjectivity and expert judgement demanded by the task of making forward-looking statements about the severity of 1-in-200 year events means that there is always likely to be some critical and fundamental choices that drive the model's conclusions. The results of the sensitivity testing can be important in helping the firm determine where these choices have been made, and how the model's robustness may be improved in the future. It can also aid regulators and senior managers in developing their understanding of the limitations of the capital modelling and its conclusions.

It is again useful to consider the recent reporting of risk model calibration methods used at some banks which mentioned the "common tactics used by banks to reduce capital requirements where historical data windows and weighting schemes where selected to minimise VaR results". This provides a warning of what can happen when principle-based models do not undergo rigorous validation and supports the view that there is a useful role for sensitivity testing.

5. Stress and scenario Testing

Global principle-based regulation has increasingly made use of stress and scenario testing as an important element of the overall capital assessment framework. In banking, US and EU regulators have made extensive use of multi-year stress testing in recent years[11]. Insurance regulators in the European Union and North America have published proposals around Own Risk and Solvency Assessment that make significant use of stress testing[12]. In these examples, firms will implement a stress testing programme to understand the impact of stresses on the firm's capital position. The firm can undertake an analysis of the impact of single large events (stress test) and combinations of events (scenario test). The stress and scenario testing process can also play an important part in the validation of the firm's capital model by showing how the balance sheet impacts produced by stress and scenario tests relate to the level of capital that the firm believes it requires. Stress and scenario testing is set out in Solvency II's validation guidance as one of the key approaches to validation[13].

A fundamental challenge in stress and scenario testing is the identification of appropriate scenarios. The scenarios are not prescribed and there generally is not a specific level of probabilistic strength that they are required to meet. Firms are required to develop their own stress and scenario tests, taking in to account their particular business and risk profile. The stress tests and scenarios can be considered in two types:

3. **Historical stress events**, such as major stock market crashes, banking crises and so on. Given that the stress tests are applied to today's balance sheet, an additional complexity arises in the "translation" of the historical stress event to a stress to be applied in the current economic environment. For example, if an interest rate stress is based on an historical event when rates fell from an unusually high level, then it may be inappropriate to apply the absolute size of this shock directly to today's yield curve.
4. **Forward-looking stresses** – this is generally more subjective and should be specific to the insurer's risk profile. The development of forward-looking stresses is undoubtedly more challenging and views may be sought from a wide range of individuals including risk managers, asset managers, economists and academics, to decide on the appropriate stresses. This has a strong connection to the ORSA process which will also consider an assessment of the impact of a set of specified scenarios, which will likely include stresses over a medium-term horizon of, say, 5 years as well as instantaneous stresses.

In both cases the balance sheet impacts of the risk factor(s) outcomes in the specified scenario can be compared with the probability distribution produced by the firm's internal capital model. This can be a useful way of relating the output from the

[11] In the US the Comprehensive Capital Analysis & Review (CCAR) conducted by the Federal Reserve includes a supervisory stress test. In Europe the European Banking Authority has conducted EU-wide stress tests to assess the resilience of financial institutions to adverse market developments.

[12] Stress testing is specifically referred to in: NAIC's ORSA Guidance Manual (March 2013); EIOPA's report on public consultation on ORSA (July 2012); OSFI's draft guidelines E-19 on ORSA (December 2012)

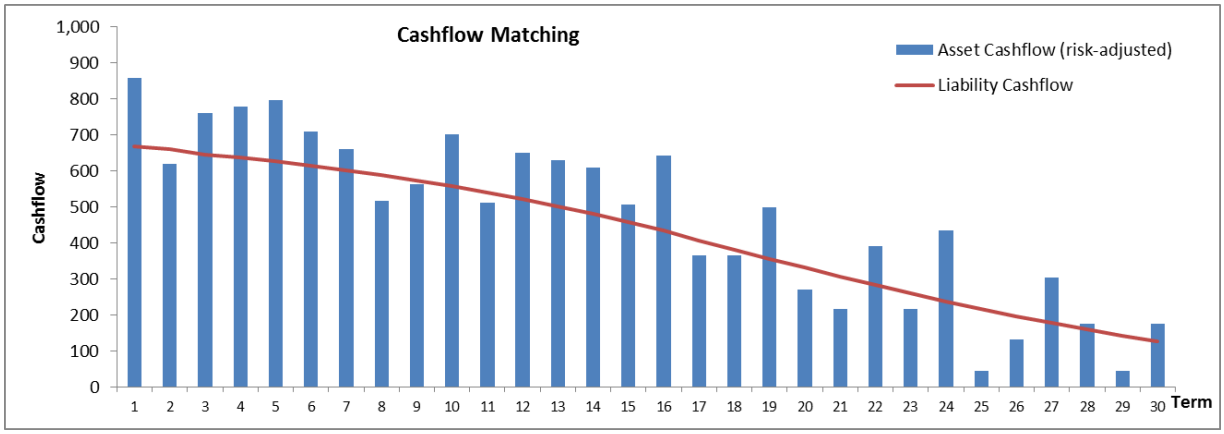
[13] Refer to guidelines 23 and 50 of the EIOPA consultation paper CP-13/011 (March 2013)

model to "real-world" events, helping users of the model's output to gain a deeper understanding of the model behaviour. However, as with back-testing and sensitivity testing, it is unlikely to provide absolute conclusions about model performance.

5.1 Example: 2008 financial crisis

This example considers a UK annuity portfolio where the backing assets are invested in fixed interest, with 90% in A-rated corporate credit and 10% in risk-free government bonds (the latter being used to achieve closer cash flow matching at longer terms). We assume the liabilities are valued with reference to the GBP swap curve, with no allowance for liquidity or matching premium and no counter-cyclical premium. The liabilities are valued at 10,000 and assets amount to 13,000 with the surplus of 3,000 invested in short-dated gilts. As the capital requirement is calculated by considering the variability of the total balance sheet, we model both the backing assets (90/10 corporate/gilts) and the surplus assets. The degree of cash flow matching for the backing assets is illustrated in Figure 3 – the duration of the assets is shorter than the liabilities by around one year.

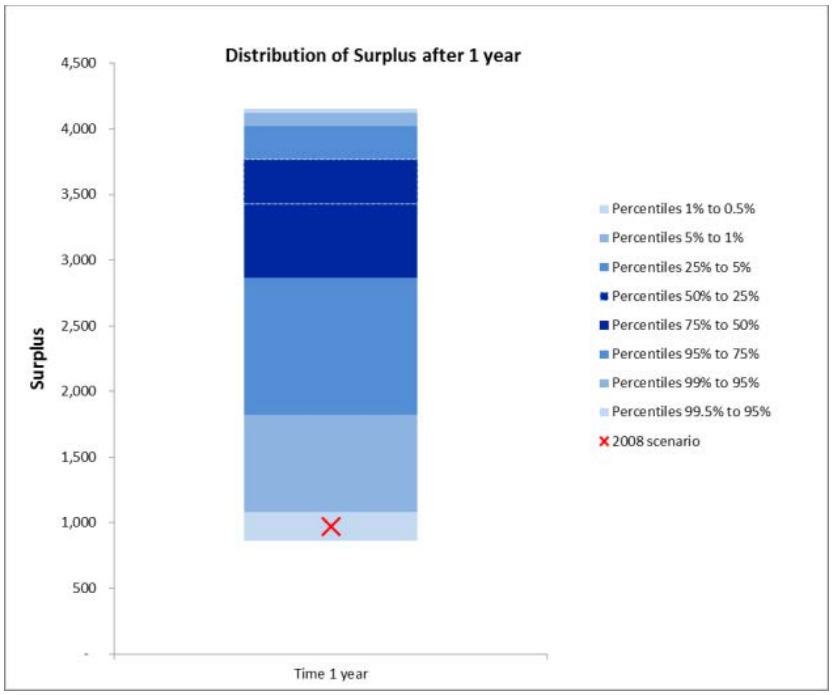
Figure 3 Asset and liability cash flow profile



We consider an historical scenario test based on the 2008 financial crisis. The scenario is therefore defined by the performance of corporate credit and gilts, and the movement in the swap curve observed over the calendar year of 2008. The scenario will be applied to the portfolio assuming the end-2012 start position. To do this we first apply the swap curve movement observed during 2008 to the stressed liability valuation. This could be implemented in the scenario test as either an absolute movement or a relative movement, and in this example we have assumed it is the absolute movement that is applied to the end-2012 swap curve (this provides a stronger scenario test given the yield curve was lower at end-2012 than in 2008). The corporate bonds are assumed to perform consistently with that observed during 2008 – the total return is (12)% based on the 2008 total return indices produced by iBoxx Sterling non-gilts 'A'-rated indices (weighted by maturity bucket).

Using projected real-world scenarios generated from the firm's risk factor model, the capital position over the next year can be modelled to obtain the probability distribution. In this example interest rates have been modelled using the extended 2-factor Black-Karasinski model and credit has been modelled using an extension of the Jarrow-Lando-Turnbull model with credit spreads modelled as a Cox-Ingersoll-Ross process. The key assumption in the dependency structure is the correlation between credit spread movements and nominal interest rates. This is assumed to be -0.5 in this example. The probability distribution of capital is shown below, together with the result produced by the scenario test, which is represented by the red cross.

Figure 4 Probability distribution of capital produced by the firm's internal model



Applying the 2008 scenario to the portfolio implies a reduction in the available capital of 2,030. This is very slightly less than the 99.5th percentile implied by the example firm's distribution above. So in this example the internal model has (implicitly) mapped the 2008 financial crisis to the 0.5th percentile of the distribution. This can provide a useful and intuitive insight to management on the strength of the risk modelling and calibration, and also on the risk profile of the business and its capacity to withstand a similar event to the credit crunch of 2008.

This example has only considered a single scenario but a realistic stress and scenario test implementation would be expected to apply several cases, both historical and forward-looking scenarios, and a fuller picture should emerge from this process of the types of scenario which pose most risk to the insurer's solvency position. For example, if scenario testing included the 1929 Wall Street crash and 1930s Great Depression, 1986 Japanese asset price bubble, 2000 Dot-Com bubble, 2008 financial crisis, 2011 sovereign debt crisis etc.., and, in all cases, the insurer's capital requirement covers the realised outcome, users may gain substantial comfort from this (though it is worth noting that the risk models that would be designed at each of these dates may have passed all the earlier scenario tests and yet failed to adequately capture the unique features of the 'new' outcome). However, the lack of an explicit probabilistic target in stress testing can create uncertainty about how much adversity should be incorporated into firms' stress and scenario tests and hence what conclusions can really be taken from the results of the analysis.

The example above applies scenario testing to a fixed annuity portfolio but can be extended to more complex liabilities. For example, in the case of variable annuities with a dynamic hedging programme, scenario testing on the internal model can be used to assess whether the hedge did perform as effectively as assumed in the model. The greater the complexity of the risk management strategy and the modelling that quantifies its residual risks, the more demanding and important the validation process will be.

6. Conclusions

Principle-based capital assessment requires rigorous and robust validation of firms' internal models. This is important both in 'policing' the considerable freedoms that firms can have in assessing capital requirements under a principle-based system, and in the development of the firms' own understanding of the model methodologies weaknesses and limitations. The latter can play an important role in prioritising the firm's ongoing development of the internal model methodology. Validation can improve the

transparency of the model behaviour, aiding management's and regulators' understanding of what information can (and cannot) be reliably obtained from the model.

This paper has focused in particular on the risk factor model element of the internal models that firms use for principle-based capital assessment. We have paid particular attention to the 1-year risk horizon used in the 1-year VaR framework that is emerging as a global standard in economic capital in the insurance sector. The paper has discussed the forms of validation that we would expect to be present in a best practice validation process, and has used a series of examples to illustrate the challenges that need to be considered when implementing these validation methods for 1-year risk factor models.

Across each of the validation methods described in the paper – back-testing, sensitivity testing and stress and scenario testing – some general conclusions emerge in risk factor model validation:

» When the model is being used to estimate a 1-year 99.5th percentile, **the limited volume of relevant historical data creates significant challenges for risk factor model validation** as well as for the underlying calibration that is being validated. For example, even 300 years of *back-testing* can provide only very limited statistical information about the performance of the model. It can tell us if the model is profoundly under-estimating the 99.5th percentile risk, but cannot offer much more. And even in that instance, it is tempting for the modeller to conclude that back-testing failures should be viewed as irrelevant historical anomalies.

» *Sensitivity testing* can do a better job of analysing risk methodologies that make significant use of expert judgement to fill the vacuums created by the inadequacy of historical data. For this validation approach to be effective, a deep understanding of the risk model methodology is required – **in risk factor tail modelling, differentiating between evidence-based science, expert judgement and Knightian uncertainty can be a daunting challenge**. But the key to insightful sensitivity testing is the assessment of how much variation in the risk model and calibration results can be created by reasonable alternative methodology choices, and this requires judgement of how much inherent uncertainty there is in an appropriate risk methodology.

» *Stress and scenario testing* can provide an intuitive and pragmatic way of probing the performance of the risk factor model and calibration. **The somewhat arbitrary and non-probabilistic way that stress and scenario tests are identified can be viewed both as this validation approach's greatest strength and its greatest weakness**. Recent experience in the banking sector has arguably highlighted that stress testing processes can create a false sense of comfort by using stress tests that are not actually very stressful. But attaching a probability to a stress test requires a model, and such a model would need to be different to the one that the stress test is being used to validate. So a probabilistic approach to stress testing really just reduces to considering alternative model forms, which is essentially what sensitivity testing is.

Despite the above challenges and limitations, we believe that a rigorous validation process that utilises all three of the above validation approaches is likely to be critically important to the successful performance of principle-based capital assessment. Validation, when effectively implemented, has an invaluable role to play in identifying the limitations and uncertainties inherent in the risk modelling methodology.

Appendix: Risk-free interest rates back-testing example

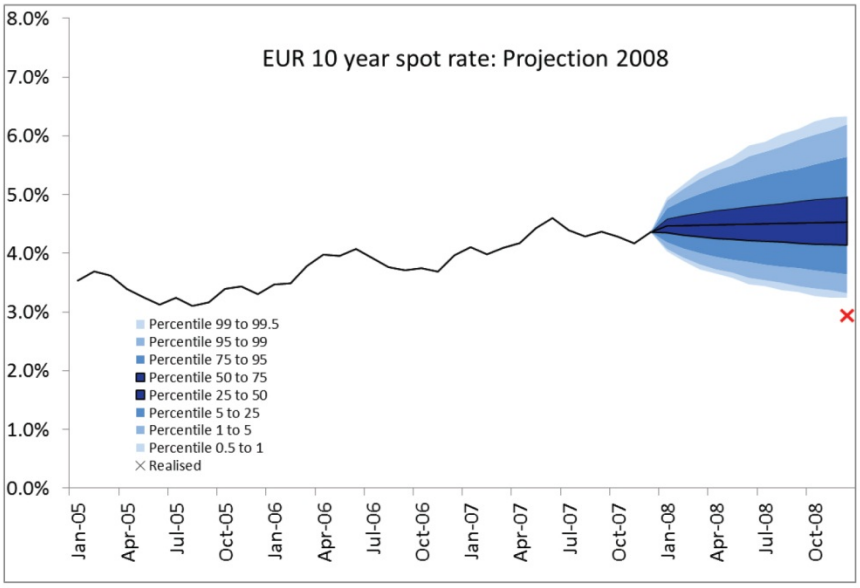
This section considers back-testing in the context of interest rate modelling. We develop an example using an interest rate model calibrated to EUR government bonds[14] and consider in particular the probability distribution produced by the model for the 10-year government yield. For the purposes of this example, our ESG's extended 2-factor Black-Karasinski model is used (this is a lognormal short rate model). The standard 1-year calibrations of the model are used throughout. Note these are Point-in-Time calibrations that use swaption-implied option volatilities in the calibration process. So the volatility of the projected interest rate distribution in the model is a function of both the level of yields at the calibration date (as it is a lognormal, proportional volatility model), and the level of swaption-implied volatilities.

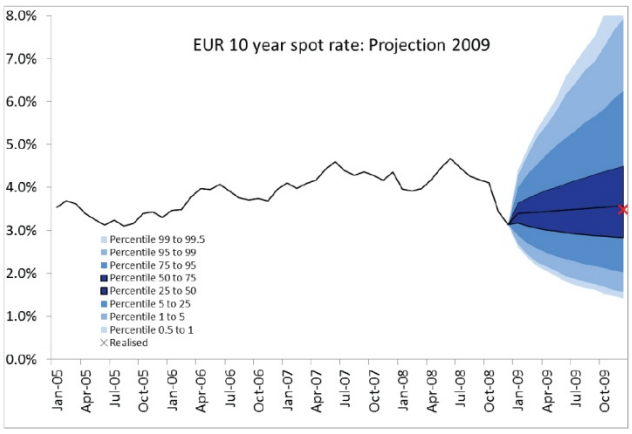
A back-testing process similar to that discussed in section 3.3 is developed here: model distributions are produced at the same five calibration dates; and these are compared with the realised outcomes for the variable in question, i.e. the 10-year bond yield at the end of the year.

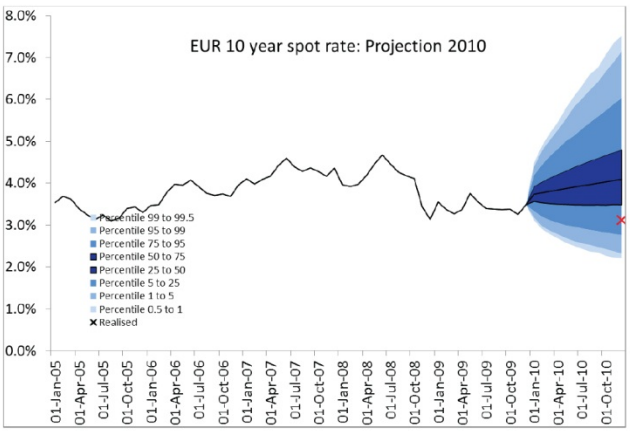
The results of the back-test are presented overleaf.

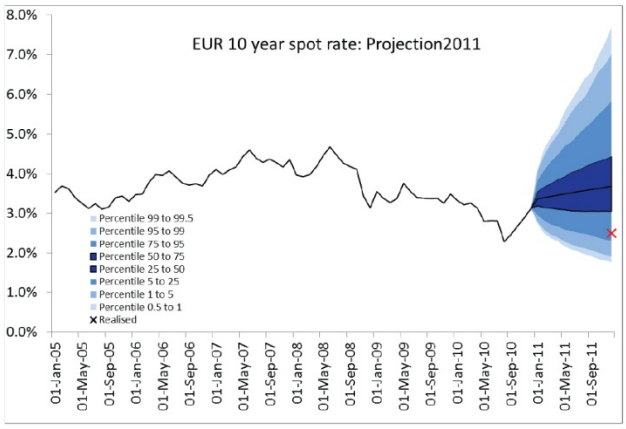
[14] For this analysis the calibration data set only includes French and German government bonds

Exhibit A.1: Back-tests of 1-year projection of 10-year EURO spot rate over last 5 calendar years









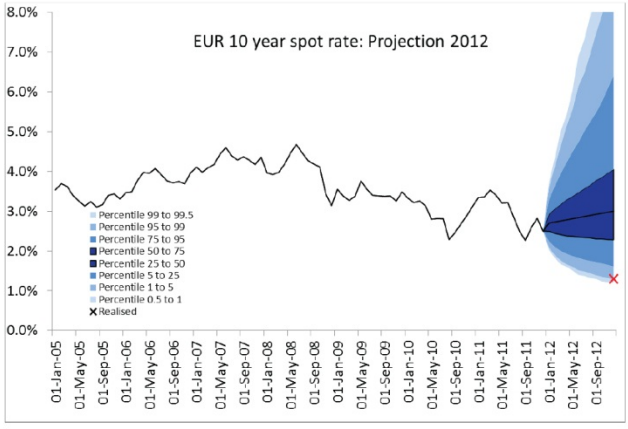


EXHIBIT A.2: SUMMARY OF RESULTS OF BACK-TESTS OF 1-YEAR PROJECTION OF 10-YEAR EURO SPOT RATE

Year	Actual change in 10-year government yield	Map realised outcome to percentile on forecast distribution	0.5th percentile of distribution
2008	-142bp	0.1st %tile	-122bp
2009	+35bp	47th %tile	-198bp
2010	-36bp	12th %tile	-150bp
2011	-63bp	8th %tile	-157bp
2012	-120bp	1st %tile	-147bp

Unlike the earlier equity return back-tests, this example does produce observed outcomes outside the 0.5th percentile of the distribution projected by our risk factor model. Again, using only 5 back-tests, it is difficult to come to robust statistical conclusions. It could be argued that 2008 was a '1-in-200 year' event – though the model's suggestions that it was a 1-in-1000 year event might raise some eyebrows! Furthermore, having two of the five observations in the 1st percentile or beyond is again statistically highly unlikely (a simple binomial calculation again suggests the probability of this occurring is less than 1 in 1000). So, even with only 5 data points, these back-test results raise significant flags with respect to the model's performance in capturing the extreme left-hand tail of 1-year changes in long-term interest rates.

As mentioned at the start of this section, the 2-factor Black-Karasinski model assumes interest rates are lognormally distributed. This arises from the definition of the model – interest rate volatility is defined to be proportional to interest rate level. The assumption of proportional volatility drives the shape of the interest rate distribution and it is particularly important to understand the implications of this feature in a low interest rate environment, i.e. the absolute volatility approaches zero as rates approach zero. An alternative yield curve model such as the B&H Economic Scenario Generator's LMM+ model has more flexibility to control the distribution of rates and can capture fatter left-hand tails in interest rate distributions.

The above example is a good illustration of where the validation process can identify the limitations of the implemented modelling approach and where a more sophisticated model can perform more robustly. It is then up to the individual firm to determine if the model limitations are likely to have a material impact on the ability to assess the capital requirements of the firm's risk profile, and whether a more complex and sophisticated model is therefore desirable.

Finally, it should be noted that the term structure of interest rates means that back-testing of interest rate models is more complex than back-testing of equity models. The above analysis has only considered a single point on the yield curve – the 10-year spot rate. Similar analysis can be performed at other points on the curve. The back-testing can also consider the joint movements of different points on the yield curve and validate whether this is adequately captured by the model. It is possible that the validation of individual yield curve points does not raise any concerns, but that the model fails to capture the more complex yield curve behaviour. For example, a 1-factor model may be sufficiently "strong" in terms of the output stresses to individual yield curve points but will not capture the exposure from yield curve twists.