**A Holistic Model Validation Framework for Current Expected Credit Loss (CECL) Model Development and Implementation**

Michael Jacobs, Jr.[[1]](#footnote-1)

Draft: December 15th, 2019

**Abstract**

The CECL revised accounting standard for credit loss provisioning is the most important changes to U.S. accounting standard in recent history. In this study we survey and assess practices in the validation of models that support CECL, across dimensions of both model development and model implementation. On the development side, this entails the usual SR 11-7 aspects of model validation, but highlighted in the CECL context are the impact of several key modeling assumptions upon loan loss provisions. We also consider the validation of CECL model implementation or execution elements, which assumes heightened focus in CECL given the financial reporting implications. As an example of CECL model development validation, we investigate a modeling framework that we believe to be very close to those being contemplated by institutions, which projects loan losses using time series econometric models, for an aggregated “average” bank using FDIC Call Report data. In this example we assess the accuracy of 14 alternative CECL modeling approaches, and further quantify the level of model risk using the principle of *relative entropy*. Apart from the illustration of several model validation issues and practices that are of particular relevance to CECL, the empirical analysis has some potentially profound policy and model risk management implications. Namely, implementation of the CECL standard may lead to under-prediction of credit losses and that the goal of mitigating the pro-cyclicality in the provisioning process that motivated CECL may fail to materialize.

Keywords: Accounting Rule Change, Current Expected Credit Loss, Allowance for Loan and Lease Losses, Credit Provisions, Credit Risk, Financial Crisis, Model Risk.

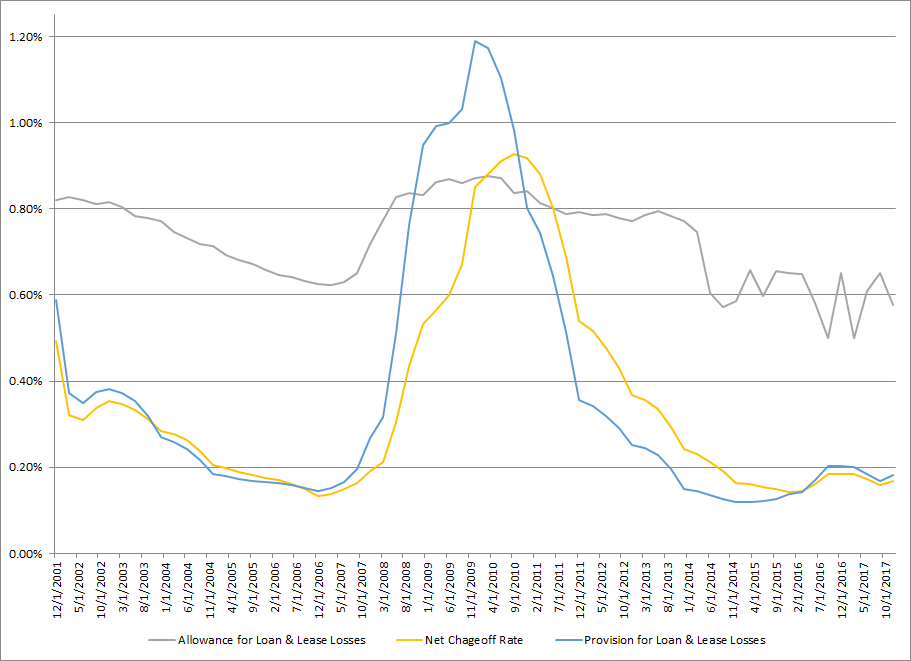
JEL Classification: G21, G28, M40, M48.**1 Introduction and Discussion**

In the U.S. the *Financial Accounting Standards Board* (“FASB”) issues of the *Generally Accepted Accounting Principles* (“U.S. GAAP”), financial accounting standards intended to ensure the provision of useful information for corporate stakeholders. In this study we focus on the *Allowance for Loan and Lease Losses* (“ALLL”), the financial reserves that firms exposed to credit risk set aside. The recent revision to these standards, the *Current Expected Credit Loss* (“CECL”; FASB, 2016) standard, is expected to substantially alter the management, measurement and reporting of loan loss provisions by financial institutions and other firms exposed to credit risk[[2]](#footnote-2).

The prevailing principle governing the ALLL has been that of *incurred loss[[3]](#footnote-3)*, wherein credit losses are either recognized only when materialized, else for performing loans the credit risk is assessed at the point of financial reporting. The key feature of this construct is that future events are not considered, which impairs the capability of managing reserves proactively based upon the expected economic outlook. The static nature of this framework implies that provisions are likely to exhibit *procyclicality*, which means that provisions rise and regulatory capital ratios decrease prior to economic downturns. This inflation in the ALLL at the trough of an economic cycle that the incurred loss standard engenders is considered detrimental to a bank from a safety and soundness perspective. Furthermore, from a systemic risk perspective this phenomenon represents a risk to the banking system and economy, as this could result in a contraction of credit in an environment of businesses and consumer vulnerability, resulting in a negative feedback loop adversely impacting the real economy.

This phenomenon is illustrated in Figure 1 where we plot *Net Charge-off Rates* (“NCORs”), the *rovision for Loan and Lease Losses* (“PLLL”) and the ALLL for the period 4Q01 to 4Q17. This data represents all insured depository institutions in the U.S. as made available by the Federal Deposit Insurance Company in the forms FR Y-9C, which are commonly termed the “Call Reports”. We show here an outstanding loan balance weighted average across all banks, or a depiction of an “average bank”. The figure shows that as NCORs and PLLs both rise at the beginning of the downturn period starting in 2007, the ALLL lags this increase and persists in rising well into the economic recovery period to peak in 2010. This build-up in the ALLL occurred just as bank asset quality was in decline, with worsening capital ratios, resulting in both declining earnings and lending activity and added to the severity of the so-called Great Recession.

The consensus view is that the motivation of FASB to consider the CECL standard was the procyclicality of the incurred loss standard, a narrative fortified by the severe consequences of the last downturn. This sentiment was rather broad based, emanating from various corners of the credit risk community spanning accounting, risk audit and supervision. The first mover in this initiative



**Figure 1: Net Charge-off Rates, Loan Loss Provisions and the ALLL as a Percent of Total Assets – All Insured Depository Institutions in the U.S. (Federal Deposit Insurance Corporation Statistics on Depository Institutions Report – Schedule FR Y-9C)**

by the standard setters was by the *International Accounting Standards Board* (“IASB”) in with the *International Reporting for Financial Statement Number 9* (IASB, 2104; “IRFS9”). IRFS9 represented a framework that was still *expected loss* (‘EL”) based for credit risk provisioning, as was the incurred loss standard, with a key difference being the incorporation of macroeconomic forecasts. FASB followed suite with the release of the new CECL (FASB, 2016), which while as with IRFS9 had the forecasting element, contained other fundamental differences from the legacy standard and IRFS9. In the case of performing loans, CECL prescribes a projection of EL over the anticipated tenor of the instrument, whereas under IRFS9 this life-of-loan expectation is limited to impaired assets. In the case of performing loans, IRFS9 limits the forecasting horizon to 1 year, similarly to the EL approach under the incurred loss standard, albeit minus the forecasting component in the case of impaired assets. IRFS9 also differs from CECL in that when the likelihood of impairment on an instrument increases materially, it contains conditions whereby the forecasting horizon switches from 1-year to the expected tenor of the instrument.

The scope of CECL spans all financial assets carried booked as *held-for-investment* (“HFI”) or *held-to-maturity* (“HTM”), or on an *amortized cost* basis. This is colloquially known as the *banking book* and constitutes most of the assets held by insured depositories. As mentioned previously, an important difference in CECL from incurred loss standard is that it is an EL methodology having a forecasting component that is applied to an expected life of instruments at the date financial reporting. A final point to note on CECL details is that the ALLL in this framework is a *valuation account*, defined as a difference in the *amortized cost basis* versus the expected accounting income to be received on loans. Operationally, from an EL perspective this means that increases expected lifetime credit losses imply a greater deduction from the amortized cost basis, and result in a build of reserves.

CECL requires that forecasts of EL are conditioned spot portfolio characteristics, the corresponding historical data used to estimate credit risk models (exposures, risk ratings and loan/obligor risk factors that drive *probability-of-default* – “PD”, *exposure-at-default* – “EAD” and *loss-given-default* – “LGD”) mapped to the latter and macroeconomic forecasts. FASB further requires that the credit risk model and macroeconomic forecasts be formed over a horizon that is deemed to be a *reasonable and supportable forecasting* (“RSF”) period, wherein it is expected that forecast errors would be tolerable and that institutions would be able to substantiate the quality of these forecasts.

FASB has not prescribed the econometric methodologies for either the credit risk or macroeconomic models, nor what constitutes reasonable and supportable assumptions underlying these. There are similarities between this state of affairs and the supervisory guidance around *Comprehensive Capital Analysis and Review* (“CCAR”) or the *Advanced Models Approach* (“AMA”) for Basel. A principles based guidance is meant to provide scalability in the approaches used across a diversity of institutions with respect to size or complexity, which offers the flexibility to choose methodologies grounded in reasonableness and supportability, albeit at the cost of potentially sacrificing comparability across entities. A good example of such elements is the RFS period, which is left unspecified but needs to be empirically or conceptually supportable. It is expected that the duration of the RSF period should reflect assets’ contractual tenors, and in cases where these terms are not known a prior (e.g., revolvingor unfunded commitments) they need to be estimated, which from a model validation perspective adds another potential source of model error to the process. On the flip side, when terms are known but there is empirical evidence that maturities will be extended (e.g., in periods of stress troubles loans might be held on the books longer than anticipated), CECL strictures do not allow banks to include this information in their forecasts, which also creates a model validation challenge in terms of the conceptual soundness of this construct.

The purpose of the ALLL is to provide information of use to critical stakeholders of an institution – such as investors, auditors and supervisors – pertaining to instruments exposed to credit risk. The nature of this information differs between CECL and the legacy accounting standard, with the former being a forward-looking view and the latter a backward-looking perspective. As a result, under the incurred loss standard we would expect changes in the provision to be driven by an exposure’s fundamental risk factors at the time of reporting (albeit with a 1-year horizon if an EL approach is used for the performing book), such as credit risk parameters (e.g., PD or LGD) or the variables that enter those models. In contrast, additional elements enter to drive the variation in the ALLL for the CECL case, namely the macroeconomic forecasting models. The practice well understands the modeling aspects that appear in both the CECL and incurred loss settings. Such constructs include loan or obligor features that are inputs to credit risk parameter models, such as company financial information or exposure structure details, as well as segmentations by industry or geography or qualitative characteristics defining management or the business environment. These models are calibrated to historical reference data-sets and then applied on a static basis over the forecasting horizon, where the loss factors underlying the models are held fixed throughout the projection.

Estimation of expected credit loss for CECL purposes features a critical difference from alternative settings (e.g., credit risk scorecards for underwriting or early warning, risk parameter development for regulatory or economic capita). A requirement of CECL is the estimation of EL over the expected tenor of instruments based upon reasonable and supportable forecasts of macroeconomic conditions. This in turn mans that models have to be constructed to forecast macroeconomic scenarios. The model validation consideration in this regard is that this nested set of models in the CECL engine engenders several challenges. One is the attribution of changes to the ALLL (e.g., scenarios, portfolio composition, credit model changes, etc.). Another challenge is the introduction of additional elements of model risk through the compounding or model errors and the likely questioning of model results by third party reviewers such as model validation, audit or supervision. Such modeling choices around the macroeconomic scenarios, which can be subjective or idiosyncratic in nature, are not common in current models supporting financial reporting or in other areas that incorporate macroeconomic forecasts such as CCAR where projections are generally sourced from the regulators, and hence such modeling considerations are not under scrutiny[[4]](#footnote-4).

In model validation of a CECL model we have to consider the main elements that distinguish this exercise from either the legacy loan provisioning approach or other credit risk modeling applications. The first consideration is the specification of the credit risk models that connect loan or obligor risk factors and macroeconomic variables to loan losses. The other key consideration is the scenario generation models that project macroeconomic forecasts and expected scenarios, which are as critical as the credit model specifications in driving outcomes. As it is known from related modeling exercises such as *stress testing* (“ST”), used by supervisors to assess the reliability of credit risk models in the revised Basel framework (Basel Committee on Banking Supervision, 2006) or the Federal Reserve’s CCAR program (Board of Governors of the Federal Reserve System, 2009), such models are subject to heightened supervisory scrutiny. A supervisory concern is that the complex models underlying these exercises may be subject to *model risk*, understood to be the danger that these constructs fail to quantify the risks that they have been designed to model (Board of Governors of the Federal Reserve System, 2011). In the case of CECL, it is expected that such reviews will be more rigorous as compared to other contexts such as Basel or CCAR given the financial statement reporting implications of CECL.

In this paper we contribute to the literature by considering and illustrating a model validation framework in the CECL setting. First, focusing on the model development dimension of model validation, we outline critical aspects of this set of activities, including the evaluation of conceptual soundness, theoretical framework and assumptions testing for CECL models. Second, we discuss a framework for the assessment and verification of the implementation and execution of CECL models. Finally, we present an example of CECL model validation, in which we analyze the impact of model specification and scenario dynamics upon expected credit loss estimates in CECL, through implementing a stylized framework borrowed from the ST modeling practice. In this exercise we perform a model selection of 14 alternative CECL specifications in a top-down framework, using Call Reports data to construct an aggregate or average hypothetical bank, with the target variable being the NCOR and the explanatory variables constituted by Fed provided macroeconomic variables as well as bank-specific controls for idiosyncratic risk.

In the example of CECL model validation, we study not only the impact on the ALLL estimate under CECL for alternative model specifications, but also the impact of different frameworks for scenario generation: the Fed baseline assumption, a Gaussian *Vector Autoregression* (“VAR”) model and a *Markov Regime Switching VAR* (“MS-VAR”) model, following the studies of Jacobs et al (2018) and Jacobs (2019). While the models all perform well (and similarly) by industry standards on an in-sample basis, the out-of-sample accuracy analysis exhibits severe underprediction across all models. Furthermore, this inaccuracy is accentuated in either the more complex specification of the credit risk models, or in the either the Fed or VAR as opposed to the MS-VAR macroeconomic scenario models, where the assumption of a linear process for the joint distributions of the macroeconomic variables is at odds with reality. The next phase of the model validation exercise quantifies the level of model risk across specifications through generating worse case loss projections using the principle of *relative entropy*, which attributes the contribution of model risk emanating from the credit risk models versus the macroeconomic scenario generation models. The finding in this regard is that the while more highly parametericized credit risk models give rise to greater measured model risk, it is in fact the less complex macroeconomic scenario generation models (assuming Gaussian dynamics) that also lead to increased model error. The model validation implications of these findings is that CECL model developers are advised to err on the side of either more parsimonious credit risk model specifications, or more realistic dynamics to characterize the distribution of macroeconomic variables, in order to mitigate underprediction or measured model risk.

This paper proceeds as follows. In Section 2 we review the related literature. In Section 3 we present a framework for CECL model validation of the conceptual soundness aspect, as well as. outlines our approach for the validation of CECL model implementation and execution. The econometric methodologies are outlined in Section 4. Section 5 presents an example of CECL model validation, where we assess the modeling data and empirical results. In Section 6, we perform our model risk quantification exercise for the various loss model and scenario generation specifications. Section 7 concludes and presents directions for future research.

**2 Review of the Literature**

The procyclicality of the incurred loss standard for the provisioning of expected credit losses has been extensively discussed by a range of authors: Bernanke and Lown (1991); Kishan and Opiela (2000); Francis and Osborne (2009); Berrospide and Edge (2010); Cornett, McNutt, Strahan, and Tehranian (2011); Carlson, Shan, and Warusawitharana (2013).

We note some key studies of model risk and its quantification that address issues in the implementation of the supervisory guidance on model risk (Board of Governors of the Federal Reserve System, 2011; “SR 11-7”). Glasserman and Xu (2013) quantify the impact of model risk through measuring and minimizing risk in a way robust to model error, starting from a baseline model to find the worst case error incurred through a deviation from this. This is based on a precise constraint on the plausibility of the deviation, using relative entropy to constrain model distance, leading to an explicit characterization of worst-case model errors. This approach transcends measuring the effect of estimation errors to consider errors in underlying stochastic assumptions of the model to characterize the greatest vulnerabilities to error in a model.

Jacobs (2015) contributes to the model risk management discipline by shifting the focus from individual models towards aggregating firmwide model risk, in line with SR 11-7 guidance that specifically focuses on measuring such risk individually and in aggregate. The author discusses various approaches to measuring and aggregating model risk across an institution, and also presents an example of model risk quantification in the realm of stress-testing, where he compares alternative models in the Frequentist and Bayesian classes, for the modelling of stressed bank losses.

In the practitioner realm, a whitepaper by Accenture Consulting (Jacobs et al, 2015a) notes that financial institutions are continuously examining their target state model risk management capabilities to support the emerging regulatory and business agendas across multiple dimensions, and that the field continues to evolve with organizations developing robust frameworks and capabilities. The authors observe that industry efforts to date have focused primarily on model risk management on an individual basis, and now more institutions are shifting focus to the aggregation of firm-wide model risk. They provide background on issues in model risk management, including an overview of supervisory guidance and discuss various approaches to measuring and aggregating model risk.

Jacobs et al (2015b) contribute to the literature by developing a Bayesian-based credit risk stress-testing methodology, which can be implemented by small-to-medium-sized banks, as well as presenting empirical results using data from the recent CCAR implementations. Through the application of a Bayesian model, they formally incorporate exogenous scenarios and also quantify the uncertainty in model output that results from stochastic model inputs. The authors compare the proportional model risk buffer measure of the severely adverse cumulative nine-quarter loss estimate obtained from the empirical implementation of a Bayesian as compared to a Frequentist model, finding it to be 40% higher in the former than in the latter. In the model validation exercise, they find that the Bayesian model outperforms the frequentist model statistically significantly, according to the cumulative percentage error metric, by 2% (1.5%) over the entire sample (downturn period).

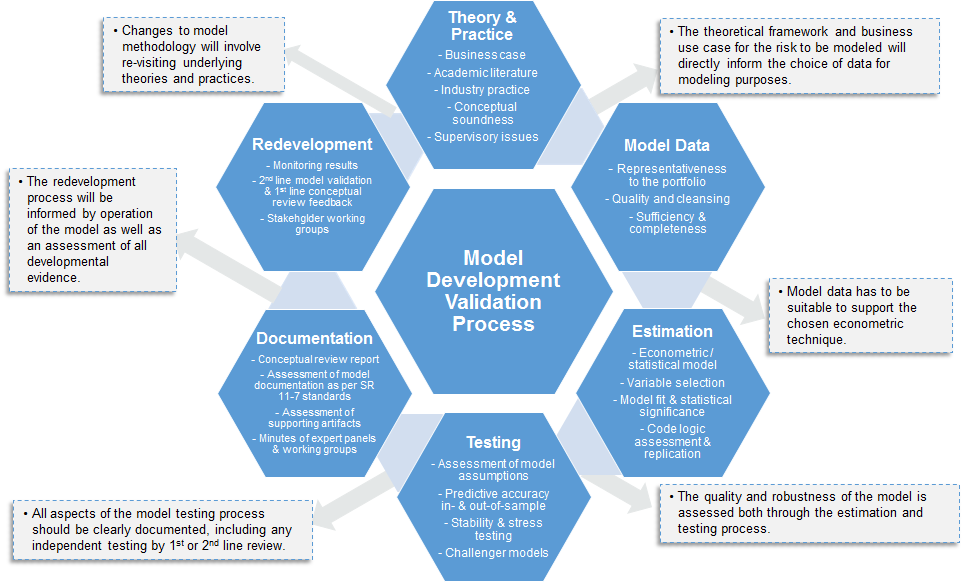
Skoglund (2018) studies the quantification of model risk inherent in loss projection models used in the macroeconomic stress testing and impairment estimation, which is of significant concern for both banks and regulators. The author applies relative entropy techniques that allow model misspecification robustness to be numerically quantified using exponential tilting towards an alternative probability law. Using a particular loss forecasting model, he quantifies the model worst-case loss term-structures, to yield insight into what represents in general an upward scaling of the term-structure consistent with the exponential tilting adjustment. The author argues that this technique can complement the traditional model risk quantification techniques where specific directions or range of reasons for model misspecification are usually considered.

There is rather limited literature on scenario generation in the context of stress testing. Jacobs et al (2018) examine this in the context of CCAR and credit risk by conducting an empirical experiment using data from regulatory filings and Federal Reserve macroeconomic data. The authors find that the a MS-VAR model performs better than a standard VAR model, both in terms of producing severe scenarios more conservative as well as showing superior predictive accuracy.

Jacobs (2019)r investigates a stress testing modeling framework believed to be very close to that employed by the regulators, which projects various financial statement line items for an aggregated “average” bank. He assesses the accuracy of alternative stress test modeling approaches, in particular simple single equation as compared to more complex multiple equation approaches. The results show potential inaccuracies in stress test model forecasts, in particular multi-equation models that do not properly account for the dependency structure amongst the input and target variables. Further the results highlight the public policy need for reconsidering the existent regulations that fail to place limits on the use of regulatory stress tests, and the need for supervisory models to be subject to model validation and governance standards.

In a study in the limited literature on CECL, Chae et al (2018) notes that CECL is intended to promote proactive provisioning, as loan loss reserves can be conditioned on expectations of the economic cycle. They study the degree to which a single modeling decision, expectations about the path of future house prices, affects the size and timing of provisions for first-lien residential mortgage portfolios. The authors find that while CECL provisions are generally less pro-cyclical as compared to the current incurred loss standard, the revised standard may complicate the comparability of provisions across banks and time.

Jacobs (2019) establishes that in general the CECL methodology is at risk of not achieving the stated objective of reducing the pro-cyclicality of provisions relative to the incurred loss standard. Across model specifications, he observes chronic underprediction of losses in the last 2-year out-of-sample period, arguably late in the economic cycle. It is found that the amount of procyclicality exhibits significant variation across model specifications and scenario generation frameworks. Relative to the perfect foresight benchmark, the MS-VAR model for macroeconomic scenario generation produces a lower level of variation in the model performance statistics as compared to the VAR or Fed models. As a second exercise, the author quantifies the level of model risk in an approach that uses the principle of relative entropy, finding that more elaborate modeling choices, such as more highly parametricized credit loss models, tend to introduce more measured model risk. However, the regime switching specification for scenario generation generates less models risk as compared to the Fed or vector autoregressive frameworks.



**Figure 2: Overview of a General Framework for the Validation of Model Development – Model Data, Estimation, Conceptual Review and Model Testing**

**3 Model Validation of CECL Model Development and Implementation**

In this section we discuss a general framework for model validation of the model development and implementation process in the CECL context. In Figure 2 above we present our framework graphically and illustrate the model validation process as following the model development life-cycle. The starting point in reviewing model development is an understanding of the *theory and practice* that underlies exercise. There must be a clear business case supporting the decision to embark upon model development. In the case of CECL, this is based upon accounting requirements, supervisory guidance and credit risk management policy considerations. The next critical element is a review of the relevant academic literature in conjunction with prevalent industry practice. In the case of CECL, we note that academic research may be limited, as compared to credit risk in general. As a consequence, we would rely more heavily upon peer bank practice, which can be comprehended through participation in industry associations and benchmark study exercises. Finally, this leads to an evaluation of the *conceptual soundness* around the approach chosen for CECL purposes. In this paper, we illustrate such developmental evidence in the CECL background review of Section 2, the literature review of Section 3 and the model theory / mathematics in Section 4.

The next pillar in this sequence is a review of the *model data*. A critical consideration here is the representativeness of the data supporting model estimation for the bank’s current portfolio, which has long been focus of supervisors going back to Basel implementation of the previous decade, as well as CCAR/DFAST in this decade. We would add that in the CECL context the challenge is to ensure that data is not unduly sampled from the stressed period, an element of conservatism seen in Basel or CCAR/DFAST. In some cases, this may necessitate the use of external or industry data, if internal data is not sufficient to meet this requirement. In this validation activity, an assessment of data quality and any data cleansing processes is critical, as illustrated in Section 7 of this paper.

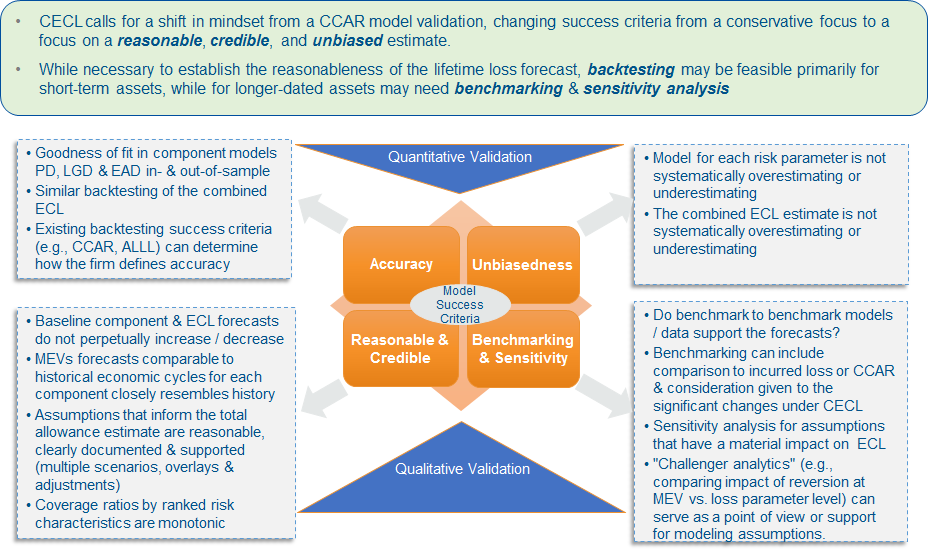
Transitioning from model data to *estimation*, we note that the complexity of the statistical or econometric model must be supported by the granularity or richness of the data. Considerations of data availability and quality will also drive segmentation and variable selection decisions. There is a wide range of practice amongst financial institute in this regard, and it is a key task of model validators to critically evaluate these choices. For example, some banks look at industry segmentations, while others do so by lines of business, and combinations of these. Another dimension is the level of modeling granularity, going from loan-level, to risk rating and then to segment level “top-of-the-house” approaches (i.e., the choice illustrated in this paper). Finally, the econometric approaches may vary and need to be justified, as we have choices such as ordinary least squares, time series, hazard rate, etc.; this is illustrated in Section 4 of this paper. Then, as shown in Section 5 of this paper, from a risk model validation perspective we must critically assess the model estimation results, aspects such as signs of coefficient estimates, measures of model fit, statistical significance of estimates, etc. Given the focus of accuracy and unbiasedness in the CECL context, the paramount model validation consideration is the quality of model fit in terms of results not showing evidence of chronic over- or under-prediction of losses. We will also highlight the role of code review in conjunction with the assessment of estimation results as part of model validation. This is not only the replication of estimation results, but a critical assessment of the structure of the code and the quality of the algorithms (e.g., is the code well organized and commented and are the algorithms robust and run time reasonable?).

In the subsequent *testing* aspect of model validation we have the evaluation of model performance along dimensions such as predictive accuracy, rank ordering and stability. While in principle this should be done on a both an in- and out-of-sample basis, a challenge in the CECL context is limitations in the length of the historical data supporting the latter analysis. Nevertheless, as illustrated in this study, even with data going back to the last downturn, it is feasible to look at the last few years as a holdout sample. That said, given how short a 2 year out-of-sample period is, we have to be cautious in arriving at strong conclusions. Therefore, we should supplement this basic backtesting by other exercises such as *stability* and *sensitivity* analysis (e.g., estimation omitting various time periods), as well as *stress testing* of the model (e.g., analysis of results under stress scenarios and comparison to historical periods of stress). Another element of model testing illustrated in this study is *challenger model* analysis, the comparison of results across different econometric specifications, either alternative variable selections or econometric specifications. Challenger model analysis can also include estimation of models under alternative segmentations in cases where models are estimated in industry or line of business groupings Finally, we could also have bottom-up challengers to segments level models, for example a loan level hazard rate formulations versus a rating migration model construct. A final point in this regard is that testing should be geared toward the objectives of CECL, which are oriented toward model accuracy as opposed to conservatism in the CCAR setting; and related to this point, champion and challenger models should all meet requirements under CECL (e,g,, in the case of LGD, there should be no discounting of workout recoveries).

The capstone of the model development cycle is documentation, the review and assessment of which is a critical ingredient in model validation. The ideal construct is documentation that would enable a 3rd party reviewer to replicate the model development having only the model development data in hand, but admittedly in the vast majority of situations model documentation is the industry falls short of this standard, especially in the case of complex and high risk models prevalent in CECL. Therefore, for practical purposes we prefer to take a more expansive view, and consider other artifacts of developmental evidence, such as various kinds of supporting documentation. Nevertheless, in general there are certain industry standards that model developers adhered to in order that the model documentation be considered fit for purpose:

* Clear articulation of the business purpose and portfolio profile pertaining to the model
* Comprehensive discussion of the model history
* Rigorous review of academic and practitioner literature relevant to the model
* Precise exposition of model mathematics and functional forms
* Complete examination of model data sourcing, cleansing processes and data quality analysis
* Thorough narrative on the segmentation and variable selection process, encompassing both statistical analyses and expert panel dialogue
* Informative description of estimation results, including comparison to challenger and candidate models
* Detailed analysis of all model testing that while supporting the choice of the champion model while also honestly representing potential weaknesses
* Understandable detail around any model adjustments, overlays and risk mitigation elements layered onto model output

In sum, model documentation should be a self-contained compendium of the elements of developmental evidence that we have discussed previously, and which are the object of the validation review. That said, this does not mean that model documentation should err on the side of being excessively voluminous, as has become the trend in the industry. In turn, the documentation of model validation should follow these standards, with the additional burden that there should be a value add in terms of analysis, observations, recommendations and conclusions.



**Figure 3: Challenges in the Validation of CECL Model Development -Benchmarking and Backtesting**

Model validation documentation should enable the improvement of the model development process, in addition to being a vehicle for managing and measuring model risk, and be comprehensible by both a technical as well as a lay-person audience. Turning to the CECL perspective, given the sensitivity of financial reporting and heightened scrutiny by 3rd party reviewers around this process, high quality documentation becomes a greater priority. Finally, we would submit that the structure of this study showcases several productive elements in what would be considered best practice model validation documentation by industry standards.

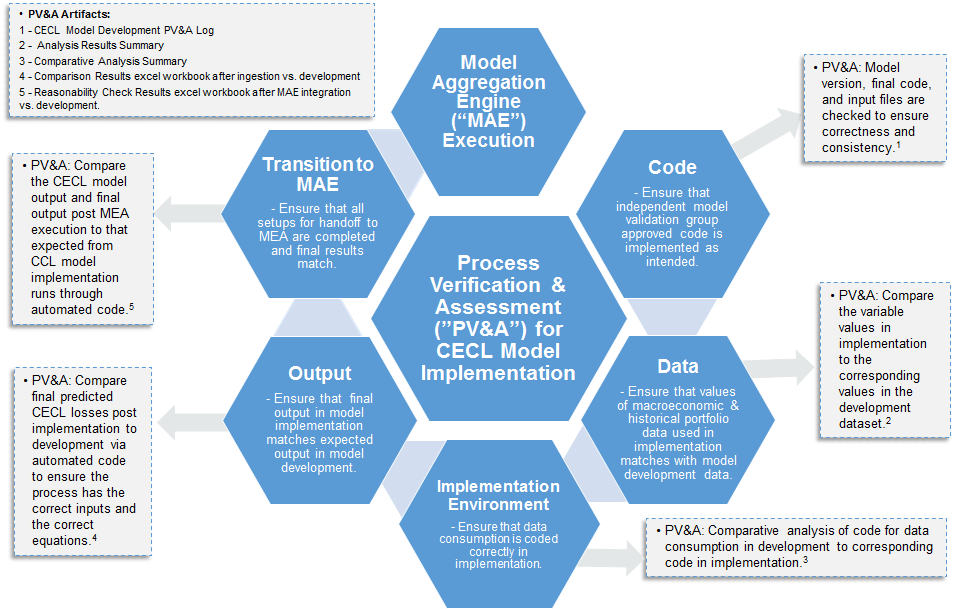
Finally for our model validation framework in Figure 2 we address model *redevelopment*, closing the loop on the model lifecycle process. The direction of future incarnations of a model will be based upon a combination of considerations informed by validation observations and recommendations, monitoring outcomes and user feedback. The role of model validation in this regard is to ensure that model developers judiciously weigh all these considerations in light of business needs, changes in the portfolio, advancements in the science of credit risk, availability of new or alternative data as well as the economics of redevelopment. The outcome of decisions made at this stage will have a direct bearing on the theoretical and practical directions taken. It is critical that model validation be involved in this redevelopment process, and moreover the criticality of this partnership with model developers is accentuated in the CECL context where the supporting models have implications for financial reporting.

Next we turn the focus on some model validation considerations that are rather particular to CECL within the testing phase of the process, namely challenges in backtesting and benchmarking, a framework for thinking about such illustrated in Figure 3 above. Whereas in stress testing and CCAR we have an emphasis of conservatism, and in credit ratings accuracy is balanced by considerations of stability or rating philosophy (e.g., a *through-the-cycle* orientation for underwriting of Basel application), in CECL the paramount considerations are the *reasonableness*, *credibility* and *unbiasedness* of estimates. In the graphic for the sake of exposition we organize these *model success criteria* concepts into 4 quadrants, labeling the positive and negative halves of the y-axis as quantitative versus qualitative considerations, respectively; and we add an additional element in the south-east quadrant, the qualitative elements of *benchmarking & sensitivity*, as such exercises are at the intersection of bias and credibility considerations.

Turning first to quantitative criteria, regarding accuracy we have the extant goodness-of-fit criteria of the CECL component models (RMM, PD, LGD or EAD), both in- & out-of-sample, as well as for the combined ECL. This aspect of CECL backtesting may leverage existing success criteria (e.g., CCAR, ALLL) in determining how the firm defines accuracy. However, in the context of credit parameter or bottoms-up models, the nuance for total ECL testing in the CECL setting for is that the consequence of combining these estimates is that the interaction of model errors may impact accuracy in the aggregate, and this is more problematic in the CECL as opposed to the CCAR application. Regarding the quantitative criterion of unbiasedness, the expectation is that the model for each risk parameter, as well as the combined ECL estimate, does not systematically overestimate or underestimate. This quantitative model success criterion is related to accuracy yet a distinct concept, in that for CECL there will be particular scrutiny of cumulative percent errors over certain key historical periods, such as the recent benign period that purportedly could precede a downturn.

Next we turn to discussion of the qualitative model success criteria. First considering the reasonability / credibility criterion for CECL estimates, a desirable characteristic is that expected component risk model or aggregate ECL forecasts neither increase nor decrease in perpetuity. Depending upon the economic environment that we are expected to enter, after a period of either elevated or muted losses, there should be a reversion to a level consistent with average experience over the long term. This requires a reversion mechanism taking us from the *reasonable and supportable forecast* (“RSF”) to the *long-run average* (“LRA”) period that yields such an intuitive pattern, which is usually a mechanism that is not completely statistical in nature. For example, we may have immediate, linear or curvilinear reversion paths of some kind. That leads to the challenge in determining how to specify ECL curves in the LRA – options include reverting to long-run loss rates in segments, injecting long-run averages of macroeconomic variables into forecasting models or leveraging long-run averages for risk parameter components; and the justification for these choices rests largely upon judgmental and not purely quantitative considerations. Another qualitative element of this dimension is that for macroeconomic forecasts intended to be comparable to certain historical economic cycles, such corresponding risk components should closely resemble history, which is something that cannot always be established through model selection due to inherent limitations in driver availability or linear specifications. Furthermore, any assumptions that inform the total allowance estimate should be reasonable, clearly documented and rigorously supported. For example, any weightings on multiple scenarios should be intuitive with respect to where we are in the economic cycle, and any overlays or adjustments to the modeled ECL estimates should be linked to an identified model weakness and be substantiated by analysis. Another consideration is that coverage ratios by ranked risk characteristics should exhibit monotonic patterns, and this element can inform the latter qualitative adjustments if reasonable ordering is not evident from the modeled output. Finally, we consider benchmark or challenger constructs, which could be models, segmentations or data. We can ask the question of if these alternatives, which should be broadly comparable to our champion constructs, provide sufficient support for the modeled CECL forecasts. Benchmarking can include comparison to incurred loss estimates or CCAR baseline forecasts, with due consideration given to the significant changes under CECL. Another critical activity is sensitivity analysis for assumptions that have a material impact on the ECL, which can not only give us a sense of the variability in the CECL estimates, but also can inform overlays or adjustments to the model. A related exercise here is *attribution analysis*, where we decompose changes in model estimates over cycles into what is driven by forecasts, portfolio quality or modifications to the model. A last item to consider here is *challenger analytics*, for example comparing impact of reversion at macroeconomic variable versus the loss parameter level, which can serve as a point of view or support for modeling assumptions.

We conclude this section by a note on the roles of 2nd line versus 1st line of defense model validation. The framework that we have described, that is broadly in line with SR 11-7 standards, is applicable to model validation occurring within the model development (“1st line model validation”), as well as the same as performed by an independent 2nd line function. It is recognized that validation activities commonly occur within model development groups, with model development staff cross-checking work amongst themselves, to ensure quality and correctness. A trend in the industry is to add formality around this 1st line validation function, and a degree of independence by making it a separate function (albeit under the same reporting structure), in order to have an enhanced and improved process. This trend has accelerated with the advent of CECL, as having such a formalized 1st lines validation process enables and organization to have a more robust CECL process, in the sense of resiliency in the face of multiple 3rd party review.



**Figure 4: Framework for Process Verification and Assessment of CECL Model Implementation and Execution**

In Figure 4 above we depict at a high level both the process for implementing and executing a CECL model, the PV&A steps that are embedded in that process, as well as the evidences or artifacts of the PV&A . As a first step, the model implementation team checks that the code (e.g., the SAS and Python scripts, as well as the input files) deployed in the implementation process is that approved by the *independent model risk management group* (“IMRMG”), and this is recorded in the *CECL Model Development PV&A Log*. If the same programming languages are used in implementation as in development, this is a rather straightforward process, else there is a step to translate the development codes into implementation codes that needs to be verified and assessed. Second, there is a check of the data that drive the CECL forecasts, namely the risk factors in the econometric models, and the macroeconomic variables used in the regression models.

This is accomplished by a comparison that is recorded in an *Analysis Results Summary* document. The third step of the process involves establishing the implementation environment, which means ensuring that the code and data are ingested correctly by the model implementation platform. The artifact supporting the PV&A of this step is recorded in the *Comparative Analysis Summary* document. In the fourth step, there is an evaluation of the model implementation output, which is again verification that the risk factors and predicted credit losses which drive the CECL model forecasts match those in the model development stage. The correct completion of this control is evidenced by a *Comparison Results Excel* workbook. The final step involves the transfer of CECL models implementation outputs to the *model aggregation engine* (“MAE”), which is the execution stage of the CECL model, and the that these are complete and that results match, and the PV&A activity associated with this is the comparison of CECL model output with that of the output post MAE implementation to that expected from CECL model development runs. The artifact associated with this element of PV&A is an automated code that resides in in a *Reasonability Check* excel workbook.

**4 Example of a CECL Modeling Framework: Time Series VAR Methodologies for Credit Model Estimation and Macroeconomic Scenario Generation**

Stress testing is concerned principally concerned with the policy advisory functions of macroeconomic forecasting, wherein stressed loss projections are leveraged by risk managers and supervisors as a decision-support tool informing the resiliency institutions during stress periods[[5]](#footnote-5). Traditionally the way that these objectives have been achieved ranged from high-dimensional multi-equation models, all the way down to single-equation rules, the latter being the product of economic theories. Many of these methodologies were found to be inaccurate and unstable during the economic tumult of the 1970s as empirical regularities such as Okun’s Law or the Phillips Curve started to fail. Starting with Sims (1980) and the VAR methodology we saw the arrival of a new paradigm, where as opposed to the univariate AR modeling framework (Box and Jenkins, 1970; Brockwell and Davis, 1991; Commandeur and Koopman, 2007), the VAR model presents as a flexible multi-equation model still in the linear class, but in which variables can be explained by their own and other variable’s lags, including variables exogenous to the system. We consider the VAR methodology to be appropriate in the application of stress testing, as our modeling interest concerns relationships and forecasts of multiple macroeconomic and bank-specific variables. We also consider the MS-VAR paradigm in this study, which is closely related to this linear time-invariant VAR model. In this framework we analyze the dynamic propagation of innovations and the effects of regime change in a system. A basis for this approach is the statistics of *probabilistic functions of Markov chai*ns (Baum and Petrie, 1966; Baum et al., 1970). The MS-VAR model also subsumes the *mixtures of normal distributions* (Pearson, 1984) and *hidden Markov-chain* (Blackwell and Koopmans, 1957; Heller, 1965) frameworks. All of these approaches are further related to *Markov-chain regression models* (Goldfeld and Quandt, 1973) and to the statistical analysis of the *Markov-switching models* (Hamilton 1988, 1989). Most closely aligned to our application is the theory of *doubly stochastic processes* (Tjostheim, 1986) that incorporates the MS-VAR model as a Gaussian autoregressive process conditioned on an exogenous regime generating process.

Let  be a -dimensional vector valued time series, the output variables of interest, in our application with the entries representing some loss measure in a particular segment, that may be influenced by a set of observable *input variables* denoted by, an -dimensional vector valued time series also referred as *exogenous variables*, and in our context representing a set of macroeconomic or idiosyncratic factors. This gives rise to the  (“vector autoregressive-moving average with exogenous variables”) representation:

 (1)

Which is equivalent to:

 (2)

Where ,  and  are *autoregressive lag polynomials* of respective orders , s and , respectively, and **** is the *back-shift operator* that satisfies **** for any process ****. It is common to assume that the input process  is generated independently of the noise process [[6]](#footnote-6). The autoregressive parameter matrices  represent sensitivities of output variables to their own lags and to lags of other output variables, while the corresponding matrices  are model sensitivities of output variables to contemporaneous and lagged values of input variables. It follows that the dependency structure of the output variables , as given by the autocovariance function, is dependent upon the parameters , and hence the correlations amongst the  as well as the correlation amongst the  that depend upon the parameters . In contrast, in a system of univariate  (“autoregressive-moving average with exogenous variables”) models, the correlations amongst the elements of  are not taken into account, hence the parameter vectors  have a diagonal structure (Brockwell and Davis, 1991). In this study we consider a *vector autoregressive model with exogenous variables* (“VARX”), denoted by , which restricts the *Moving Averag*e (“MA”) terms beyond lag zero to be zero, or :

 (3)

The rationale for this restriction is three-fold. First, in MA terms were in no cases significant in the model estimations, so that the data simply does not support a VARMX representation. Second, the VARX model avails us of the very convenient DSE package in R, which has computational and analytical advantages (R Development Core Team, 2019). Finally, the VARX framework is more practical and intuitive than the more elaborate VARMAX model, and allows for superior communication of results to practitioners.

We now describe how the parameter matrices of the  models are estimated. We may obtain consistent and efficient estimates using the method of *conditional maximum likelihood estimation* (“CMLE”), which in the case of the  model recursively optimizes the following conditional log-likelihood function:

 (4)

where , which is estimated by , and the are obtained recursively from:

 (5)

As well as the use of any necessary observed initial values of . CMLE estimates of the unknown parameter matrices , given by

 (6)

may be obtained by iterative numerical methods such as Newton-Raphson for the VARMAX model. In the special case of the  model (4.9), and where there are no constraints on the parameter matrices , we obtain a significant simplification of the estimation algorithm as the CMLE estimates are now given by linear multivariate least squares regression methods. In particular , we can express the  model (6) as follows:

 (7)

where  and . It follows that the CMLE estimator of  is given by:

 (8)

Where  and . Under standard stationarity conditions, the asymptotic distribution of is given by:

 (9)

We now consider the MS-VARX generalization of the VARX methodology with changes in regime, where the parameters of the VARX system  will be time-varying. However, the process might be time-invariant conditional on an unobservable regime variable , denoting the state at time  out of  feasible states. In that case, then the conditional probability density of the observed time series  is given by:

 (10)

Where  is the VAR parameter matrix in regime  and  are the observations . Therefore, given a regime , the conditional  system in expectation form can be written as:

 (11)

We define the innovation term as:

 (12)

The innovation process  is a Gaussian, zero-mean white noise process having variance-covariance matrix :

 (13)

If the  process is defined conditionally upon an unobservable regime  as in equation (4), the description of the process generating mechanism should be made complete by specifying the stochastic assumption of the MS-VAR model. In this construct,  follows a discrete state homogenous Markov chain:

 (14)

Where  denotes the parameter vector of the regime generating process. We estimate the MS-VAR model using MSBVAR the package in R (R Development Core Team, 2019). Finally note that in the remainder of the document outside this section we will use the acronyms VAR and MS-VAR instead of VARX and MS-VARX to refer to our competing modeling methodologies.

We close out this section by commenting on a view of this methodology from a model validation of conceptual soundness perspective. We would submit that for CECL purposes, and in a top-down modeling setting (i.e., where the target variables are losses in modeling segments, as opposed to loan or obligor level risk factors like LGD or PD, respectively), that the VAR or MS-VAR methodology is fit-for-purpose. These techniques are well understood by the academic and practitioner communities, and have been widely used in the industry for credit modeling purposes including CCAR, so that such techniques are rather suitable for CECL modeling. That said, in validating these models we need to be mindful of whether the model code is correct and well-constructed, and that assumptions of these statistical models are satisfied. Indeed, in our model development process, we have followed industry and SR 11-7 standards in testing our model code and the assumptions of the econometric models, including supporting diagnostics such as sensitivity and stress testing of the models; and the construction of challenger models[[7]](#footnote-7).

**5 Example of CECL Model Development Validation: Analysis of Model Data, Estimation Results, Challengers and Performance Testing**

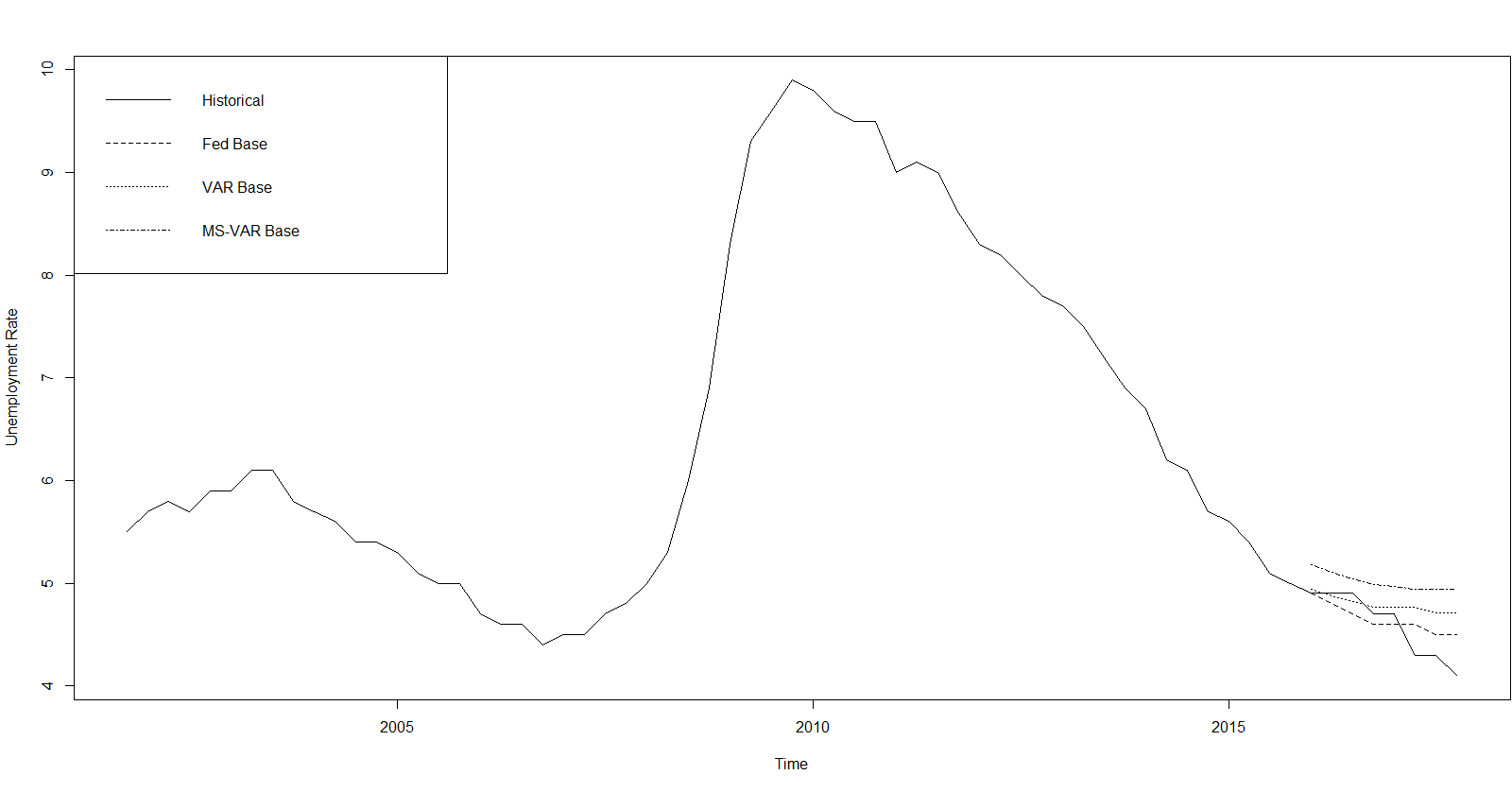
The data in this empirical estimation exercise are sourced from the *Statistics on Depository Institutions* (“SDI”) report, which is available on the *Federal Deposit Insurance Corporation*'s (“FDIC”) research website[[8]](#footnote-8). This bank data represents all insured depository institutions in the U.S. and contains income statement, balance sheet and off-balance sheet line items. We use quarterly data from the 4th quarter of 1991 through the 4th quarter of 2017. The models for CECL are specified and estimated using a development period that ends in the 4th quarter of 2015, leaving the last 2 years 2016 and 2017 as an out-of-sample test time period. The model development data are Fed macroeconomic variables, as well as aggregate asset-weighted average values bank financial characteristics for each quarter, the latter either in growth rate form or normalized by the total value bank assets in the system.

The Federal Reserve’s CCAR stress testing exercise requires U.S. domiciled top-tier financial institutions to submit comprehensive capital plans conditioned upon prescribed supervisory, and at least a single bank-specific, set of scenarios (base, adverse and severe). The supervisory scenarios are constituted of 9 quarter paths of critical *macroeconomic variables* (“MVs”). In the case of institutions materially engaged in trading activities, in addition there is a requirement to project an instantaneous market or counterparty credit loss shock conditioned on the institution’s idiosyncratic scenario, in addition to supervisory prescribed market risk stress scenarios. Additionally, large custodian banks are asked to estimate a potential default of their largest counterparty. Institutions are asked to submit post-stress capital projections in their capital plan starting September 30th of the year, spanning the nine-quarter planning horizon that begins in the fourth quarter of the current year, defining movements of key MVs. In this study we consider the MVs of the 2015 CCAR, and their base scenario for CECL purposes:

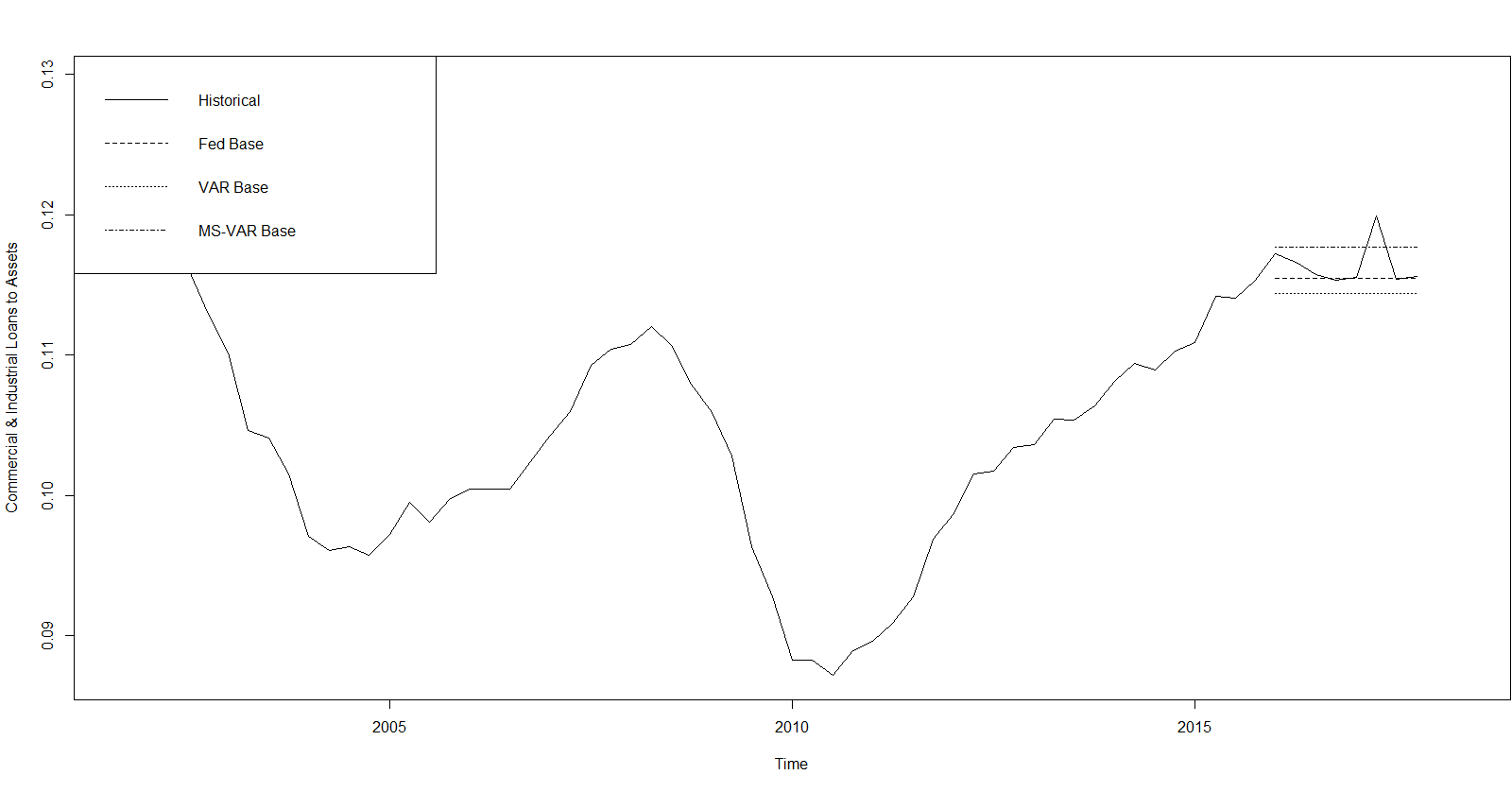
* Real Gross Domestic Product Growth (“RGDP”)
* Real Gross Domestic Investment (“RDIG”)
* Consumer Price Index (“CPI”)
* Real Disposable Personal Income (“RDPI”)
* Unemployment Rate (“UNEMP”)
* Three-month Treasury Bill Rate (“3MTBR”)
* Ten-year Treasury Bond Rate (“10YTBR”)
* BBB Corporate Bond Rate (“BBBCR”)
* Dow Jones Index (“DJI”)
* National House Price Index (“HPI”)
* Nominal Disposable Personal Income Growth (“NDPIG”)
* Mortgage Rate (“MR”)
* CBOE’s Equity Market Volatility Index (“VIX”)
* Commercial Real Estate Price Index (“CREPI”)

Our model selection process imposed the following criteria in selecting input and output variables across both VAR and MS-VAR models for the purposes of scenario generation[[9]](#footnote-9):

* Transformations of chosen variables should indicate stationarity
* Signs of coefficient estimates are economically intuitive
* Probability values of coefficient estimates indicate statistical significance at conventional confidence levels
* Residual diagnostics indicate white noise behavior
* Model performance metrics (goodness of fit, risk ranking and cumulative error measures) are within industry accepted thresholds of acceptability
* Scenarios rank order intuitively (i.e., severely adverse scenario stress losses exceeding scenario base expected losses)

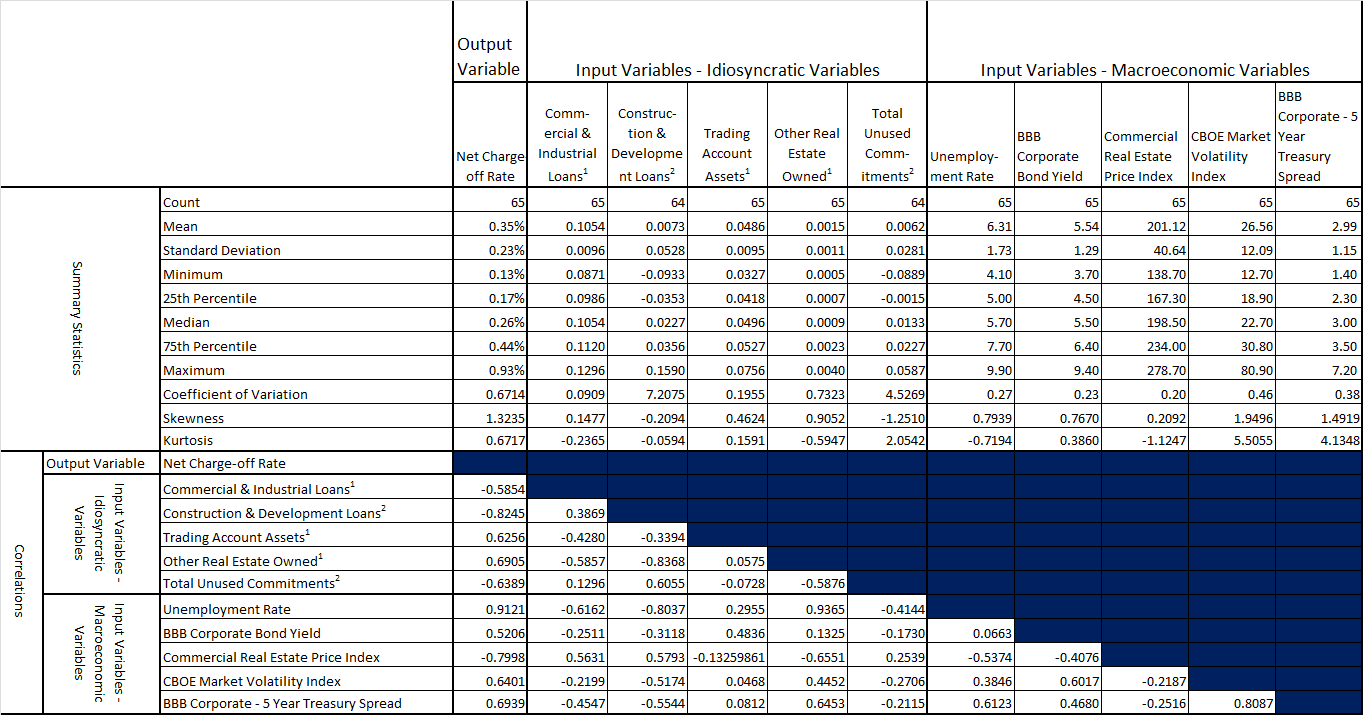


**Figure 1** Time Series and Base Scenarios - Unemployment Rate (Federal Reserve Board 4Q91-4Q15 and Jacobs et al (2018) Models).



**Figure 2** Time Series and Base Scenarios - Commercial and Industrial Loans to Total Assets (FDIC SDI Report, Federal Reserve Board 4Q91-4Q15 and Jacobs et al (2018) Models).

**Table 1.** Summary Statistics and Correlations of Historical Y9 Credit Loss Rates, Banking System Idiosyncratic Variables and Macroeconomic Variables (FDIC SDI Report and Federal Reserve Board 4Q91-4Q15).



We considered a diverse set of macroeconomic drivers representing varied dimensions of the economic environment, and a sufficient number of drivers (balancing the consideration of avoiding over-fitting) by industry standards (i.e., at least 2–3 and no more than 5–7 independent variables). According to these criteria, we identify the optimal set focusing on 5 of the 9 most commonly used national Fed CCAR MVs as input variables in the VAR model:

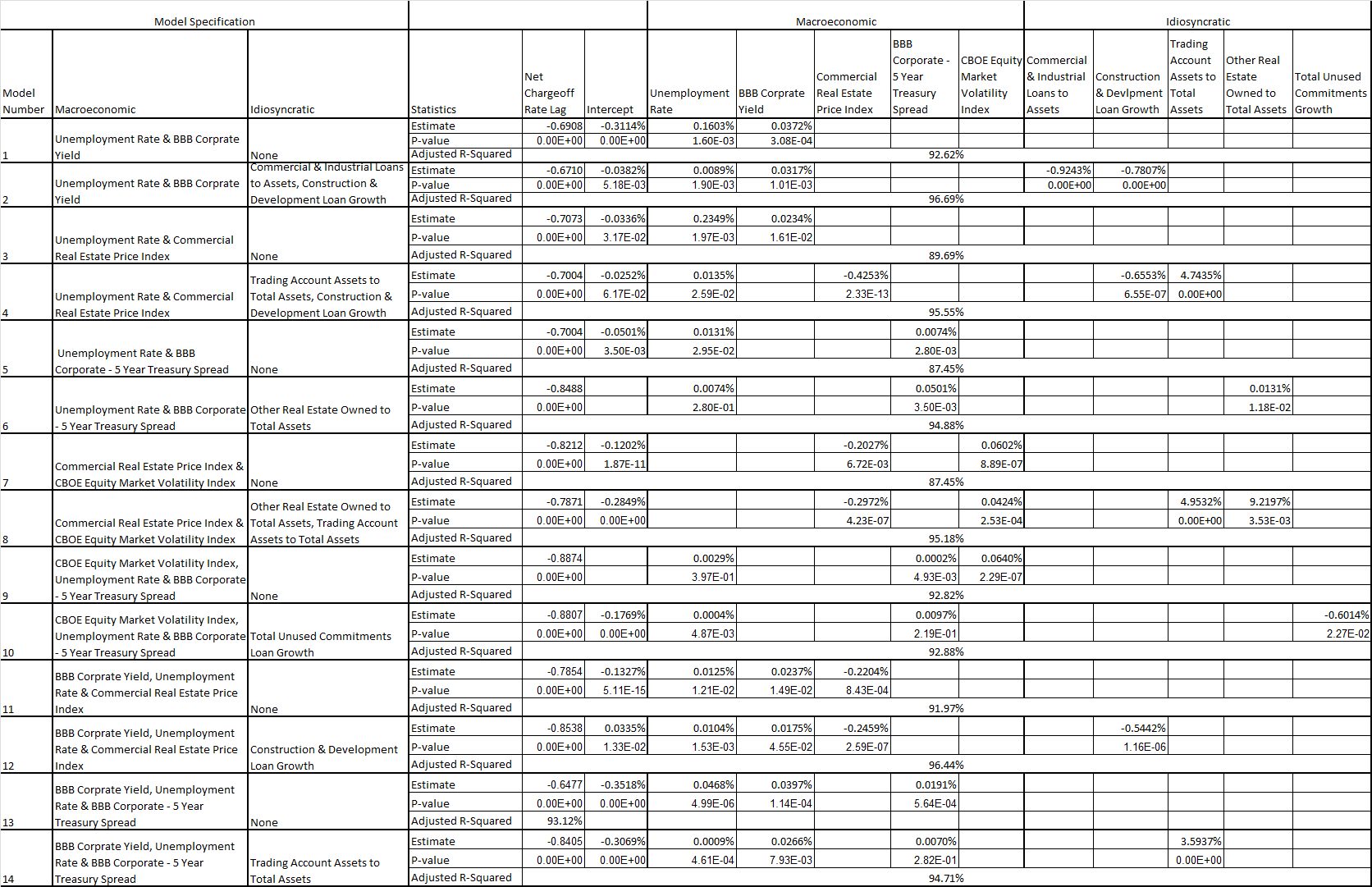
* UNEMP
* BBBCY
* CREPI
* VIX
* CORPSPR

Similarly, we identify the following balance sheet items, banking aggregate idiosyncratic factors, according to the same criteria:

* Commercial and Industrial Loans to Total Assets (“CILTA”)
* Commercial and Development Loans Growth Rate (“CDLGR”)
* Trading Account Assets to Total Assets (“TAATA”)
* Other Real Estate Owned to Total Assets (“OROTA”)
* Total Unused Commitments Growth Rate (“TUCGR”)

This historical data, 65 quarterly observations from 4Q01 to 4Q17[[10]](#footnote-10) are summarized above in Table 1 in terms of distributional statistics and correlations. In Figures 1 and 2 above we show as examples the macroeconomic variable UNEMP and the idiosyncratic variable CILTA, respectively, both the historical time series as well as the 3 scenarios generation model forecasts (Fed, VAR and MS-VAR) for the period 1Q16-4Q17. Across all series, we observe that the credit cycle is clearly reflected, with indicators of economic or financial stress (health) displaying peaks (troughs) in the recession of 2001–2002 and in the financial crisis of 2007–2008, with the latter episode dominating in terms of severity by an order of magnitude. However, there are some differences in timing, extent and duration of these spikes across macroeconomic variables and loss rates. These patterns are reflected in the percent change transformations of the variables as well, with corresponding spikes in these series that correspond to the cyclical peaks and troughs, although there is also much more idiosyncratic variation observed when looking at the data in this form. First, we will describe main features of the distributional statistics of all variables, then the correlations with the variables and the NCORs, followed by the dependency structure within the group of input macroeconomic variables, then the same for the input bank idiosyncratic variables, and finally the cross-correlations between these two groups. We observe that all correlations have intuitive signs and magnitudes that suggest significant relationships, although the latter are not large enough to suggest any issues with multicollinearity. Considering the correlations within and across the sets of macroeconomic and idiosyncratic variables, we note that signs are all economically intuitive, and while magnitudes are material they are not so high in order to result in concerns about multicollinearity.

**Table 2.** Vector Autoregressive CECL Model Estimation Results Compared - Historical Y9 Credit Loss Rates, Banking System Idiosyncratic and Macroeconomic Variables (FDIC SDI Report and Federal Reserve Board 4Q91-4Q15).

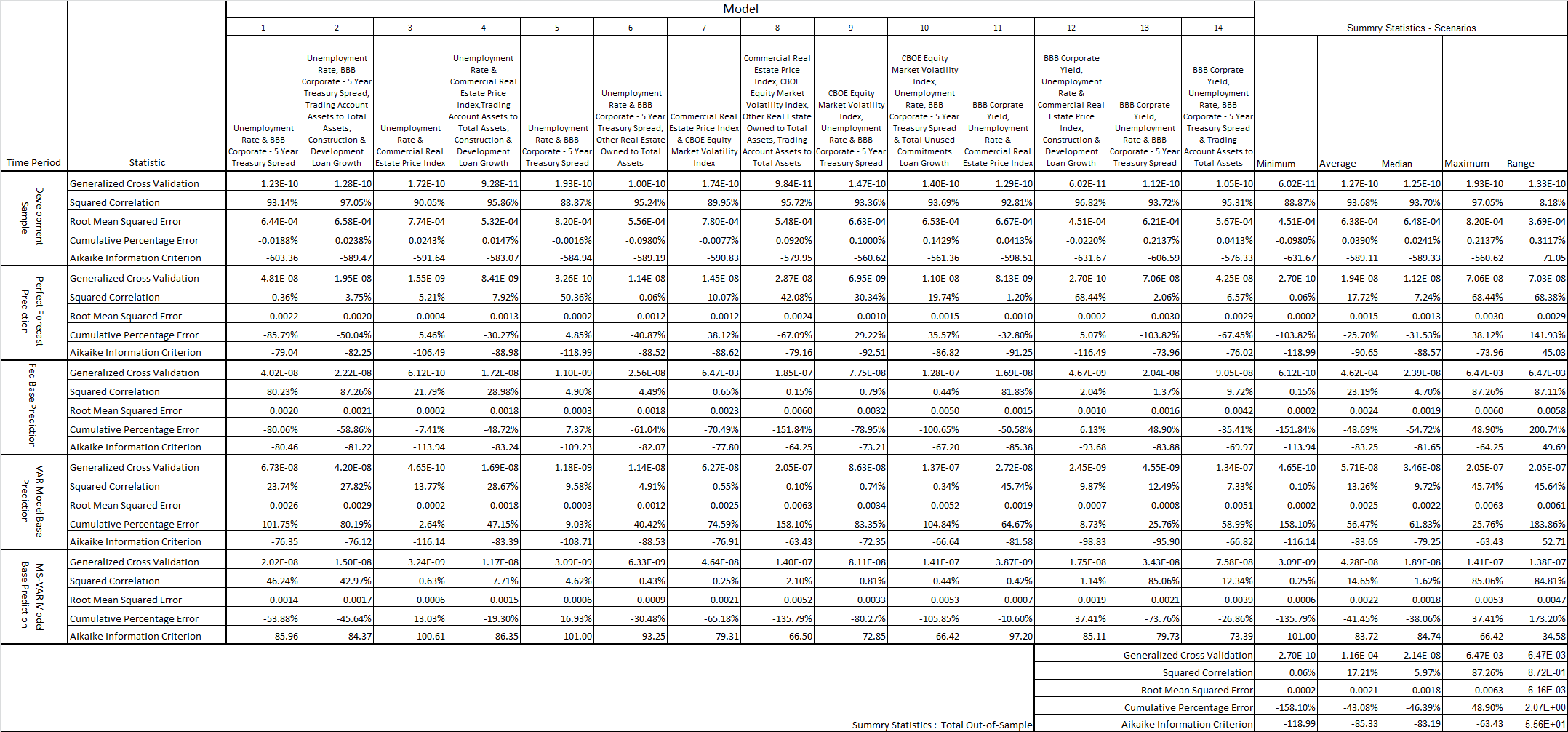


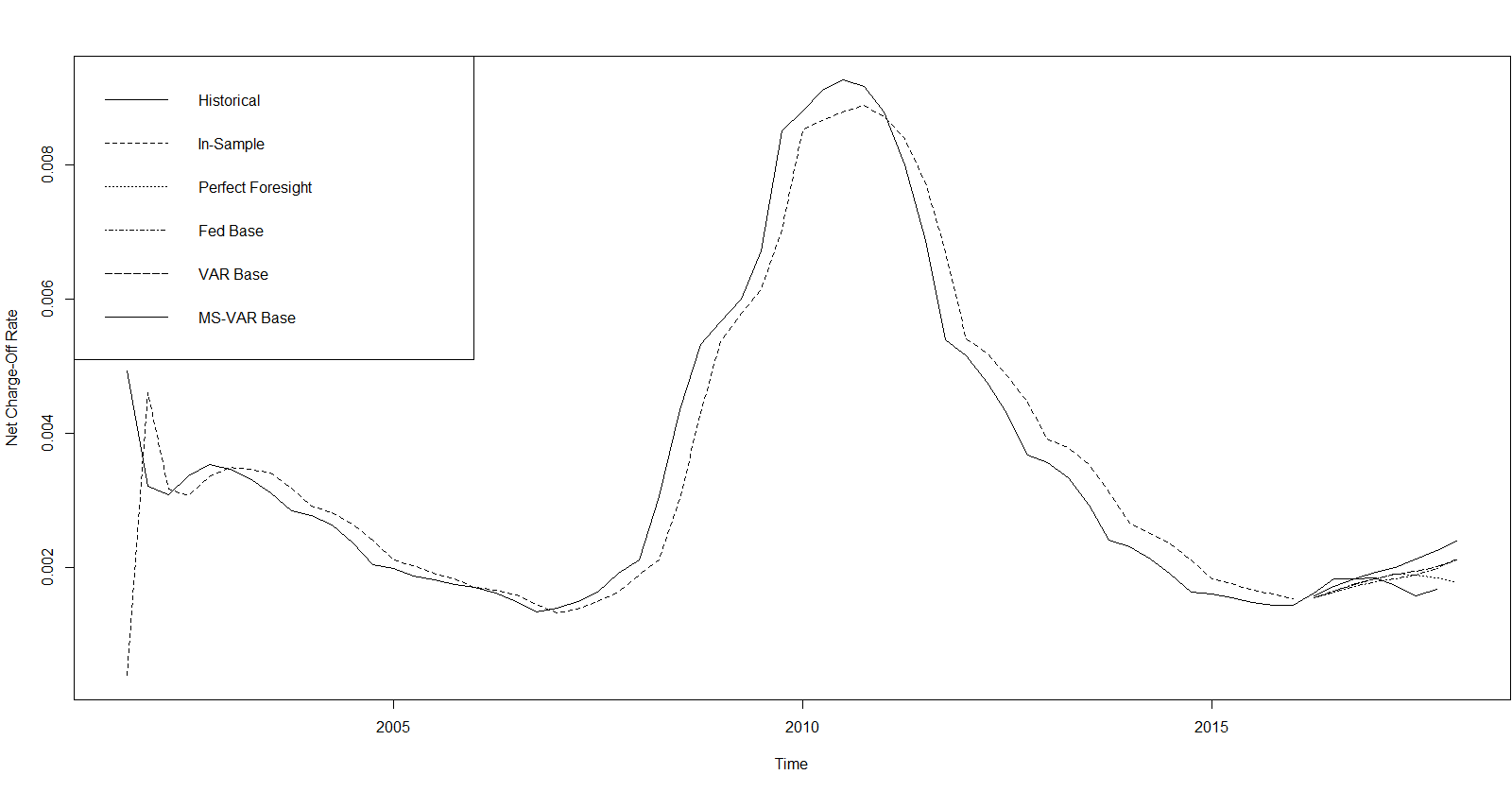
A critical modeling consideration for the MS-VAR estimation is the choice of process generation distributions for the normal and the stressed regimes. As described in the summary statistics of Jacobs (2018), we find that when analyzing the macroeconomic data in percent change form, there is considerable skewness in the direction of adverse changes (i.e., right skewness for variables where increases denote deteriorating economic conditions such as UNEMP). Furthermore, in normal regimes where percent changes are small we find a normal distribution to adequately describe the error distribution, whereas when such changes are at extreme levels in the adverse direction we find that a log-normal distribution does a good job of characterizing the data generating process.[[11]](#footnote-11) Another important modeling consideration with respect to scenario generation is the methodology for partitioning the space of scenario paths across our 10 macroeconomic and idiosyncratic variables for the Base scenario. In the case of the Base scenario, we take an average across all paths in a given quarter for a given variable. The scenarios are shown in Figures 11.10 to 11.14 of the appendix, where we show for each macroeconomic variable the Base scenarios for the VAR and MS-VAR models[[12]](#footnote-12), and also compare this to the corresponding Fed scenarios, along with the historical time series.

The VAR (1) estimation results of our CECL models are summarized above in Table 2. We identify 14 optimal models according to the criteria discussed previously, 7 combinations of macroeconomic variables (4 bivariate and 3 trivariate specifications), as well as versions of these incorporating idiosyncratic variables:

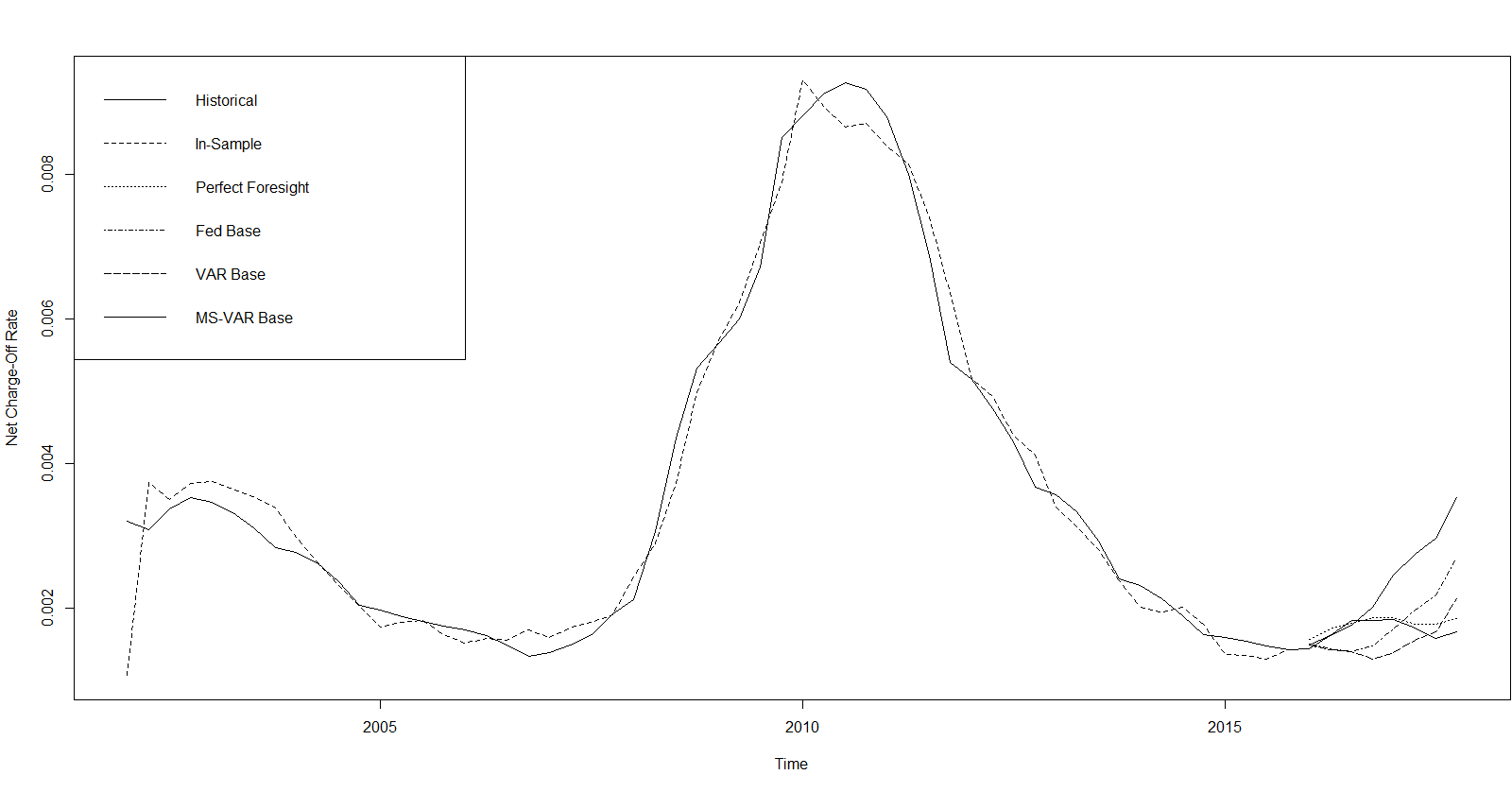
* Model 1: Macroeconomic - UNEMP & BBBCY; Idiosyncratic - none
* Model 2: Macroeconomic - UNEMP & BBBCY; Idiosyncratic - CILTA & CDLGR
* Model 3: Macroeconomic - UNEMP & CREPI; Idiosyncratic - none
* Model 4: Macroeconomic - UNEMP & CREPI; Idiosyncratic - TAAA & CDLGR
* Model 5: Macroeconomic - UNEMP & CORPSPR; Idiosyncratic - none
* Model 6: Macroeconomic - UNEMP & CORPSPR; Idiosyncratic - OROTA
* Model 7: Macroeconomic - CREPI & VIX; Idiosyncratic - none
* Model 8: Macroeconomic - CREPI & VIX; Idiosyncratic - TAAA & OROTA
* Model 9: Macroeconomic - CORPSPR , UNEMP & VIX; Idiosyncratic - none
* Model 10: Macroeconomic - CORPSPR , UNEMP I & VIX; Idiosyncratic - TUCGR
* Model 11: Macroeconomic - BBBCY , UNEMP & CREPI; Idiosyncratic - none
* Model 12: Macroeconomic - BBBCY , UNEMP & CREPI; Idiosyncratic - CDLGR
* Model 13: Macroeconomic - BBBCY , UNEMP & CORPSPR; Idiosyncratic - none
* Model 14: Macroeconomic - BBBCY , UNEMP & CORPSPR; Idiosyncratic - TAAA

**Table 3.** CECL Vector Autoregressive Model In- and Out-of-Sample Performance Measures Compared - Historical Y9 Credit Loss Rates, Banking System Idiosyncratic and Macroeconomic Variables (FDIC SDI Report, Federal Reserve Board 4Q91-4Q15 and Jacobs et al (2018) Models).





**Figure 3** Net-Chargeoff Rate Model Accuracy Plots – Unemployment Rate and BBB Corporate - 5 Year Treasury Bond Spread (FDIC SDI Report, Federal Reserve Board 4Q91-4Q15 and Jacobs et al (2018) Models).



**Figure 4** Net-Chargeoff Rate Model Accuracy Plots – Unemployment Rate, Commercial Real Estate Price Index, BBB Corporate Bond Yield and Commercial Development Loan Growth (FDIC SDI Report, Federal Reserve Board 4Q91-4Q15 and Jacobs et al (2018) Models).As we can see in Table 2, all of the candidate models satisfy our basic requirements of model fit and intuitive sensitivities. Model fit, as measured by Adjusted R-Squared (“AR2”), this metric ranges from 87% to 97% across models, which is good performance by industry standard and broadly comparable. The best fitting model is number 2, the bivariate macroeconomic specification with UNEMP and BBBCY, with idiosyncratic variables CILA and CDLG, having an AR2 of 97.7%. The worst fitting model is number 7, the bivariate macroeconomic specification with CREPI and VIX, with no idiosyncratic variables, having an AR2 of 87.5%. The autoregressive coefficient estimates all show that the NCORs display significant mean reversion, having values ranging from -0.89 to -0.65 and averaging -0.77.

The model performance metrics of the 14 models are above presented in Table 3. We tabulate these statistics for the development sample 4Q01-4Q15, and out-of-sample 1Q16-4Q17, and the latter is evaluated under the three scenario generation models (Fed, VAR as MS-VAR) as well as perfect foresight (i.e., assuming we anticipated the actual paths of the macroeconomic and idiosyncratic variables). We also depict graphically the in- and out-of-sample fits in the model accuracy plots of Figures11.11 through 11.24 of the Appendix. We consider the following industry standard model performance metrics:

* Generalized Cross-Validation (“GCV”)
* Squared Correlation (“SC”)
* Root Mean Squared Error (“RMSE”)
* Cumulative Percent Error (“CPE”)
* Akaike Information Criterion (“AIC”)

We observe from Table 3, as well as from the examples above in Figure 3 (Unemployment Rate and BBB Corporate - 5 Year Treasury Bond Spread) and Figure 4 (Unemployment Rate, Commercial Real Estate Price Index, BBB Corporate Bond Yield and Commercial Development Loan Growth), that there is great diversity of out-of-sample model performance across econometric specifications as well as across models for baseline economic forecasts, both in absolute terms and in comparison to the perfect foresight benchmark. We further note that in summary, out-of-sample performance across models is not assessed as being favorable by industry standards, even in the perfect foresight scenario. Measures of model (GCV, SC, RMSE & AIC) fit tend to be poor, and there is chronic underprediction of NCORs according to the CPE metric. For example, the average SC in-sample is 93.7% ranging in 88.9% to 87.1%, but out-of-sample it averages 17.2% and ranges in 0.06% to 87.3%. On the other hand, in sample the CPE averages 0.04% and ranges in -0.09% to 0.21%, while out of sample the average is -43.0% and ranges in -158.9% to 48.9%.

While this is not conclusive, if we assume that we are near the end of an economic expansion, these results imply that most banks models will under-provision as we enter a downturn, which is evidence that the CECL standard may fail to address the problem of procyclicality that was the intent of the new accounting guidance. Among the loss model specifications, we find that generally the loss prediction specifications having a lower parameterization, as well as the MS-VAR scenario generation model, tends to produce better measures of model fit and a lower degree of under-prediction. Furthermore, relative to the perfect foresight benchmark, the MS-VAR model produces a lower level of variation in the model performance statistics across loss predictive model specifications.

**6 The Quantification of Model Risk According to the Principle of Relative Entropy**

Risk measurement relies on modelling assumptions, the errors in which exposing such models to model risk. In this paper we apply a tool for quantifying model risk and making risk measurement robust to modeling errors. As simplifying assumptions are inherent to all modelling frameworks, the prime directive of model risk management is to assess vulnerabilities to and consequences of model errors. Therefore, a well-designed model risk measurement framework is capable of bounding the effect of model error on specific measures of risk, given a baseline nominal model for measuring risk, as well as identifying the sources of model error to which a measure of risk is most vulnerable, and furthermore isolating which changes in the underlying model have the greatest impact on this risk measure.

In this paper, consistent with the objectives of credit loss measurement in CECL, we focus on the both objectives through calculating an upper bound on the range of credit risk values that can result over a range of model errors within a certain distance of a nominal model, for a range of credit loss model and economic scenario generation models. This bound is somewhat analogous to an upper confidence bound, but whereas a confidence interval quantifies the effect of sampling variability, the robustness bound that we develop quantifies the effect of model error. The simple example of a standard error estimate should help illustrate this idea, as a conventional measure of credit risk in a loan or bond portfolio. Measuring standard deviation prospectively requires assumptions about the joint distribution of the returns of assets, or default correlation, in a credit portfolio.

In light of the first objective listed above and our focus in the CECL context, we would want to bound the values of standard deviation that can result from a reasonable degree of model error. In practice, model risk is sometimes addressed by comparing the results of different models, but more often if it is considered at all, model risk is investigated by varying model parameters. Crucially, the tools applied here go beyond parameter sensitivity to consider the effect of changes in the probability law that defines an underlying model, enabling us to identify vulnerabilities to model error that are not reflected in parameter perturbations. For example, the main source of model risk might result from an error in a joint distribution of returns that cannot be described through a change in a covariance matrix. To work with model errors described by changes in probability laws, we need a way to quantify such changes, and to this end we deploy the *principle of relative entropy* (Hansen and Sargent (2007), Glasserman and Xu (2013).) In Bayesian statistics, the relative entropy between posterior and prior distributions measures the information gained through additional data. In characterizing model error, we interpret relative entropy as a measure of the additional information required to make a perturbed model preferable to a baseline model. Thus, relative entropy becomes a measure of the plausibility of an alternative model. It is also a convenient choice because the worst-case alternative within a relative entropy constraint is typically given by an *exponential change of measure*.

We quantify the model risk with respect to a champion, or null, model  such that the *Kullbck-Leibler Relative Entropy* divergence measure from a challenger, or reference, model  is given by:

 (15)

In this construct, the  mappings are our set of estimated CECL loss distribution models, and the  benchmark.is some kind of alternative, such as the perfect foresight loss forecast. We can define the *likelihood ratio*  characterizing our modeling choice according to the relationship:

 (16)

It is standard in the literature to express (9) in terms as an equivalent expectation of a relative deviation in likelihood:

 (17)

Where  represents a relatively small upper bound on model risk deviations, dictated by the model risk appetite of the organization with respect to a particular model type (e.g., a model performance threshold). A key property of relative entropy is that and only if. Given a set of reference models and a relative distance measure, the solution the model error can be quantified by the following change of numeraire (Glasserman and Xu, 2013):

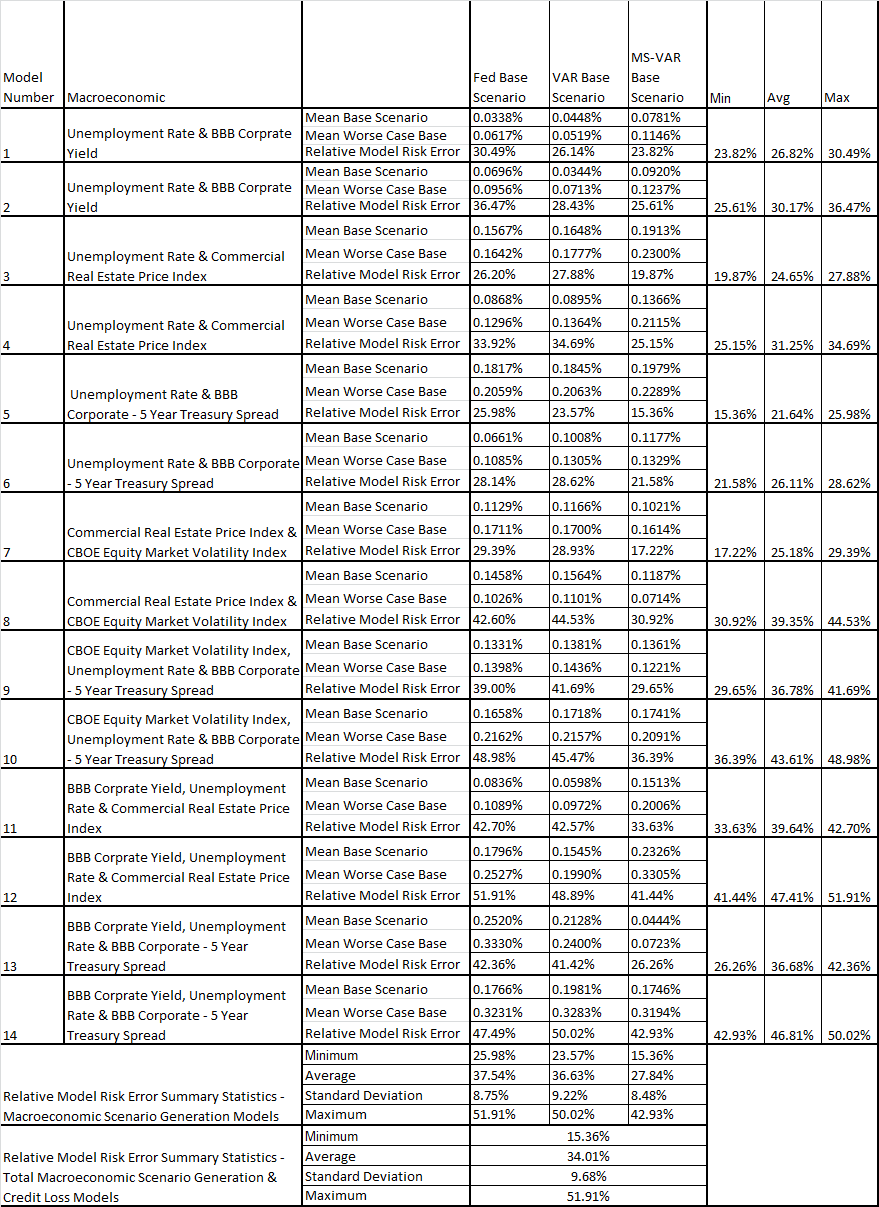
 (18)

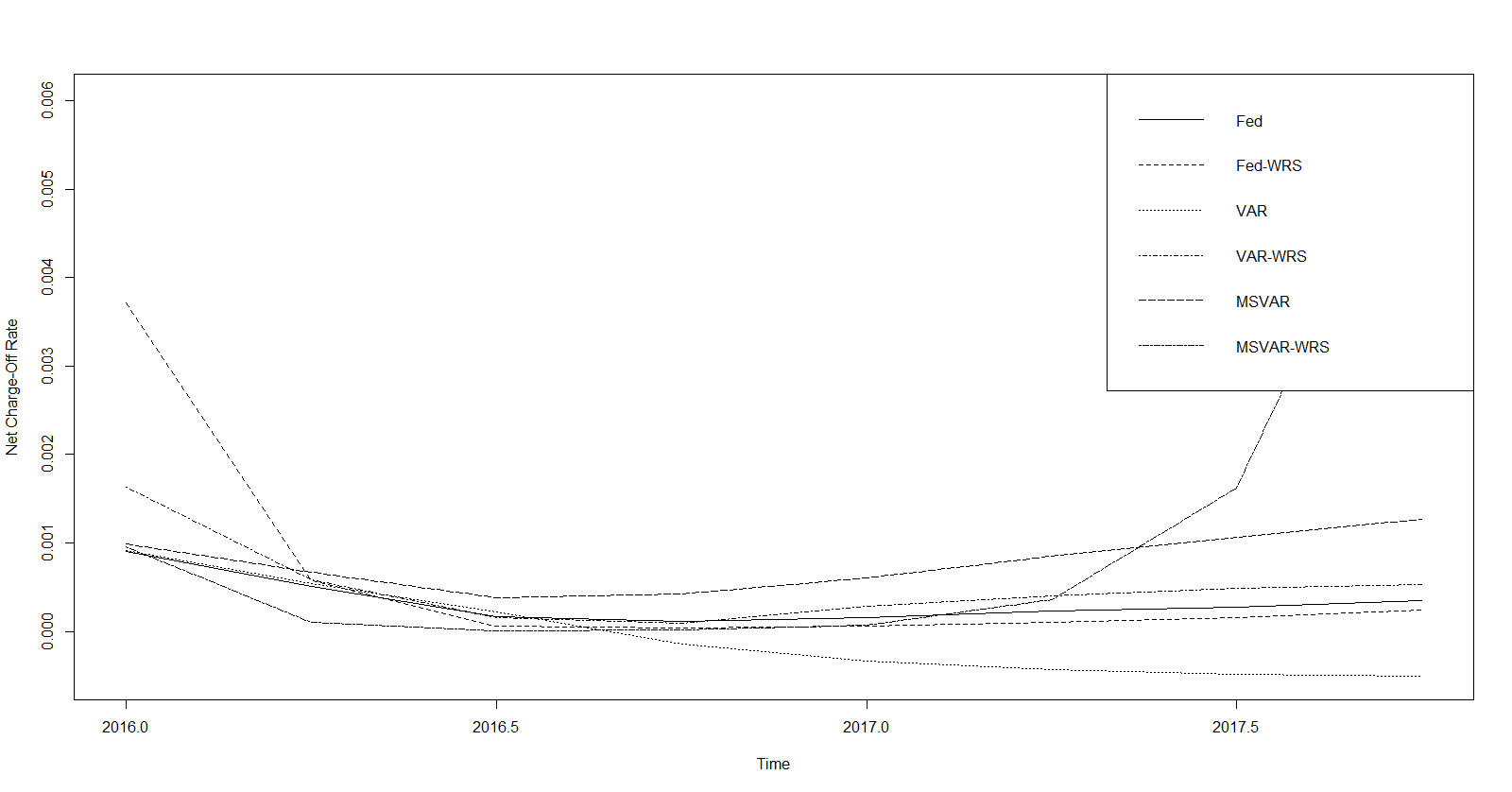
Where (25) is the solution or inner supremum to the optimization problem:

 (19)

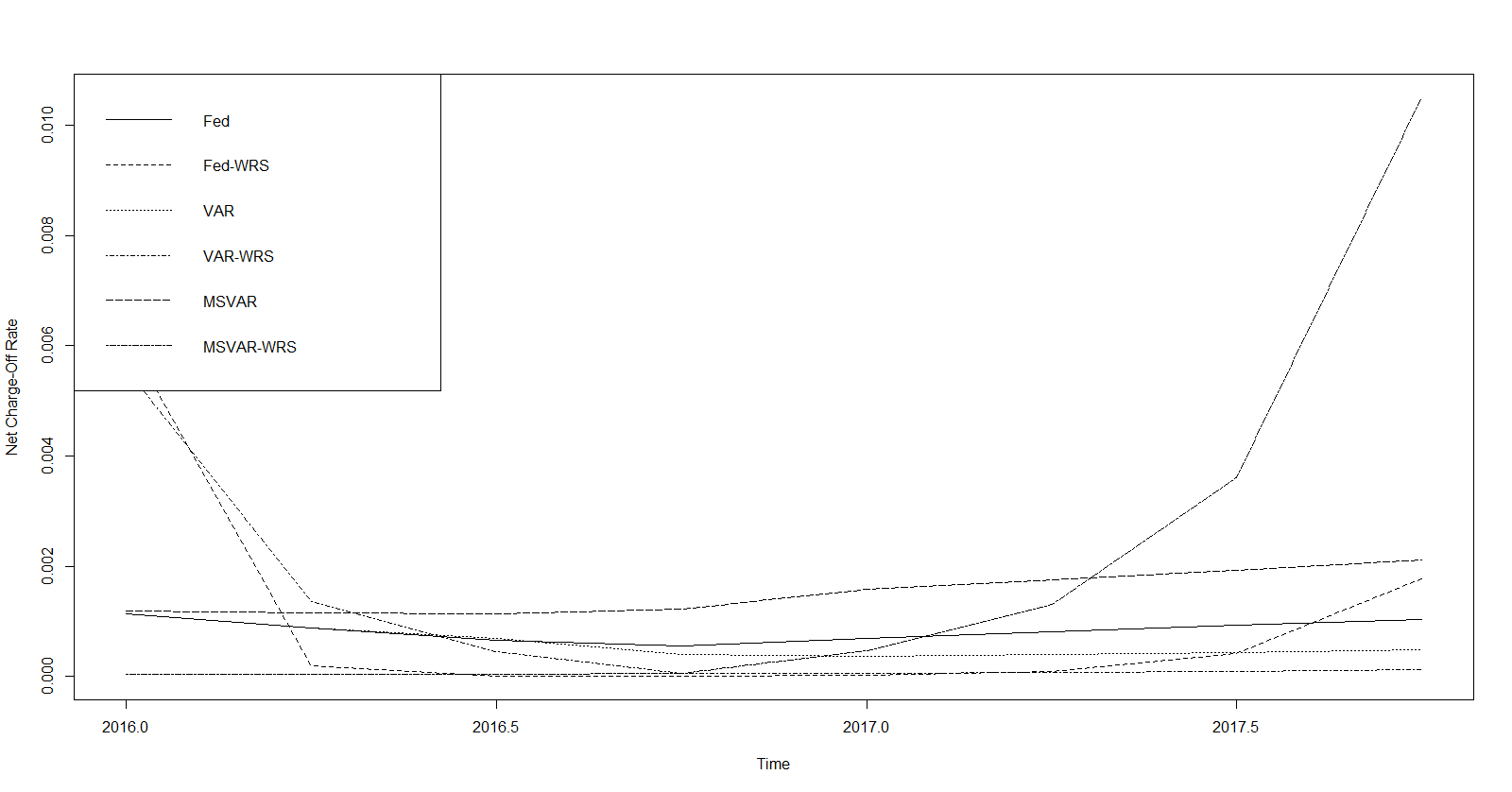
In (19) the model risk measure is parametricized by  , such that  corresponds to the best case of zero model risk, and  the worst case of maximal model risk. An important

**Table 4.** Kullback-Leibler Relative Entropy Worse Case Loss for Model Risk Quantification Measures - Vector Autoregressive CECL Models Out-of-Sample Compared (FDIC SDI Report, Federal Reserve Board 4Q91-4Q15 and Jacobs et al (2018) Models).





**Figure 5** Kullback-Leibler Relative Entropy Worse Case Loss for Model Risk Quantification Plot – Unemployment Rate and BBB Corporate Bond Yield (FDIC SDI Report, Federal Reserve Board 4Q91-4Q15 and Jacobs et al (2018) Models).



**Figure 6** Kullback-Leibler Relative Entropy Worse Case Loss for Model Risk Quantification Plot – Unemployment Rate, CBOE Equity Volatility Index, BBB Corporate - 5 Year Treasury Bond Spread and Total Uncommitted Loan Growth (FDIC SDI Report, Federal Reserve Board 4Q91-4Q15 and Jacobs et al (2018) Models).

property of the change of measure (16) is that it is *model free*, or independent of the reference model. This is critical from a model validation perspective, as this implies that this procedure is robust to model mispecification.

We summarize the empirical implementation results of our model risk quantification in Table 4 above and show examples in Figure 5 (Unemployment Rate and BBB Corporate Bond Yield) and Figure 6 (Unemployment Rate, CBOE Equity Volatility Index, BBB Corporate - 5 Year Treasury Bond Spread and Total Uncommitted Loan Growth) above. In Table 4 we tabulate the mean CECL loss under the model, the same in the worst case, and the *relative model risk error* (“RMRE”) measure defined as:

(20)

Where  is a CECL loss estimate for the vector of macroeconomic variables  in loss model  (our 14 models VAR NCO models) and economic scenario generation model  (Fed, VAR and MS-VAR), and  is the worst case version at time  in the forecast period, so this represents an empirical forecast error metric where the expectation with respect to the null model is replaced by a sample average.

Table 4 shows that across credit loss and macroeconomic scenario generation models, the average RMRE 34.01% is substantial, and varying widely across both credit loss and macroeconomic scenario generation specifications from 15.4% to 51.9%. Considering scenario generation frameworks, we observe that the MS-VAR model has a consistently lower RMSE measure as compared to the Fed or VAR models, averaging 27.8% in the former as compared to 27.5% and 36.6%, respectively, in the latter. Another pattern that we observe is that, in the majority of cases, credit loss models having either more macroeconomic variables, or include idiosyncratic factors in addition to a set of macroeconomic factors, have higher model risk measures. The 2-variable credit loss models with no idiosyncratic variables have RMSEs ranging in 21.6% to 25.2.4%, while those having 3 macroeconomic factors have a range of 36.7% to 39.6%. On the other hand, among the version of these models having idiosyncratic variables, the 2-variable credit loss models have RMSEs ranging in 24.7% to 39.4%, while those having 3 macroeconomic factors have a range of 43.6% to 47.4%.

There is a profound model risk management policy implication of this analysis that speaks to our second objective in measuring model risk, namely identifying the sources of vulnerability in assumptions that give rise to model risk. There are two sources at play herein – the joint distribution of the macroeconomic and idiosyncratic variables, and the assumptions on the error terms in the joint distribution of losses, and with respect to the latter whether they are Gaussian (i.e., Fed or VAR) or follow a heavy-tailed distribution (i.e., MS-VAR). We conclude from these results that the less parsimonious are the models, the greater risk is there of model mis-specification that manifests in the higher RMRE measures. On the other hand, while we may think that the additional parameter of the MS-VAR model would give rise to more model risk, the fact that this model is better capable of modeling the fat-tailed distribution of credit losses is realized in a lower MS-VAR measure, regardless of the credit loss model specification. This should be no surprise, as we saw that the Fed and VAR models had rather more egregious under-prediction than the MS-VAR model, which is at odds with the historical distribution of losses. On the other hand, while the credit models with more variables fit only slightly better on an in-sample basis, most exhibited relatively poor performance out-of-sample. The conclusion is that practitioners may wish to err on the side of more parsimonious models that can also accommodate non-normality.

**7 Conclusion and Future Directions**

In this study we have presented an holistic model validation framework for CECL model development and model implementation execution, and have also illustrated elements of CECL model development validation, including the quantification of model risk. In our example of the latter validation of CECL model development, we have analyzed the impact of model specification and scenario dynamics upon expected credit loss estimates in CECL, through implementing a highly stylized framework borrowed from the ST modeling practice. We performed a model selection of alternative CECL specifications in a top-down framework, using FDIC FR-Y9C data and constructing an aggregate or average hypothetical bank, with the target variable being NCORs and the explanatory variables constituted by Fed provided macroeconomic variables as well as bank-specific controls for idiosyncratic risk. We studied not only the impact of the ALLL estimate under CECL for alternative model specifications, but also the impact of different frameworks for scenario generation: the Fed baseline assumption, a Gaussian VAR model and a mixture-of-distributions MS-VAR model, following the study of Jacobs et al (2018).

The implications for CECL model validation and governance is our results suggest that in general the CECL methodology is at risk of not achieving the stated objective of reducing the pro-cyclicality of provisions relative to the legacy incurred loss standard, as across models we observe chronic underprediction of losses in the last 2-year out-of-sample period, which arguably is a period that is late in the economic cycle. Furthermore, we have illustrated that the amount of such procyclicality exhibits significant variation across model specifications and scenario generation frameworks. In general, we have found that the MS-VAR scenario generation framework produces the best performance in terms of fit and lack of under-prediction relative to the perfect foresight benchmark, which is in line with the common industry practice of giving weight to adverse but probable scenarios, which the MS-VAR regime switching model can produce naturally and coherently as part of the estimation methodology that places greater weigh on the economic downturn. We have also found that for any scenario generation model, across specifications the more lightly parameterized models tend to have better out of sample performance. Furthermore, relative to the perfect foresight benchmark, the MS-VAR model was found to have produced a lower level of variation in the model performance statistics across loss predictive model specifications.

As a second and related exercise, we quantified the level of model risk in this hypothetical CECL exercise, using the principle of *relative entropy*. We found that more elaborate modeling choices, such as more highly parametricized models in terms of more macroeconomic or idiosyncratic covariates, tend to introduce more measured model risk. However, the more highly parametricized macroeconomic scenario generation framework that can accommodate heavy tails in the credit loss distribution (the MS-VAR model, which has separate parameters for normal and stressed conditions), tends to introduce less measured model risk than Gaussian approaches (Fed or VAR models). The implication is that while banks may wish to err on the side of more parsimonious approaches in order to manage the increase model risk that the introduction of the CECL standard gives rise to, they are advised to balance this against the need to model the non-normality in the credit loss distribution.

The policy implication of this analysis is that the volume of lending and the amount of regulatory capital held may vary greatly across banks, even when it is the case that the respective loan portfolios have very similar risk profiles. Another consequence of this divergence of expected loan loss estimates under the CECL standard is that supervisors and other market participant stakeholders may face challenges in comparing banks at a point of time or over time. There are also implications for the influence of modeling choices in specification and scenario projections on the degree of model risk introduced by the CECL standard.

There are several directions in which this line of research could be extended:

* More granular classes of credit risk models, such as ratings migration or PD/LGD scorecard/regression
* Alternative data-sets, for example bank or loan level data
* More general classes of regression model, such as logistic, semi-parametric or machine learning / artificial intelligence techniques (Jacobs, 2018)
* Applications related to stress testing, such as regulatory or economic capital

**8 References**

Basel Committee on Banking Supervision (2006) ‘International Convergence of Capital Measurement and Capital Standards: A Revised Framework’, The Bank for International Settlements, Basel, Switzerland.

Basel Committee on Banking Supervision (2009) ‘Principles for Sound Stress Testing Practices and Supervision - Consultative Paper No. 155’, The Bank for International Settlements, Basel, Switzerland.

Batchelor, R. A., & Dua, P. (1990). Product differentiation in the economic forecasting industry. *International Journal of Forecasting 6(3)*, 311-316.

Bernanke, B. S., & Lown C. S. (1991). The credit crunch. *Brookings Papers on Economic Activity 2*, 205-247.

Berrospide, J., & Edge, R. (2010). The effects of bank capital on lending: What do we know, what does it mean. *International Journal of Central banking 6(4)*, 5-55.

Box, G., & Jenkins, G. (1970). *Times series analysis: forecasting and control*. San Francisco,CA: Holden-Day.

Board of Governors of the Federal Reserve System. (2011). ‘Supervisory Guidance on Model Risk Management’, Supervisory Letter 11-7, Washington, D.C., April 4th.

Brockwell, P.J., & Davis, R.A. (1991). *Time series: theory and methods.* New York, N.Y.: Springer-Verlag.

Carlson, M., Shan, H., & Warusawitharana, M. (2013). Capital ratios and bank lending: A matched bank approach, *Journal of Financial Intermediation 22*, 663-687.

Chae, S. Sarama, R. Vocjtech, C., &.Wang, J. (2018). The impact of the current expected credit loss standard (CECL) on the timing and comparability of reserves, *Finance and Economics Discussion Series 2018-020*.

Commandeur, J. J. F., & Koopman, S.J. (2007). *Introduction to state space time series analysis*. New York, N.Y: Oxford University Press.

Cornett, M., McNutt, J., Strahan, P., & Tehranian, H. (2011). Liquidity risk management and credit supply in the financial crisis, *Journal of Financial Economics 101*, 297-312.

Deloitte. (2017). CECL implementation insights.

Financial Accounting Standards Board (1975) “Accounting Standards Update No. 1975-5, Accounting for Contingencies,” December.

Financial Accounting Standards Board (1993) “Accounting Standards Update No. 1993-114, Accounting by Creditors for Impairment of a Loan,” December.

Financial Accounting Standards Board (2012) “Accounting Standards Update No. 2012-260, Financial Instruments—Credit Losses (Subtopic 825-15): Measurement of Credit Losses on Financial Instruments,” December.

Financial Accounting Standards Board (2016) “Accounting Standards Update No. 2016-13, Financial Instruments—Credit Losses (Topic 326): Measurement of Credit Losses on Financial Instruments,” June.

Francis, W.B., & Osborne, M. (2009). Bank regulation, capital and credit supply: measuring the impact prudential standards, *Occasional Paper 36, Financial Services Authority*.

Glasserman, P., & Xu, X. (2013). Robust risk measurement and model risk. *Quantitiative Finance, 14(1),* 29-58.

Global Credit Data. (2019). Current Expected Loss (CECL) accounting: benchmarking study.

Hanan, E.J. (1971). The identification problem for equation systems with moving average errors. *Econometrica 39,* 751-766.

Hanan, E.J. (1988). The statistical theory of linear systems. New York, N.Y.: John Wiley.

Hansen, L.P., & Sargent, T.J., (2007). *Robustness*. Princeton, N.J.: Princeton University Press.

Hirtle, B., Kovner, A., Vickery, J., &.Bhanot, M. (2015). Assessing financial stability: the capital and loss assessment under stress scenarios (CLASS) model. *Federal Reserve Bank of New York Staff Report No. 663 (July)*.

International Accounting Standards Board. (2014). “International Reporting for Financial Statement Number 9”, July.

Jacobs, Jr., M., 2015, The quantification and aggregation of model risk: perspectives on potential approaches, *The Journal of Financial Engineering and Risk Management 2(2)*, 124–154.

Jacobs, Jr., M., Klein, L., & Merchant, A., 2015a (September), Emerging trends in model risk management. *Accenture Consulting*.

Jacobs, Jr., M., Karagozoglu, A.K , & Sensenbrenner, F.J. (2015b). Stress testing and model validation: application of the Bayesian approach to a credit risk portfolio. *The Journal of Risk Model Validation, 9(3)*, 41–70.

Jacobs, Jr., M.. (2017). A mixture of distributions model for the term structure of interest rates with an application to risk management. *American Research Journal of Business and Management, 3(1)*, 1-17.

Jacobs, Jr., M., & Sensenbrenner, F.J. (2018a). A comparison of methodologies in the stress testing of credit risk – alternative scenario and dependency constructs. *Quantitative Finance and Economics 2(2)*, 294–324.

Jacobs, Jr., M. (2018b). The validation of machine learning models for the stress testing of credit risk. *The Journal of Risk Management in Financial Institutions 11(3)*, 1-26.

Jacobs, Jr., M. (2019). An analysis of the impact of modeling assumptions in the Current Expected Credit Loss (CECL) framework on the provisioning for credit loss. *The Journal of Risk and Control 6(1)*, 65–113.

Jacobs, Jr., M. (2019). The accuracy of alternative supervisory methodologies for the stress testing of credit risk. (August). Forthcoming *The Journal of Financial Engineering and Risk Management*.

Joaquin M., & Mukerji, J. (2003). MDA Guide. Object Management Group, Inc., Version 1.0.1., June.

Jouault, F.: Loosely Coupled Traceability for ATL In: Proceedings of the European Conference on Model Driven Architecture (ECMDA) workshop on traceability, Nuremberg, Germany (2005).

Kishan, R., and Opiela, T (2000). Bank size, bank capital, and the bank lending channel. *Journal of Money, Credit, and Banking 32*, 121-141.

Mellor, S., & Balcer, M. (2002). Executable UML A Foundation for Model-Driven Architecture. Indianapolis , I.N.: Addison-Wesley.

Oliver Wyman. (2017). Current Expected Loss (CECL) accounting: it’s time to get moving.

R Development Core Team, 2019: “R: A Language and Environment for Statistical Computing.” R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

SAS Institute. (2015). The current state of banking stress testing technology. *Whitepaper No. 107977*.

Sims, C.A. (1980). Macroeconomics and reality. *Econometrica* *48*, 1-48.

Skoglund, J., 2018 (April), "Quantification of Model Risk in Stress Testing and Scenario Analysis", *SAS Institute*.

Stock, J.H., &. Watson, M.W. (2001). Vector autoregressions. *Journal of Economic Perspectives* *15(4)*, 101-115.

1. Corresponding author: Michael Jacobs, Jr., Ph.D., CFA, Lead Quantitative Analytics & Modeling Expert, Head – C&I 1st Line Model Risk Management, PNC Financial Services Group – Balance Sheet Analytics & Modeling / Model Development, 340 Madison Avenue, New York, N.Y., 10022, 917-324-2098, [michael.jacobsjr@pnc.com](mailto:michael.jacobsjr@pnc.com). The views expressed herein are those of the author do not necessarily represent an official position of PNC Financial Services Group. [↑](#footnote-ref-1)
2. The international counterpart to CECL is codified in International Accounting Standards Board (2014; “IRFS9”), which while bearing similarities to CECl, in that the component for performing loans is an expected loss construct incorporating macroeconomic forecasts, has several fundamental differences. [↑](#footnote-ref-2)
3. These are given by FASB (1975) and FASB (1993) for performing and impaired instruments, respectively. [↑](#footnote-ref-3)
4. Several challenges are associated with macroeconomic forecasting, related to changes in the structure of the economy, measurement errors in data as well as behavioral biases (Batchelor and Dua, 1990). [↑](#footnote-ref-4)
5. Refer to Stock and Watson (2001) for a discussion of the basic aspects of macroeconomic forecasting (i.e., characterization, forecasting, inferences and policy advice regarding macroeconomic time series and the structure of the economy.) [↑](#footnote-ref-5)
6. In fact, the exogenous variables  can represent both stochastic and non-stochastic (deterministic) variables, examples being sinusoidal seasonal (periodic) functions of time, used to represent the seasonal fluctuations in the output process, or intervention analysis modelling in which a simple step (or pulse indicator) function taking the values of 0 or 1 indicates the effect of output due to unusual intervention events in the system. [↑](#footnote-ref-6)
7. The full model development code package, an R project containing all data and code, is available upon request. [↑](#footnote-ref-7)
8. These are available at <https://www5.fdic.gov/sdi/main.asp?formname=standard> [↑](#footnote-ref-8)
9. We perform this model selection in an R script designed for this purpose, using the libraries “dse” and “tse” to estimate and evaluate VAR and MS-VAR models (R Core Development Team, 2019). [↑](#footnote-ref-9)
10. We leave out the last 2 years of available data, 1Q16–4Q17, in order to have a holdout sample for testing the accuracy of the models. We also choose to start our sample in 2001, as we believe that the earlier period would reflect economic conditions not relevant for the last decade and also because in the financial industry this is a standard starting point for CCAR and DFAST stress testing models. [↑](#footnote-ref-10)
11. This is similar to the findings of Loregian and Meucci (2016) and Jacobs (2017a) in the context of modelling U.S. Treasury yields. We observe that this mixture well characterizes the empirical distributions of the data in this paper. [↑](#footnote-ref-11)
12. Estimation results for the VAR and MS-VAR model are available upon request. The models are all convergent and goodness of fit metrics in with industry standards. Signs of coefficient estimates are in line with economic intuition and estimates are all significant at conventional levels. We use the dse, tseries and MSBVAR libraries in R in order to perform the estimations (R Development Core Team, 2019). [↑](#footnote-ref-12)