Journal of Risk Model Validation **15**(3), 1–39
DOI: 10.21314/JRMV.2022.023

Research Paper

Quantification of model risk with an application to probability of default estimation and stress testing for a large corporate portfolio

Michael Jacobs Jr.

PNC Financial Services Group, 340 Madison Avenue, New York, NY 10022, USA;
email: michael.jacobsjr@pnc.com

(Received November 1, 2021; revised May 19, 2022; accepted August 12, 2022)

ABSTRACT

This paper addresses the building of obligor-level hazard rate corporate probability of default models for stress testing, departing from the predominant practice in wholesale credit modeling of constructing segment-level models for this purpose. We build models based upon varied financial, credit rating, equity market and macroeconomic factors, using an extensive history of large corporate firms sourced from Moody's. We develop distance-to-default (DTD) risk factors and design hybrid structural/Merton reduced-form models as challengers to versions of the models containing other individual variables. We measure the model risk attributed to various modeling assumptions according to the principle of relative entropy and observe that the omitted-variable bias with respect to the DTD risk factor, neglect of interaction effects and incorrect link function specification has the greatest, intermediate and least impact, respectively. Given the sensitive regulatory uses of these models and the concerns raised in the industry about the effect of model misspecification on capital and reserves, our conclusion is that validation methods chosen in the stress testing context should be capable of testing model assumptions. Our research adds

Print ISSN 1753-9579 | Online ISSN 1753-9587

to the literature in that it offers state-of-the-art techniques as viable options in the arsenal of model validators, developers and supervisors seeking to manage model risk.

Keywords: probability of default (PD); stress testing; Current Expected Credit Loss (CECL); credit risk; model validation; model risk.

1 INTRODUCTION AND SUMMARY

The importance of stress testing in assessing the credit risk of bank loan portfolios has grown over time. Currently, these exercises are accepted as the primary means of supporting capital planning, business strategy and portfolio management decision-making (Financial Services Authority 2008). Such analysis gives us insight into the likely magnitude of losses in an extreme but plausible economic environment, conditional on varied drivers of loss. It follows that such activity enables the computation of unexpected losses that can inform regulatory or economic capital according to Basel III guidance (Basel Committee on Banking Supervision 2011).

Most recently in this domain, in satisfaction of the Current Expected Credit Losses (CECL) accounting standards (Financial Accounting Standards Board 2016),[1] or in compliance with the Federal Reserve's Dodd–Frank Act Stress Test (DFAST) program (Board of Governors of the Federal Reserve System 2016),[2] we observe that the predominant types of models used in the industry differ slightly from those used in the context of the 2007–9 global financial crisis, as their application must meet particular capital adequacy and accounting requirements that were not previously a consideration (Global Credit Data 2019).

In this study, we quantify the degree of model risk due to several forms of model misspecification or violations of model assumptions by utilizing the principle of relative entropy. This methodology studies the distance of an alternative model to a reference model according to some suitable loss metric (see Hansen and Sargent 2007; Glasserman and Xu 2014) and can capture the dimensions of model uncertainty error beyond parameter estimation error. This framework for measuring model risk is applied to a CECL stress testing exercise of default risk for a corporate portfolio. The importance of this application is rooted in the sensitivity of CECL or DFAST results from the perspective of prudential supervision as well as of accounting policy. It is observed that third-party reviewers, such as model validators or regulators, often

[1] See Jacobs (2015) for a conceptual framework addressing the measurement of model risk and Jacobs (2020) for the quantification of model risk in the context of a top-down credit model for CECL.

[2] See Jacobs (2019) for an assessment of the accuracy of alternative supervisory methodologies for DFAST stress testing.

question the impact of faulty model assumptions on capital and reserve projections. We find that omitted-variable bias due to leaving out a critical risk factor has the most impact, that incorrect specification of the model equation function is of minimal importance and neglected interaction terms among the explanatory variables are of intermediate influence for the measured model risk in terms of forecast bounds.

To elaborate on this finding, the standard manner in which credit models have been adapted for stress testing is through the modification of probability of default (PD) models at the disposal of financial market participants. PD models are designed to accurately measure an obligor's ability and willingness to meet future debt obligations over some horizon, and are typically associated with a credit score or rating. The majority of PD risk rating methodologies or models currently used in the industry are characterized by a dichotomy of outcomes: point-in-time (PIT) versus through-the-cycle (TTC). In the so-called PIT rating philosophy such PD models should incorporate a complete set of borrower-specific and macroeconomic risk factors that will measure default risk at any point in the economic cycle. In contrast, according to the TTC rating philosophy, the model should abstract from the state of the economy or from cyclical effects and measure default risk over a more extended time horizon that incorporates a variety of macroeconomic states. This TTC orientation implies that ratings derived from the model should show "stability", wherein material changes in ratings can be ascribed to fundamental, as opposed to transient, factors. PIT PD models are typically deployed in loan pricing and early warning systems, while TTC PD models feature prominently in regulatory capital, credit underwriting and portfolio management applications.

Note that this distinction prevails in particular for wholesale credit asset classes (eg, large corporate or middle market commercial and industrial (C&I) loans), where instead of using PIT PD models directly for stress testing, alternative approaches that are prevalent in the industry are used. One such common methodology involves adding sensitivity to macroeconomic variables in the TTC PD models (Global Credit Data 2019). Such TTC PD models are commonly found in the Basel III framework or for use in credit underwriting, as previously mentioned. The predominant manner in which TTC PD models are used in stress testing is through a rating transition model construct (Gross *et al* 2020), where the level of modeling is at the rating level and credit ratings are aggregated for different modeling segments across a bank's portfolio.

These considerations point us to an overarching conceptual question regarding stress testing and how such models are validated. Demonstrating the fitness for purpose of downstream models that are dependent upon upstream credit risk models is a challenge within the industry. Often, model development teams will find it infeasible to test the downstream fitness of upstream models during redevelopment, and it could be the case that they become subject to overarching validation issues that

are not resolved until the next redevelopment cycle. This could then involve lags of multiple years and be costly in terms of enhanced model monitoring requirements. In the case of wholesale portfolios, in which it is more common to have this disconnect, this leads to challenges in demonstrating the conceptual soundness of credit risk and stress testing models to model validators and supervisors (Global Credit Data 2019). The type of PD models that we investigate in this paper address this issue, as they are directly applicable to stress testing. To the best of our knowledge, there is no academic or practitioner literature addressing this issue that is particular to the wholesale credit asset class, which is where our research presents its main contribution to the literature.

The position of this research in the academic literature is at the intersection of two streams of inquiry. First, there are a number of empirical studies that focus on the factors that determine corporate default and its forecasting (see, for example, Altman 1968, Jarrow and Turnbull 1995 and Duffie and Singleton 1999). At the other end of the spectrum, there are mainly theoretical studies that focus on modeling frameworks for either understanding corporate default (see, for example, Merton 1974) or perspectives on the TTC versus PIT dichotomy (see, for example, Aguais *et al* 2008). In this paper, we blend these considerations of theory and empirics while also addressing the previously discussed conceptual validation issue facing practitioners in stress testing for wholesale portfolios.

This research is further distinguished from the existing literature by its utilization of an obligor-level and dynamic modeling framework that considers financial, credit rating and macroeconomic variables that are time varying.[3] We estimate these models over a history that contains several economic cycles and apply them to a CECL stress testing exercise. We implement this exercise through the construction of discrete-time hazard rate models of default, a class of dynamic PD models, by utilizing a data set of corporate ratings and defaults. This methodology has the benefit of accommodating the discrete character of our data (which are quarterly snapshots), and while the use of discrete-time survival models appeared previously in the prediction of corporate defaults, prior studies have not incorporated macroeconomic risk factors or been applied to stress testing (Shumway 2001; Cheng *et al* 2010). This is, to the best of our knowledge, the first study to do so. Finally, we employ what is, compared with other models in this class, a less complex estimation algorithm. While the computational advantages of our data design approach may not be the primary benefit of our methodology, given the availability of inexpensive computational power, they do allow more resource-constrained model developers to perform rapid prototyping prior to investing in costlier techniques.

[3] See Bellotti and Crook (2013) for an application of this methodology in a retail credit context.

We present an innovation to the literature by introducing a structural Merton model style distance-to-default (DTD) measure into a stress testing application. Further, we construct discrete-time hazard rate models of PD, featuring a modeling data design that allows for a computationally efficient estimation technique, which is of particular value to modeling practitioners. Extending the literature along another dimension, the DTD risk factor, derived from equity prices and accounting measures of leverage, admits the design of challenger hybrid structural–reduced-form models, which are compared with the versions of these models containing all the other variables under consideration except for the DTD. It is demonstrated that the challenger models result in improved model performance (ie, measures of discriminatory power and predictive-level PD accuracy) and have a comparable quality of CECL scenario forecasts and that introduction of the structural DTD risk factor does not result in the other variables being rendered statistically insignificant.

The remainder of this study proceeds as follows. Section 2 constitutes a review of the relevant literature, such as studies on hazard rate modeling to predict binary outcomes, either the probability of occurrence or the time to the event, and we specify the discrete-time version within this class of models that we employ in the PD context. Section 3 is an outline of the methodology for our modeling exercise, where we discuss the general framework and its different subclasses, leading to the particular technique employed in this research. In Section 4 we present the empirical results of this study, including descriptive statistics of the modeling data set, estimation results, model performance metrics and the exercise in which we quantify model risk. Finally, in Section 5 we summarize our study and discuss future directions for this line of inquiry.

2 REVIEW OF THE LITERATURE

Rating or scorecard models have historically focused on estimating the PD, as opposed to the severity of losses, in the event of default or loss given default. Default is usually defined as a "failure", such as bankruptcy, liquidation, failure to pay, deemed unlikely to pay, etc. This construct does not consider downgrades or upgrades in credit ratings, as considered in mark-to-market (MTM) models of credit risk. These default mode (DM) credit risk models project credit losses only due to events of default, unlike MTM models that consider as credit events all credit quality changes. Among such DM models we can identify three broad categories: expert-based systems (eg, artificial neural networks), risk rating methodologies (eg, agency credit ratings from S&P Global Ratings or Moody's) and credit scoring models (eg, scorecards developed by banks or FICO scores).

The PD scoring model is most prevalent among the credit measurement methodologies used historically. One of the first models in this class was a multiple dis-

criminant analysis (MDA), as illustrated in the classic paper by Altman (1968). These types of models have the advantages of being cost-efficient to deploy and not subject to subjectivity or inconsistency, as observed in expert systems. Altman and Narayanan (1997) documented how these models became prevalent across the industry and academia and concluded that the similarities across applications are more pronounced than the differences. Another class of credit scoring models that is now widespread is the logistic regression model (LRM) (Hosmer *et al* 2013), a prime example of which is the RiskCalc model from the vendor Moody's Analytics (Dwyer *et al* 2004). This is used for commercial credit risk and considered an industry standard.

More advanced studies on the development, application and evaluation of predictive decision support models in the credit industry have been conducted beyond the aforementioned seminal academic and vendor approaches. Thomas (2010) highlights that corporate risk models employ data from balance sheets, financial ratios or macroeconomic indicators, whereas retail models use data from application forms, customer demographics and customer transaction history. He attributes these differences to specific modeling challenges that arise in consumer, as opposed to corporate, credit scoring and that lead many studies to focus on either the corporate or the retail business. García *et al* (2010) highlight how in PD scorecard development statistical hypothesis testing is often neglected or employed inappropriately, as the assumptions of parametric tests are violated in classifier comparisons, focusing on pairwise comparisons without p-value adjustments, which increase the actual probability of type I errors. Hofer (2015) addresses the research question of how to update PD scorecards in the face of new information, which is increasingly relevant to the industry as modeling data becomes more granular, especially in the retail sector but also for wholesale portfolios as there has been a movement from annual to quarterly observations. Hao *et al* (2011) propose a novel algorithm with several classifiers, paired algorithms and ensemble strategies in a factorial design, which focuses on preselected methods and omits a systematic comparison of several state-of-the-art classifiers, while Abellán and Mantas (2014) feature fewer classifiers and propose a novel algorithm that is compared with various reference methods. Finally, the benchmarking study by Lessmann *et al* (2015) compares several novel classification algorithms in state-of-the-art PD credit scoring models, in which they examine the extent to which the assessment of alternative scorecards differs across established and novel indicators of predictive accuracy.

In his seminal paper on the structural approach to credit risk, Merton (1974) models equity in a firm with leverage as a call option on assets where the strike price coincides with the face value of the debt. The PD is derived by solving for the option value numerically with the unobserved asset value and its volatility given the quantity of debt and a valuation horizon. The product of this process is a firm's DTD, which

represents the number of standard deviations separating asset and debt repayment values and is inversely related to the PD. The CreditEdge public firm model developed by Moody's Analytics is a well-established implementation of this framework that also uses historical default rates to empirically calibrate the output and produces the expected default frequency (EDF). Since the EDFs are ultimately based upon equity prices, there is consequently a heightened sensitivity to the changing financial state of the obligor, in contrast to agency credit ratings, which are more reliant on static data available at the time of underwriting or periodic reassessment of the borrower.

Another class of commonly used credit risk models arises from the previously discussed structural Merton approach and an alternative reduced-form framework originated by Jarrow and Turnbull (1995) and Duffie and Singleton (1999) that uses intensity-based models to estimate stochastic hazard rates. This school of thought differs in the methodology employed in estimating PDs. While the structural Merton approach considers an economic process that produces defaults, the reduced-form approach extracts a random intensity process that generates defaults from the prices of defaultable debt. A prominent example of a model in this class is the proprietary Kamakura Risk Manager, which incorporates an econometric methodology based on Chava and Jarrow (2004). This so-called Jarrow–Chava model (JCM) is sometimes called a hybrid approach, in that it combines the direct modeling of default, as in the LRM, with the use of either equity or debt market data, and in the case of traded debt instruments this construct has the potential to control for the distorting effect of illiquidity on the measurement of default risk. Note that a critique of the JCM is that the presence of anomalies, such as embedded options in the debt markets, can adversely impact the accuracy of these models. In this study, we circumvent this limitation by combining the use of fundamental factors, as in the LRM, with equity market information, as in the structural Merton model, in our version of a hybrid hazard rate model, as will be detailed in Section 3.

Beyond these seminal studies and vendor applications, there are several more recent and state-of-the-art studies that apply survival analysis in credit risk modeling. Baesens *et al* (2005) discuss and contrast statistical and neural network approaches for survival analysis. Several neural network survival analysis models are discussed and evaluated according to how they deal with censored observations, time-varying inputs and the monotonicity and scalability of the generated survival curves. Baesens *et al* compare the performance of a neural network survival analysis model with that of the proportional hazards model for predicting both loan default and early repayment, using data from a UK financial institution. Dirick *et al* (2015) study a special type of survival model called the mixture cure model, which facilitates the prediction of multiple events of interest, such as default and early repayment.

Stress testing based upon hypothetical scenarios, usually a blend of macroeconomic projections and the application of judgmental elements, has become a prevalent tool in the supervision of financial institutions (Financial Services Authority 2008). The qualitative aspect of the stress testing process is considered by some to be deficient, as illustrated by the critique of the exercise conducted by US supervisors during the 2007–9 global financial crisis, where the projection of unemployment in the 2009 adverse scenario fell short of the realization of this factor in under a year and was therefore deemed to be insufficiently severe (Board of Governors of the Federal Reserve System 2009). In his analysis of this incident, Baker (2009) claims that the supervisors may have underpredicted loan losses on the order of US$120 billion, placing him among those who conclude that this surprise on the part of the regulators is evidence that these tests are failures. In the retail credit risk context, Haldane (2009) attributes this weakness to either the omission of, or the underprediction of the impact of, risk factors in the dynamic macroeconomic model that was used (the so-called disaster myopia). He also points to the phenomenon termed "misaligned incentives", which means that institutions had no intention of designing realistic stress tests. Jacobs (2019) shows that, in addition to these downward biases, the prevalent econometric methodologies used by many supervisors and banks may be subject to heightened inaccuracies, which are attributed to a misspecified dependence structure between risk factors. He also highlights the challenges resulting from the paucity of default data in applying high-dimensional approaches to the stress testing of credit risk, as in the common approach investigated and improved upon, which will necessarily feature misspecified correlations that will give rise to inaccuracies.

There is a well-established literature on the applications of the survival approach to default prediction, as applied to stress testing applications, that follows the period of the pre-2010 downturn, an essentially profound paradigm shift to a stream of research presenting a variety of novel approaches, attributable to the increased importance of stress testing in the Basel III regulation and in daily practice by regulators. Bellotti and Crook (2013) estimate discrete-time survival models of borrower default for credit cards that include behavioral and macroeconomic factors across the life of the loan. They find that the dynamic models that include these variables provide statistically significant improvements in model fit and improved forecasts of default at both account and portfolio levels when applied to an out-of-sample data set. They also simulate extreme economic conditions and show how these models can be used to stress test credit card portfolios. Bellotti and Crook (2014) offer improved methodologies for scenario generation to predict default rates for retail portfolios implemented using discrete survival analysis, enabling macroeconomic conditions to be included as time-varying covariates. They depart from traditional models by employing Monte Carlo simulation to generate a distribution of estimated default rates from which value-at-risk (VaR) and expected shortfall are computed as a means of stress

testing. Breeden and Crook (2022) propose the use of survival models suitable for CECL stress testing that include lagged delinquency as a covariate and using a large sample of 30-year mortgages. They show that the proposed method is more accurate than any of the other methodologies they considered (roll rate, state transition and vintage models) for both short-term and long-term predictions of the default.

Finally, we review some of the foundational studies in the quantification of model risk according to the principle of relative entropy. Hansen and Sargent (2007) propose an alternative paradigm to the standard theory of decision-making under uncertainty based on a statistical model that informs an optimal distribution of outcomes. They adapt robust control techniques through developing a theory of model risk measurement that acknowledges misspecification in economic modeling and they apply this framework to a variety of problems in dynamic macroeconomics. Glasserman and Xu (2014) apply this framework to a financial risk measurement that relies on models of prices and other market variables that inevitably rely on imperfect assumptions that give rise to model risk. They develop a framework for quantifying the impact of model error through measuring and minimizing risk in a way that is robust to model error. Their robust approach starts from a baseline model and finds the worst-case error in risk measurement that would be incurred through a deviation from a baseline model given a precise constraint on the plausibility of the deviation. Using relative entropy to constrain model distance leads to an explicit characterization of worst-case model errors that lends itself to Monte Carlo simulation, allowing the straightforward calculation of bounds on the model error with very little computational effort beyond that required to evaluate performance under the baseline nominal model. They apply this technique to a variety of applications in finance, such as problems of portfolio risk measurement, credit risk, delta hedging and counterparty risk measured through credit valuation adjustment. Skoglund (2019) applies the principle of relative entropy to quantify the model risk inherent in loss-projection models used in macroeconomic stress testing and impairment estimation in an application to a retail portfolio and a delinquency transition model. He argues that this technique can complement traditional model risk quantification techniques, where a specific direction or range of model misspecification reasons, such as model sensitivity analysis, model parameter uncertainty analysis, competing models and conservative model assumptions, is usually considered.

3 MODEL METHODOLOGY AND CONCEPTUAL FRAMEWORK

In this section, we outline the development of the econometric framework that we employ. We proceed by first describing the most general model and defining a set of common terms. We then describe various special cases, starting with the hazard rate survival model in continuous time, followed by the discrete-time version of this

model. Finally, we describe the implementation of the landmark (or snapshot)[4] data sample design of the latter version of the hazard rate model.

Let us denote by t the calendar time, which we decompose as $t = a_i + \tau_i$. The time of origination is a_i (ie, the date on which the snapshot is measured) and the time of duration is τ_i (ie, the time from the measurement of the snapshot to the end of the forecast horizon). Units of observation or obligors are subscripted by i for $i = 1, \ldots, N$ obligors. We may record time at various granularities, in the case of C&I borrowers either quarterly or annually, and while spacing may be irregular due to the exact timing of when financial statements are spread, in general these will be in multiples of one quarter, so that in reality we are dealing with discrete sampling. Variables subscripted by i differ among borrowers but not temporally (like a segmentation characteristic of a borrower, such as a PD scorecard or an industry group). Variables subscripted by t differ through calendar time, but are common among obligors (as with a macroeconomic variable), and a term subscripted by it may be permitted to vary both temporally and cross-sectionally (as with a financial ratio). The corresponding risk factors or covariates are given by $\boldsymbol{w}_i$, $\boldsymbol{z}_t$ and $\boldsymbol{x}_{it}$, while the respective parameter vectors are given by $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$. The terms $\boldsymbol{\gamma}_{12}$, $\boldsymbol{\gamma}_{13}$ and $\boldsymbol{\gamma}_{23}$ denote matrixes of interaction term parameters between these three sets of parameters to be estimated.

The following describes a rather general and stylized econometric model of PD with respect to obligor i in a discrete-time period t. We will later impose restrictions upon this framework to arrive at a representation of the credit risk models used in practice. Let us denote by d_{it}^* a continuous latent variable representing the utility[5] gained by the default of borrower i in period t. We define the event of default as $d_{it} = 1$ if $d_{it}^* > 0$, and the event of non-default as $d_{it} = 0$ if $d_{it}^* \leqslant 0$. Suppose that this latent variable is a function linear in the risk factors and their interaction terms plus a residual term η_{it} and borrower-specific intercept term β_{0i}:

$$d_{it}^* = \beta_{0i} + \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + \boldsymbol{z}_t^{\mathrm{T}}\boldsymbol{\beta}_2 + \boldsymbol{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}_3 + \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\gamma}_{12}\boldsymbol{z}_t + \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\gamma}_{13}\boldsymbol{x}_{it} + \boldsymbol{z}_t^{\mathrm{T}}\boldsymbol{\gamma}_{23}\boldsymbol{x}_{it} + \eta_{it}, \tag{3.1}$$

[4] In the snapshot data sample design, for variables that have a different frequency such as quarterly macroeconomic variables but annual or semi-annual financial variables, for each instance of the former we create a time series that evolves while the latter is frozen (Houwelingen and Putter 2008).

[5] Since regression-based approaches to PD modeling are inherently empirical (in contrast to structural Merton model approaches), there is, strictly speaking, no theoretical model to describe. However, we can link the logistic regression approach to the economic theory of consumer or producer behavior, where the state of default can be viewed as a decision made by either a borrower or lender who is optimizing a utility function (Small and Rosen 1981).

and then define the conditional PD as

$$\mathrm{PD}_{it} = P(d_{it} = 1 \mid \boldsymbol{w}_i, \boldsymbol{z}_t, \boldsymbol{x}_{it}), \tag{3.2}$$

which implies that

$$\begin{aligned} &P(d_{it} = 1 \mid \boldsymbol{w}_i, \boldsymbol{z}_t, \boldsymbol{x}_{it}) \\ &\quad = P(\eta_{it} \leqslant \beta_{0i} + \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + \boldsymbol{z}_t^{\mathrm{T}}\boldsymbol{\beta}_2 + \boldsymbol{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}_3 + \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\gamma}_{12}\boldsymbol{z}_t + \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\gamma}_{13}\boldsymbol{x}_{it} + \boldsymbol{z}_t^{\mathrm{T}}\boldsymbol{\gamma}_{23}\boldsymbol{x}_{it}) \\ &\quad = F(\beta_{0i} + \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + \boldsymbol{z}_t^{\mathrm{T}}\boldsymbol{\beta}_2 + \boldsymbol{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}_3 + \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\gamma}_{12}\boldsymbol{z}_t + \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\gamma}_{13}\boldsymbol{x}_{it} + \boldsymbol{z}_t^{\mathrm{T}}\boldsymbol{\gamma}_{23}\boldsymbol{x}_{it}), \end{aligned} \tag{3.3}$$

where the distribution function of η_{it} is given by $F(\cdot)$. The variables subscripted by time may include lags of variable lengths.

Through the imposition of restrictions or assumptions governing various modeling aspects it may be demonstrated how this construct subsumes many PD models currently employed by practitioners. A canonical case is obtained by restricting all of the interaction terms to zero, which gives rise to the typical PIT PD model used in early warning or credit portfolio management:

$$P(d_{it} = 1 \mid \boldsymbol{w}_i, \boldsymbol{z}_t, \boldsymbol{x}_{it}) = F(\beta_{0i} + \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + \boldsymbol{z}_t^{\mathrm{T}}\boldsymbol{\beta}_2 + \boldsymbol{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}_3). \tag{3.4}$$

Usually, in this setting, t is a horizon spanning from one month to one year, and the function $F(\cdot)$ is typically the logistic link function of the LRM, $1/(1 + \exp(-x))$ (Hosmer *et al* 2013). It is common among practitioners to employ a linear transformation to the left-hand side term in (3.4), deriving a quantity interpreted as a "score", where an alternative means of deriving this is through scaling the logit estimate.

An important observation is that the model in (3.3) and (3.4) may not be suitable for an unbalanced panel data set,[6] which is exactly the format of data typical of the C&I asset classes. What this data design means is that there will be defaulted obligors over the prediction horizon, where subsequent performance is unobservable or otherwise recorded at alternative noncontiguous calendar times in the modeling data set. The consequence of this is potential bias in parameter estimates. Among the various means of accommodating this phenomenon is the survival analysis estimation technique (Kalbfleisch and Prentice 2002; Cox and Oakes 1984). This framework is prevalent in retail credit risk modeling, where the approximation of discrete time as

[6] Panel data can also be characterized as unbalanced panel data or balanced panel data. Balanced panel data sets have the same number of observations for all groups. Unbalanced panel data sets have missing values at some time observations for some of the groups. The main concern with unbalanced panel data is the question of why the data are unbalanced. If observations are missing at random, then this is not a problem. As an example, if the attrition of firms in your data over time is not random, ie, it is related to the idiosyncratic errors, then this sample selection may bias the estimates.

continuous is more likely to be realistic, as opposed to the wholesale setting, where this assumption is more likely to be tenuous.

We first review the case of the continuous-time survival model, where the target quantity is the instantaneous probability of transitioning from one state (eg, an obligor with performing loans) to another (eg, default). Denoting by T_i the duration time to obligor default implies that the conditional PD in the next instant, given that the obligor is currently not in default, may be represented by the hazard function having the duration time τ:

$$\lambda_i(\tau) = \lim_{\Delta\tau \to 0} \frac{P(T_i \in [\tau, \tau + \Delta\tau] \mid T_i \geqslant \tau)}{\Delta\tau}. \tag{3.5}$$

It follows that the survival probability (ie, the probability of not being in default over some time interval) may be expressed as a function that can be written as the following integral representation of the hazard function:

$$S(\tau) = P(T_i \geqslant \tau) = \exp\left(-\int_{q=0}^{\tau} \lambda_i(q)\,\mathrm{d}q\right). \tag{3.6}$$

A popular approach in both the academic literature and consumer credit practice is the use of the Cox proportional hazard model (CPHM) (Cox 1972) to estimate the hazard function as well as the related survival probabilities. The CPHM model further admits inclusion of dynamic covariates and can be expressed as

$$\begin{aligned}
&\lambda_i(a_i, \tau_i, \boldsymbol{w}_i, \boldsymbol{z}, \boldsymbol{x}, \boldsymbol{\beta}) \\
&\quad = \lambda_0(\tau) \exp[\beta_{0i} + \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + \boldsymbol{z}(a_i + \tau_i)^{\mathrm{T}}\boldsymbol{\beta}_2 + \boldsymbol{x}_i^{\mathrm{T}}(\tau)\boldsymbol{\beta}_3 \\
&\qquad\qquad + \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\gamma}_{12}\boldsymbol{z}(a_i + \tau_i) + \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\gamma}_{13}\boldsymbol{x}_i(\tau) + \boldsymbol{z}(a_i + \tau_i)^{\mathrm{T}}\boldsymbol{\gamma}_{23}\boldsymbol{x}_i(\tau)],
\end{aligned} \tag{3.7}$$

where the risk factors $\boldsymbol{x}_i(\tau)$ are obligor specific and dynamic across the time of duration (eg, financial variables), $\boldsymbol{z}(a_i + \tau_i)$ are risk factors varying over absolute time but constant over the cross section (eg, macroeconomic factors), and $\lambda_0(\tau)$ is a baseline hazard function of time which models the evolution of default risk independently of the other risk factors (ie, the idea being that this is a time-dependent residual of sorts). In this construct, forecasts of obligor financial or macroeconomic conditions after the beginning of a forecast period propagate through all subsequent time periods and influence the hazard function and survival probabilities over the entire forecast horizon.

There are various ways in which standard LRMs are inferior to survival models. First, survival models admit PD prediction over arbitrary forecast horizons apart from the window of the default flag used in developing the LRM, such as a one-year horizon of most PIT PD models, and thus are inherently suitable for applications such

as CECL or DFAST loss forecasting. In addition, unlike in LRMs the PD predictions in survival models are conditional on not previously having been in default, and the data would have to be modified (as discussed below) in order to accommodate conditional defaults. Finally, as survival probabilities are available across the entire projection horizon there is the potential for an application to profitability forecasting, or for use as a challenger to an economic capital construct.

In (3.3) we represent a discrete-time panel model of binary choice, which is an LRM built upon a panel data design and is equivalent to a discrete-time survival model. We previously alluded to the fact that it is most common for financial institutions to use panel data sample designs that could be used to model dynamic risk factors. Essentially, this involves estimating LRMs for a set of fixed horizons but with a different model for each forecast window (for example for quarters 1 through $3 \times 4 = 12$ for CECL/DFAST applications). In fact, this is the approach taken by the vendor Kamakura in its corporate PD model, with the obvious disadvantage that it is extremely resource intensive and requires a specialized software infrastructure. Another issue with this approach is the degradation of performance as the default horizon lengthens. This may be undesirable in an application such as CECL where there is a premium on accuracy, in the sense of predicting the level of default rates, as well as in distinguishing between defaults and non-defaults.

An alternative to estimating an LRM for each forecast horizon (as just described), which gives identical results in the academic literature and, in some cases, in retail credit risk modeling, and can be implemented in some standard software packages (eg, PYTHON `scikit-survival`), is as follows. As we have a discrete-time panel data sample design, then, given a modeling data set design matrix of appropriate form, a discrete survival model may be estimated with a hazard function of the following form:

$$\begin{aligned} h_i^{\mathrm{d}}(a_i, \tau_i, \boldsymbol{w}_i, \boldsymbol{z}, \boldsymbol{x}, \boldsymbol{\beta}) &= P(T_i \in [\tau_i - 1, \tau_i] \mid T_i \geqslant \tau_i - 1) \\ &= 1 - \frac{S(a_i, \tau_i, \boldsymbol{w}_i, \boldsymbol{z}, \boldsymbol{x}, \boldsymbol{\beta})}{S(a_i, \tau_i - 1, \boldsymbol{w}_i, \boldsymbol{z}, \boldsymbol{x}, \boldsymbol{\beta})}, \end{aligned} \tag{3.8}$$

where we denote by h_i^{d} the discrete hazard rate for the ith obligor and by $S(\cdot)$ the associated survival probability. Cox (1972) has proposed the following specification of this relationship:

$$\mathrm{logit}(h_i^{\mathrm{d}}(a_i, \tau_i, \boldsymbol{w}_i, \boldsymbol{z}, \boldsymbol{x}, \boldsymbol{\beta})) = \mathrm{logit}(h_i^{\mathrm{d}}(\tau_i)) + \boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + \boldsymbol{z}_t^{\mathrm{T}}\boldsymbol{\beta}_2 + \boldsymbol{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}_3, \tag{3.9}$$

where the logit function (or log-odds ratio function) is the inverse of the logistic link function $F^{-1}(x) = \log(x/(1-x))$ and the discrete baseline hazard function is given by $h_i^{\mathrm{d}}(\tau_i)$.

Jenkins (1995) proposes estimating the model (3.9) by specifying indicator variables corresponding to each time interval. Alternatively, Singer and Willett (1993) suggest using functions of the duration time index to capture this effect as a legitimate approach. In either case, we define a default indicator that is zero in all intervals where default is not observed and unity during the period where a default occurs. Subsequent to default the obligor does not appear in the data set so long as the default is not cured, in which case the obligor reappears in the data set as a performing entity and the indicator is reset to zero.

The specifications discussed in this section prior to the models in (3.8) and (3.9) have all assumed that time is continuous, but as we have pointed out this is not the case in reality, and, moreover, in the C&I asset class with quarterly observations this is likely to not be a realistic setup. Stepanova and Thomas (2002) find in spite of this argument that estimation results assuming continuous or discrete time show little difference. This may not be surprising, as Kalbfleisch and Prentice (2002) show that as the observation intervals tend to zero the discrete- and continuous-time models do indeed converge. Nevertheless, we believe the observation that the continuous-time model is an adequate approximation could be highly dependent on the particular data set, and in our case we do not find the differences to be immaterial in terms of either the coefficient estimates or the measures of model performance, which leads us to favor employing a discrete-time model.

Finally, we come to the approach used to estimate the hazard models in this research, which is based upon a paper by Houwelingen and Putter (2008) in the biostatistics literature. They model survival probabilities at a five-year horizon for acute lymphocytic leukemia patients after transplantation of bone marrow. This research proposes a landmark methodology and compares it with an established multistate modeling methodology in biostatistics. Houwelingen and Putter show that this technique greatly simplifies the modeling methodology, as it reduces to LRM estimation on the so-called snapshotted data set and leads to easy-to-interpret prediction rules.

At each landmark (snapshot in our terminology) point a simple Cox constant baseline hazard model is fitted on the interval, which is mathematically equivalent to estimating an LRM model on the restructured data set. This is in line with the methodologies used in the industry for PD scorecard development and does not assume continuous time, is less computationally intensive than the panel logistic approach of Kamakura and allows for simple implementation within standard software such as PYTHON. Our approach can be expressed mathematically as a modification of the LRM equation (3.4):

$$P(d_{it} = 1 \mid \boldsymbol{w}_i^S, \boldsymbol{z}_t^S, \boldsymbol{x}_{it}^S, t_i^s) = F(\beta_{0i} + f(t_i^s) + \boldsymbol{w}_i^{S\mathrm{T}}\boldsymbol{\beta}_1 + \boldsymbol{z}_t^{S\mathrm{T}}\boldsymbol{\beta}_2 + \boldsymbol{x}_{it}^{S\mathrm{T}}\boldsymbol{\beta}_3 + g(t_i^s \boldsymbol{x}_{it}^{S\mathrm{T}}\boldsymbol{\beta}_4)), \quad (3.10)$$

where the snapshotted data set is given by

$$\boldsymbol{X}_{it}^{S} = (f(t_i^s), \boldsymbol{w}_i^S, \boldsymbol{z}_t^S, \ldots, \boldsymbol{x}_{it}^S, g(t_i^s \boldsymbol{x}_{it}^S)),$$

t_i^s is the time since snapshot, $f(\cdot)$ is an appropriate transformation of the latter, $t_i^s \boldsymbol{x}_{it}^S$ is a set of time-interacting terms on the obligor-specific variables that vary over time (which can include the PD rating as well as financial ratios), and $g(\cdot)$ is some appropriate transformation of the latter. The time interactions capture the decay effects over a forecasting horizon with respect to obligor specific risk factors. The term $f(t_i^s)$ is analogous to the baseline hazard function in the Cox proportional hazard specification of (3.10).[7]

4 EMPIRICAL ANALYSIS

4.1 Description of modeling data

In this section we describe the data used in the empirical experiment. We collected data that are representative of a large corporate portfolio of borrowers, as would be held by a typical US bank. We used well-known sources that over the years would have been examined by multiple researchers and modelers, assuring good data quality at the start. Apart from this we performed the standard data cleaning procedures to further maximize the chances of having the best quality data available. We intentionally tried to maximize the historical time period and the range of variable types in order to have the most robust models as possible. The following data are used for the development of the models in this study.

Compustat. Standardized fundamental and market data for publicly traded companies, including financial statement line items and industry classifications (Global Industry Classification Standards (GICS) and North American Industry Classification System (NAICS)) over multiple economic cycles from 1979 onward. These

[7] However, a downside of this approach is that the model fit obtained that is available in standard software cannot be used to test the statistical significance of the parameter estimates. The correct standard errors can be obtained by taking into account the clustering of the data (ie, each snapshot is effectively a separate case, several such cases exist for each obligor and there is correlation over time in the former and cross-sectionally in the latter) using the so-called sandwich estimators of Lin and Wei (1989). This approach is incorporated in software packages such as SAS (the `GENMOD` or `SURVEYSELECT` procedures) or PYTHON (the generalized estimation equations in the logistic regression function of the `statsmodels` library). However, such approaches are very computationally intensive, and exhibit stability issues in the case where we have highly unbalanced panels, where the latter means nothing more than defaults being relatively rare and concentrated over time. However, in this research we obtain standard errors through bootstrapping, where we impose the proper stratification to preserve the correlation structure of the data, which is straightforward to implement in PYTHON due to vectorized operations.

data include default types such as bankruptcy, liquidation and rating agencies' default rating, all of which are part of the industry standard default definitions.

Moody's Default Risk Service (DRS) Rating History. An extensive database of rating migrations, default and recovery rates across geographies, regions, industries and sectors.

BankruptcyData New Generation Research, Inc. Provides information on corporate bankruptcies.

Center for Research in Security Prices (CRSP) US Stock Databases. This product is comprised of a database of historical daily and monthly market and corporate action data for over 32 000 active and inactive securities with primary listings on the New York Stock Exchange (NYSE), NYSE American, Nasdaq, NYSE Arca and Better Alternative Trading System (BATS) exchanges, and it includes CRSP broad market indexes.

A series of filters are applied to this Moody's population to construct a population that is closely aligned with the large corporate segment of US companies that are publicly rated and have publicly traded equity. In order to achieve this using Moody's data, the following combination of NAICS and GICS industry codes, regional codes and a historical yearly net sales threshold are used.

(1) Non-C&I obligors defined by the following NAICS codes (see Table 1) are not included in the population:

- financials;
- real estate investment trusts;
- government;
- dealer finance;
- not for profit.

(2) A similar filter is performed according to the GICS (see Table 2) classification:

- education;
- financials;
- real estate.

(3) Only obligors based in the United States or Canada are included.

(4) Only obligors with maximum historical yearly net sales of at least US$1 billion are included.

(5) There are exclusions for obligors with missing GICS codes, and for modeling purposes obligors are categorized into different industry segments on this basis.

(6) Records prior to 1991 Q1 are excluded, the rationale being that capital markets and accounting rules were different before the 1990s, and the macroeconomic data used in the model development are only available after 1990. As one-year change transformations are among those applied to the macroeconomic variables, this cutoff is advanced one year from 1990 to 1991.

(7) Records that are too close to a default event are not included in the development data set. This is an industry standard approach, the rationale being that the records of an obligor in this time window do not provide information about future defaults of the obligor, but rather they are more likely to reflect existing problems that the obligor is experiencing. This restriction corrects a range of timing issues between when statements are issued and when ratings are updated.

(8) In general, the defaulted obligors' financial statements after the default date are not included in the modeling data set. However, in some cases obligors may exit a default state or "be cured" (eg, emerge from bankruptcy), in which case the statements between the default date and the cure date are not included.

In our opinion, these data exclusions are reasonable, in line with industry standards, sufficiently documented and do not compromise the integrity of the modeling data set.

The model development time period considered for the Moody's data is 1991 Q1–2015 Q4. Table 1 shows a comparison of the modeling population by GICS industry sectors, where for each sector the defaulted obligors columns represent the defaulted obligors in the sector as a percentage of the entire population. The data are concentrated in consumer discretionary (20%), industrials (17%), tech hardware and communications (12%) and energy except Exploration & Production (E&P) (11%). A similar industry composition is shown in Table 2 according to the NAICS classification system.

The model development data set contains financial ratios and default information that are based upon the most recent data available from DRS, Compustat and BankruptcyData, so that the data are timely and, a priori, should be give the benefit of the doubt with respect to favorable quality. Further, the model development time period 1991 Q1–2015 Q4 spans two economic downturn periods and a complete business cycle, the length of which is another factor that supports a verdict of good quality. Related to this point, we plot the yearly and quarterly default rates in the model development data set in Figure 1.

TABLE 1 Moody's large corporate modeling data analysis and GICS industry segment composition for all Moody's obligors versus defaulted Moody's obligors (1991–2015).

GICS industry segment	All Moody's obligors	Defaulted Moody's obligors
Consumer discretionary	19.6	30.9
Consumer staples	8.4	6.4
Energy	7.6	5.9
Healthcare equipment & services	2.9	2.9
Industrials	31.6	15.1
Materials	10.5	11.3
Pharmaceuticals & biotechnology	2.7	0.2
Software & IT services	2.5	1.8
Technology hardware & communications	4.3	11.3
Utilities	7.6	5.6

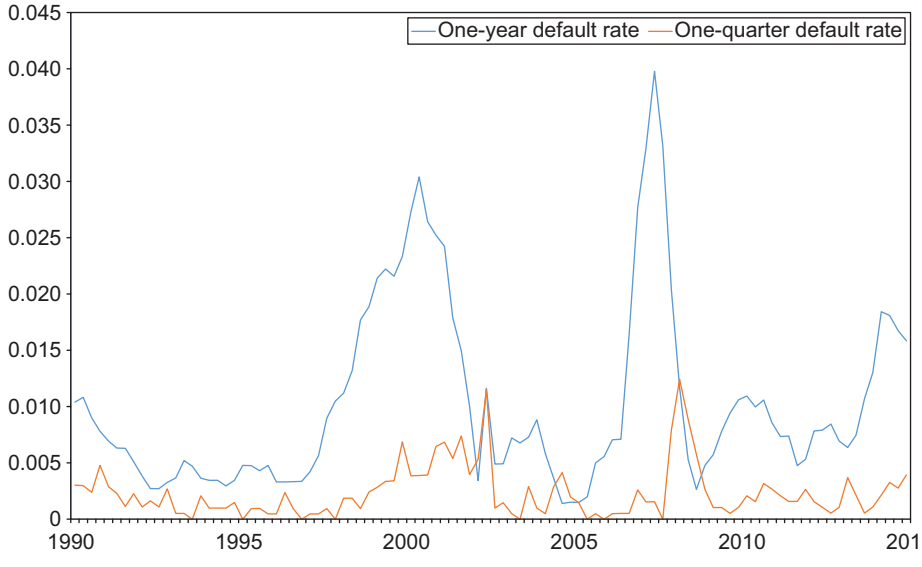
All values are given in percent.

TABLE 2 Moody's large corporate modeling data analysis and NAICS industry segment composition for all Moody's obligors versus defaulted Moody's obligors (1991–2015).

NAICS industry segment	All Moody's obligors	Defaulted Moody's obligors
Agriculture, forestry, hunting & fishing	0.2	0.4
Accommodation & food services	2.3	2.9
Waste management remediation services	2.4	2.1
Arts, entertainment & recreation	0.7	1.0
Construction	1.7	2.5
Educational services	0.1	0.2
Healthcare & social assistance	1.6	1.6
Information services	11.5	12.1
Management compensation enterprizes	0.1	0.1
Manufacturing	37.7	34.4
Mining, oil & gas	6.8	8.6
Other services (ex-public administration)	0.4	0.6
Professional, scientific & technological services	2.3	2.5
Real estate, rentals & leasing	0.9	1.6
Retail trade	9.6	12.4
Transportation & warehousing	5.4	7.0
Utilities	8.3	5.4
Wholesale trade	7.0	2.7

All values are given in percent.

FIGURE 1 PD model large corporate modeling data: one-year and one-quarter horizon default rates over time (1991–2015).



In Table 3 we present the summary statistics for the variables that appear in our final models. These final models were chosen based upon an exhaustive search algorithm in conjunction with five-fold cross-validation, and we have chosen the leading three models incorporating and omitting the DTD risk factor. The following are the categories and names of the explanatory variables appearing in the final candidate models:

- financial/liquidity current ratio (CR), net working capital to tangible assets ratio (NWCTAR), adjusted cash ratio (ACR);
- macroeconomic unemployment rate (UR), S&P 500 equity index (S&P), nonfarm employment (NFE), Baa corporate bond spread (SPR) and Dow Jones equity index (DOW);
- credit quality PD rating (PD);
- duration time since snapshot (TSS); and
- Merton structural distance-to-default (DTD).[8]

[8] All candidate explanatory variables are winsorized at either the 10th, 5th or 1st percentile levels at

TABLE 3 Summary statistics: Moody's large corporate obligor-level hazard rate models, explanatory variables and default indicators.

	Count	Mean	Standard deviation	Minimum	25th percentile	Median	75th percentile	Maximum
Unemployment rate[a] (%)	1.21E+06	−2.74	100.44	107.95	60.67	−42.99	25.48	394.92
S&P 500 equity index[b] (%)	1.21E+06	3.11	103.60	290.17	53.75	8.15	62.62	223.75
Nonfarm employment[c] (%)	1.21E+06	0.03	100.68	−424.35	−44.40	28.39	69.66	145.91
Dow Jones equity index[d] (%)	1.21E+06	2.70	103.60	−288.06	−53.64	12.75	63.83	237.01
Baa corporate bond spread (%)	1.21E+06	8.19	101.35	−129.40	−75.63	−1.97	65.19	419.18
Net working capital to tangible assets[e] (%)	1.21E+06	−15.34	88.03	−414.74	76.02	15.52	36.93	812.59
Current ratio[f] (%)	1.21E+06	−15.94	81.86	−157.96	−67.06	−28.83	12.96	706.65
Adjusted cash ratio[g] (%)	1.21E+06	−12.56	87.74	−173.44	−70.27	−23.43	25.46	600.58
PD[h] (%)	1.21E+06	87.74	41.32	0.00	69.15	109.86	109.86	109.86
Time since snapshot[i]	1.21E+06	0.8734	0.4132	0.0000	0.6931	1.0986	1.0986	1.0986
Merton DTD (%)	1.21E+06	18.74	41.53	−338.86	2.06	6.07	16.52	526.06

[a]Year-on-year change lag 2. [b]Year-on-year change lag 2. [c]Quarter-on-quarter change lag 0. [d]Year-on-year change lag 1. [e]Winsorized and standardized. [f]Winsorized and standardized. [g]Winsorized and standardized. [h]Logarithm. [i]Logarithm and winsorized.

4.2 Econometric specifications and model validation

In Tables 4 and 5 we present the estimation results and in-sample performance statistics for our final leading models; the remaining runner-up models are shown in Appendix B online as the results are qualitatively similar. In Table 4 we tabulate the results for the leading models with the DTD risk factors included as well as the other explanatory variables, whereas in Table 5 we show the best models that omit the DTD variable.

Across the models, the signs of the coefficient estimates are in line with economic intuition, and significance levels are indicative of very precisely estimated parameters. The macroeconomic variables associated with improving economic conditions (S&P, NFP and DOW) have negative signs, while those that indicate deteriorating conditions (UNP and SPR) have positive signs. The duration variable TSS has a positive sign, which is consistent with the intuition that on an unconditional basis default risk increases over time, given that the preponderance of the obligors in the sample are rated better than speculative grade. The sign on the PD rating is positive, which makes sense in that worse rated obligors have higher default risk. The financial ratios measuring borrower liquidity all have negative signs, as greater levels of such resources diminish the chances of a default, while the interaction terms with time are positive, the latter indicating sensibly that the efficacy of this factor decays over time. Finally, the negative signs on DTD indicate that firms further away from their default points have lower default risk, as expected.

Area under the curve (AUC) statistics indicate that the models have a strong ability to rank order default risk. The associated receiver operator characteristics (ROC) curves are shown in Figure 2 (Figure 4) for the leading model, which has the DTD risk factors included with (excluded from) the other explanatory variables. Regarding measures of predictive accuracy, in all cases the pseudo R-squared (PR2) indicates that all models exhibit good fit, which is confirmed by the plots of the predicted PD versus the default rates over time, as shown in Figure 3 (Figure 7) for the leading model that has the DTD risk factor included with (excluded from) the other explanatory variables. As expected, the Akaike information criterion (AIC) and PR2 predictive accuracy measures deteriorate when the DTD risk factors are omitted, but this rank ordering does not carry over to the AUC discriminatory power measure except in the case of the first leading model of each type.

In Figures 6–10 we show the 12-quarter baseline and adverse scenario macroeconomic forecasts for the models, with the average PDs. These scenarios are sourced from Moody's Analytics as of 2021 Q4. We observe that, while all the

either tail of the sample distribution, in order to mitigate the influence of outliers or contamination in data and according to a customized algorithm that analyzes the gaps between these percentiles and caps/floors where these are maximal.

TABLE 4 Hazard rate regression estimation results: Moody's large corporate financial, macroeconomic credit quality, duration and Merton DTD explanatory variables for one-quarter default model 1.

Variable	Coefficient estimate	Standard error	P-value
S&P 500 equity price index	−0.4425	0.0180	0.0000
Unemployment rate	0.1465	0.0165	0.0000
Logarithm of PD	1.0383	0.0335	0.0000
Logarithm of time	0.0375	0.4967	0.0000
(Logarithm of PD)*(logarithm of time)	−0.026	0.1540	0.0095
Net working capital to tangible assets	−0.4984	0.1791	0.0000
(Net working capital to tangible assets)*(time)	0.0198	0.1650	0.0061
DTD	−0.5786	0.2547	0.0082
Constant	−0.3050	0.1426	0.0047
Loglikelihood	−18 192.00		
AIC	36 400.63		
Pseudo-R-squared	0.161		
Area under the receiver operator curve	0.881		

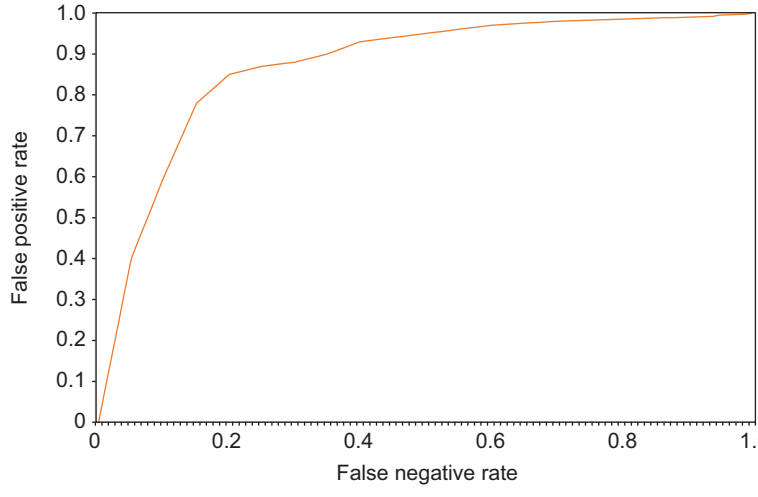
models show a reasonable pattern of stress in the adverse scenario relative to the baseline scenario, the patterns exhibit significant variation across models, so that the final model selection will be dependent upon which of these patterns is deemed preferable by business and risk management experts, for reasons other than statistical performance.

4.3 The quantification of model risk according to the principle of relative entropy

In the building of risk models we are subject to errors from model risk, one source of which is the violation of modeling assumptions. In this section we apply a methodology for the quantification of model risk that is a tool in building models robust to such errors. A key objective of model risk management is to assess the likelihood, exposure and severity of model error in that all models rely upon simplifying assumptions. It follows that a critical component of an effective model risk framework is the development of bounds upon a model error resulting from the violation of modeling assumptions. This measurement is based upon a reference nominal risk model and is capable of rank ordering the various model risks as well as indicating which perturbation of the model has maximal effect upon some risk measure.

In line with the objective of managing model risk in the context of obligor-level PD stress testing, we calculate confidence bounds around forecasted PDs spanning

FIGURE 2 Hazard rate regression ROC: Moody's large corporate financial, macroeconomic credit quality, duration and Merton DTD explanatory variables for one-quarter default model 1.



model errors in the vicinity of a nominal or reference model defined by a set of alternative models. These bounds can be likened to confidence intervals that quantify sampling error in parameter estimation. However, these bounds are a measure of model robustness that instead measure model error due to the violation of modeling assumptions. In contrast, a standard error estimate conventionally employed in managing credit portfolios does not achieve this objective, as this construct relies on an assumed joint distribution of the asset returns or default correlation.

We meet our previously stated objective in the context of stressed PD modeling through bounding a measure of loss, in this case the scenario PD forecasts, which can reflect, within reason, a level of model error. We have observed that, while among practitioners one alternative means of measuring model risk is to consider challenger models, an assessment of estimation error or sensitivity in perturbing parameters is in fact a more prevalent means of accomplishing this objective and one which captures only a very narrow dimension of model risk. In contrast, our methodology transcends the latter aspect to quantify potential model errors such as incorrect specification of the probability law governing the model (eg, the distribution of error terms or the specification of a link function in generalized linear regression, of which logistic regression is a subclass), variables belonging in the model (eg, omitted-variable bias

FIGURE 3 Hazard rate regression receiver accuracy plot: Moody's large corporate financial, macroeconomic credit quality, duration and Merton DTD explanatory variables for one-quarter default model 1.

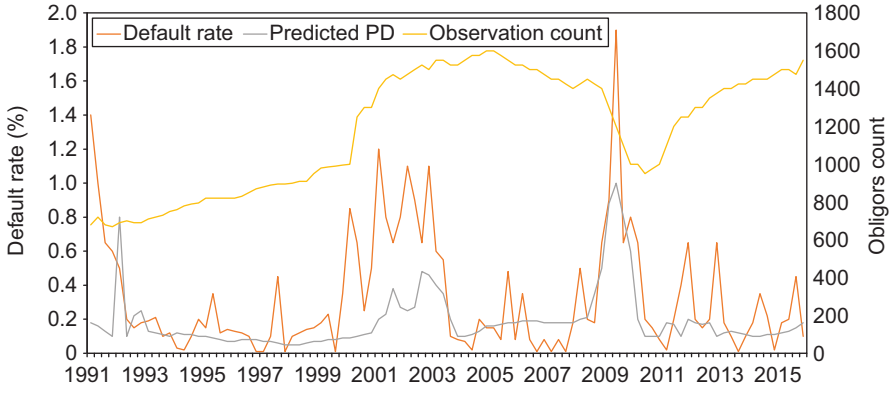


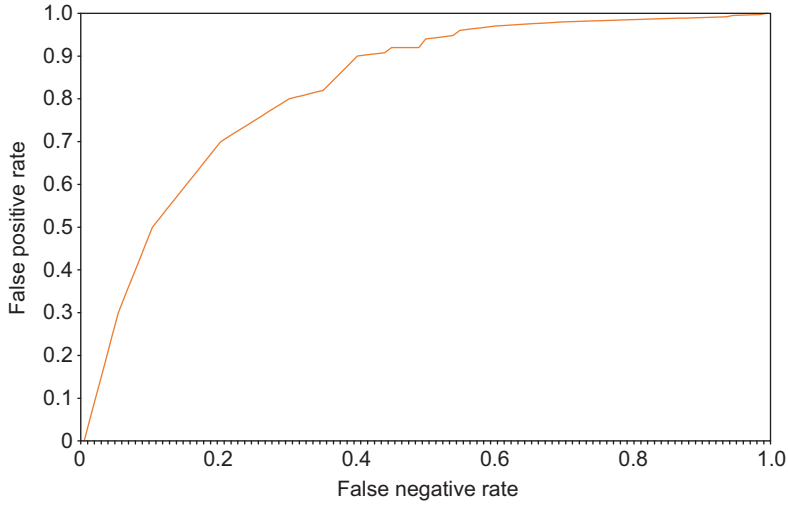
TABLE 5 Hazard rate regression estimation results: Moody's large corporate financial, macroeconomic credit quality and duration explanatory variables for one-quarter default model 1.

Variable	Coefficient estimate	Standard error	P-value
S&P 500 equity price index	−0.4608	0.0273	0.0000
Nonfarm employment	−0.0739	0.0225	0.0010
Logarithm of PD	1.2973	0.1687	0.0000
Logarithm of time	2.4152	0.4995	0.0000
(Logarithm of PD) × (logarithm of time)	−0.4115	0.1549	0.0079
Net working capital to tangible assets	−0.8437	0.1794	0.0000
(Net working capital to tangible assets) × (time)	0.4438	0.1654	0.0073
Constant	−5.0918	0.5431	0.0000
Loglikelihood	−19 857.85		
AIC	39 731.7		
Pseudo R-squared	0.136		
Area under the receiver operator curve	0.829		

with respect to the DTD) or the functional form of the model equations (eg, neglected transformations or interaction terms).

As these types of common model errors under consideration all relate to the likelihood of such an error, which in turn is connected to perturbation in the proba-

FIGURE 4 Hazard rate regression ROC: Moody's large corporate financial, macroeconomic credit quality and duration explanatory variables for one-quarter default model 1.



bility laws governing the entire modeling construct, we apply the principle of relative entropy (Hansen and Sargent 2007; Glasserman and Xu 2014). Relative entropy between a posterior distribution and a prior distribution is a measure of information gain when incorporating incremental data in Bayesian statistical inference. In the context of quantifying model error, relative entropy has the interpretation of a measure of the additional information requisite for a perturbed model to be considered superior to a champion or null model. Said differently, relative entropy may be interpreted as measuring the credibility of a challenger model. Another useful feature of this construct is that within a relative entropy constraint the so-called worst-case alternative (eg, in our case, the upper bounds on the scenario forecasts caused by ignoring some feature of the alternative model) can be expressed as an exponential change of measure.

Model risk with respect to a champion model $y = f(x)$ is quantified by the Kullback–Leibler relative entropy divergence measure to a challenger model $y = g(x)$ and is expressed as follows:

$$D(f,g) = \frac{g(x)}{f(x)} \log\left(\frac{g(x)}{f(x)}\right) f(x)\,\mathrm{d}x. \tag{4.1}$$

In this construct, the mapping $g(x)$ is an alternative PD model, and the mapping $f(x)$ is some kind of benchmark, the latter being the base PD models that we have

FIGURE 5 Hazard rate regression receiver accuracy plot: Moody's large corporate financial, macroeconomic credit quality and duration explanatory variables for one-quarter default model 1.

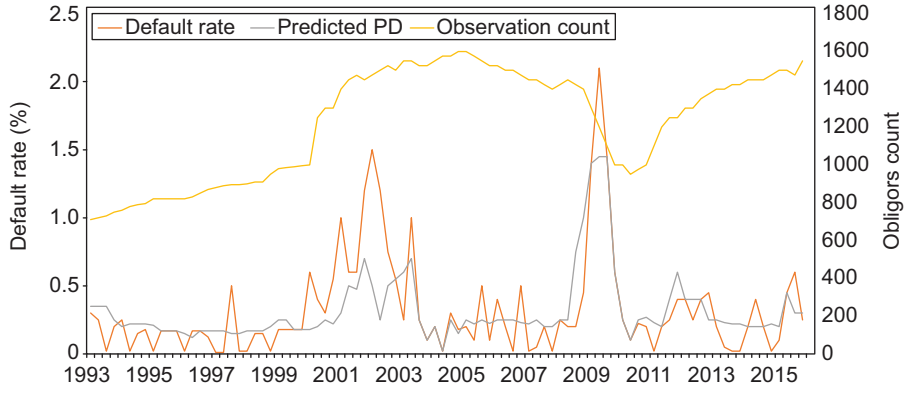


FIGURE 6 Baseline and stressed macroeconomic scenario forecasts: Moody's large corporate financial, macroeconomic credit quality, duration and Merton DTD explanatory variables for one-quarter default model 1.

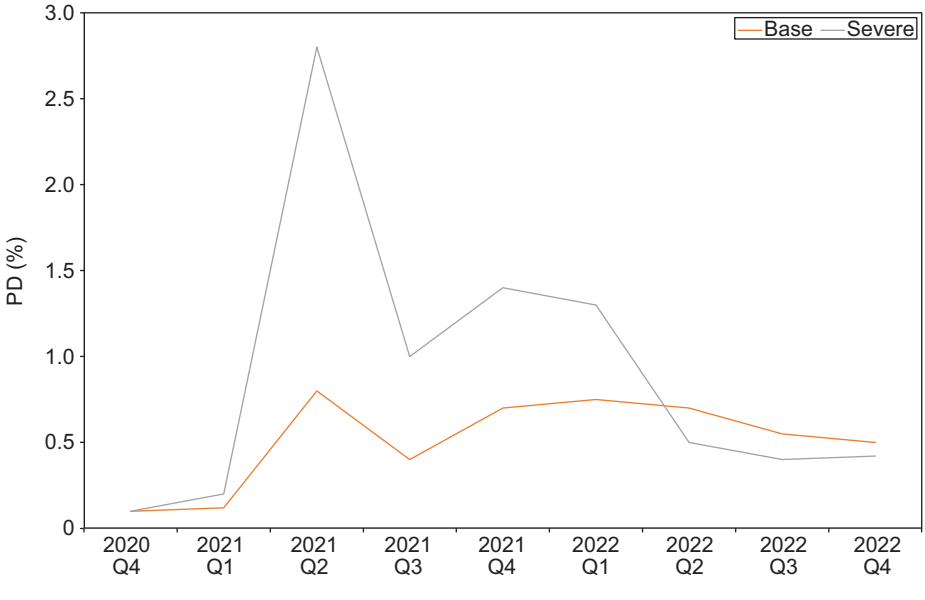
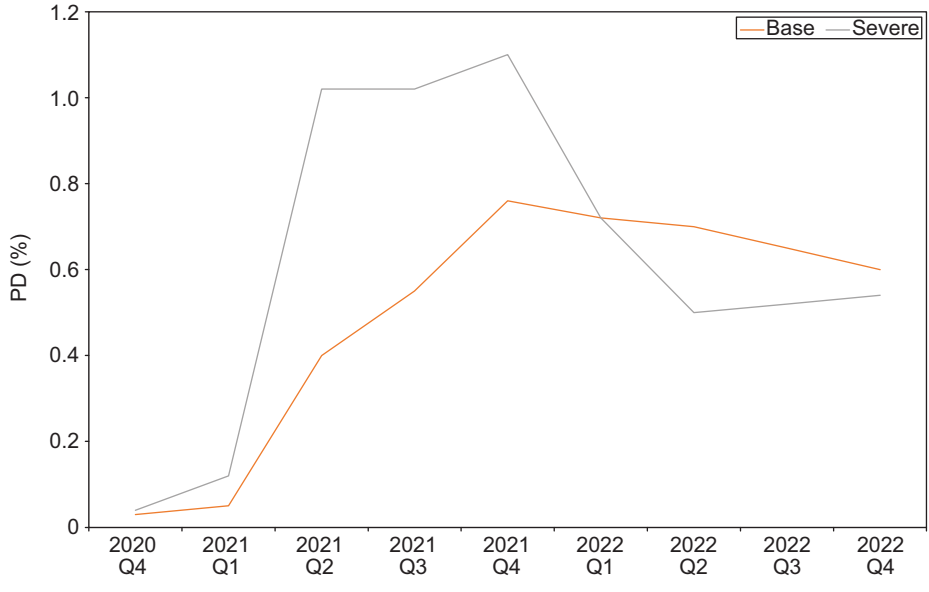


FIGURE 7 Baseline and stressed macroeconomic scenario forecasts: Moody's large corporate financial, macroeconomic credit quality and duration explanatory variables for one-quarter default model 1.



estimated in this paper that may be violating some model assumption. In a model validation context this is a critical construct as the implication of these relations is a robustness to model misspecification with respect to the alternative model (ie, we do not have to assume that either the reference model or alternative model is correct and we need only quantify the distance of the alternative from the reference model to assess the impact of the modeling assumption at play). Define the likelihood ratio $m(f, g)$, which characterizes our modeling choice and is expressed as follows:

$$m(f, g) = \frac{g(x)}{f(x)}. \tag{4.2}$$

As is standard in the literature, (4.2) may be expressed as an equivalent expectation of a relative deviation in likelihood:

$$\mathbb{E}_f[m \log(m)] = D(f, g) < \delta, \tag{4.3}$$

where δ is an upper bound to deviations in model risk (which should be relatively small), which may be determined by the model risk tolerance of an institution for a certain model type and interpretable as a threshold for model performance. A

property of relative entropy dictates that $D(f,g) \geqslant 0$ and $D(f,g) = 0$ only if $f(x) = g(x)$. Given a relative distance measure $D(f,g) < \delta$ and a set of alternative models $g(x)$, model error can be quantified by the following change of numeraire:

$$m_\theta(f,g) = \frac{\exp(\theta f(x))}{\mathbb{E}_f[\exp(\theta f(x))]}, \tag{4.4}$$

where the solution (or inner supremum) to (4.4) is formulated in the following optimization:

$$m_\theta(f,g) = \inf_{\theta>0} \sup_{m(x)} \mathbb{E}_f\left[m(x)f(x) - \frac{1}{\theta}(m(x)\log(m(x)) - \delta)\right]. \tag{4.5}$$

Equation (4.5) features the parameterization of model risk by $\theta \in [0,1]$, where $\theta = 0$ is the best case of no model risk and $\theta = 1$ is the worst case of model risk in extremis. The change in measure of (4.4) has the important property of being model-free, or not dependent upon the specification of the challenger model $g(x)$. As mentioned previously, this reflects the robustness to misspecification of the alternative model that is a key feature of this construct, and is, from a model validation perspective, a desirable property. In other words, we do not have to assume that either the champion model or the alternative model is correct and only have to quantify the distance of the alternative from the base model to assess the impact of violating the modeling assumptions.

We study the quantification of model risk with respect to the following modeling assumptions:

- omitted-variable bias;
- misspecification according to neglected interaction effects; and
- misspecification according to an incorrect link function.

Omitted-variable bias is analyzed by consideration of the DTD risk factor as discussed in the main estimation results in this paper, where we saw that including this variable in the model specification did not result in other financial or macroeconomic variables falling out of the model, and it improved model performance. The second assumption is based upon the estimation of alternative specifications that include interaction effects among the explanatory variables. Finally, we analyze the third assumption through estimation of these specifications with the complimentary log–log (CLL) as opposed to the logit link function.[9]

[9] The Logit link function used commonly in logistic regression is symmetric whereas CLL is asymmetric. When the probability of the binary or binomial response approaches 0 at a different rate

We implement this procedure in a bootstrap simulation exercise, where we develop a distribution of the baseline and adverse macroeconomic forecasts at each horizon, and study the high 95th and the low 5th percentiles of these distributions as upper and lower bounds on model risk, respectively. In each iteration, we resample the data with replacement (stratified in order that the history of each obligor is preserved), and reestimate the models considered in the main body of the paper, as well as three variants that include either DTD, interaction effects or a CLL link function. In the case of the DTD risk factor, we will be comparing the variants, ie, those considered in the main results, which have already been estimated, except that in each run the results will be perturbed according to the different bootstraps of the data set, and in the other two cases there will be alternative estimations.[10]

The results of the model risk quantification exercise are shown for the leading model with the DTD risk factor for omitted-variable bias in Figure 8, for neglected interaction effects in Figure 9 and for a misspecified link function in Figure 10. We also tabulate the summary statistics of these results for all three models in Table 6.

It is observed in the width of the model risk bounds that omitted-variable bias with respect to DTD results in the highest model risk, incorrectly specified link function has the lowest measured risk and neglected interaction effects are intermediate in the quantity of model risk. Two more notable characteristics of these results are the asymmetry in the model risk bounds, which are skewed toward greater projected PD estimates, and also that the bounds are not monotonic – these aspects are not featured in the parametric confidence bounds, which measure pure parameter uncertainty.

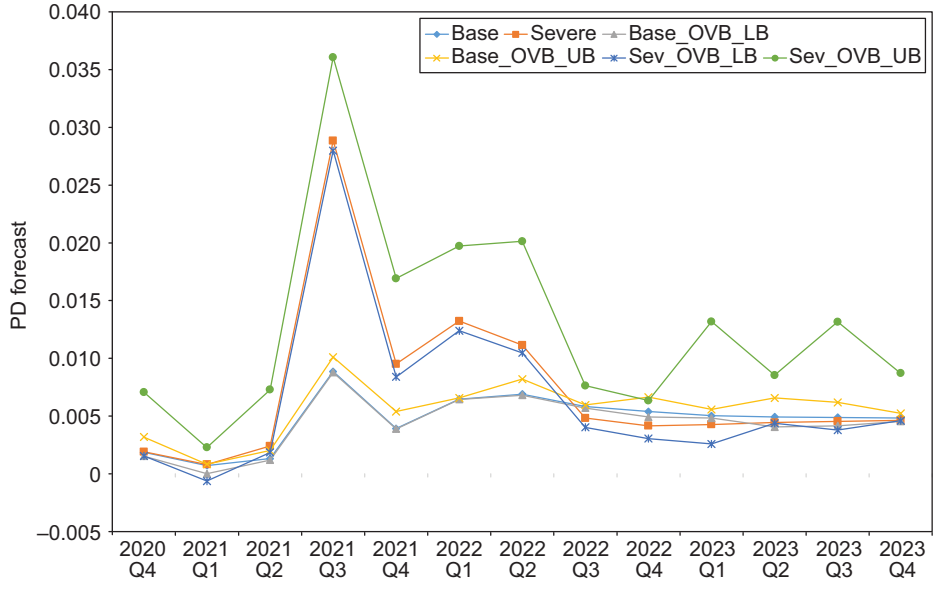
5 CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

In this study we measured the model risk attributable to various model assumptions in dynamic econometric models of large corporate default according to the principle of relative entropy. This methodology studies the distance of an alternative model to a reference model according to some suitable loss metric (see Hansen and Sargent 2007; Glasserman and Xu 2014) and can capture dimensions of model uncertainty error beyond parameter estimation error. This framework for measuring model risk was applied to a CECL stress testing exercise of PD for a corporate portfolio. It was observed that omitted-variable bias (with respect to the Merton DTD risk factor) has

than it approaches 1 (as a function of a covariate), symmetric link functions cannot be appropriate and do not always provide the best fit for the given data set in binomial regression. This is the case for the unbalanced data that we have. As defaults are very rare events, asymmetric link functions such as the CLL are sometimes good alternatives.

[10] For the sake of brevity, we do not include these results, but they are available from the author upon request. Across 100 000 iterations the results are stable and robust across the base as well as for the alternative specifications.

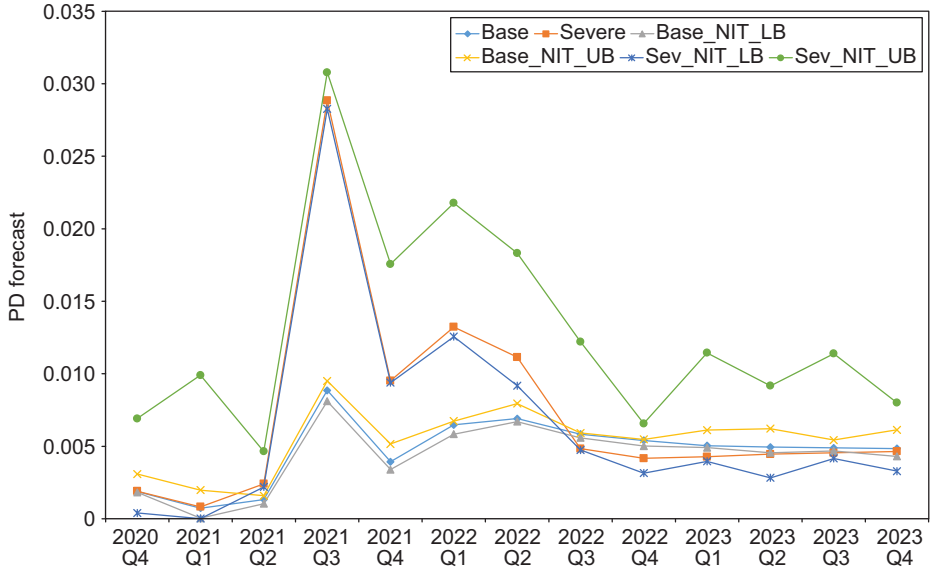
FIGURE 8 Quantification of model risk according the principle of relative entropy forecast upper and lower bounds for omitted-variable bias: Moody's large corporate financial, macroeconomic credit quality, duration and Merton DTD explanatory variables for one-quarter default model 1.



the greatest impact upon measured model risk in terms of bounds on PD forecasts, the incorrect specification of the link function has the least impact and the neglect of interaction effects between risk factors has an intermediate impact. The importance of this application is rooted in the sensitivity of the CECL or DFAST results from the perspective of prudential supervision as well as accounting policy, as model validators or regulators often question the impact of faulty model assumptions on capital and reserve projections. We addressed this issue through this research.

This quantification of model risk was accomplished through the consideration of an obligor-level hazard rate methodology for corporate PD modeling that features macroeconomic, financial, equity market, duration and credit rating variables. This methodology was applied to stress testing for CECL, where we have departed from the common practice for wholesale portfolios of adapting rating transition models where the ratings are stressed for this purpose, and to our knowledge this is one of the first studies in the literature to have done so. We have further innovated by developing explicitly discrete-time hazard models with a specialized data sample design (the landmark methodology) that is particularly tractable in computation. This

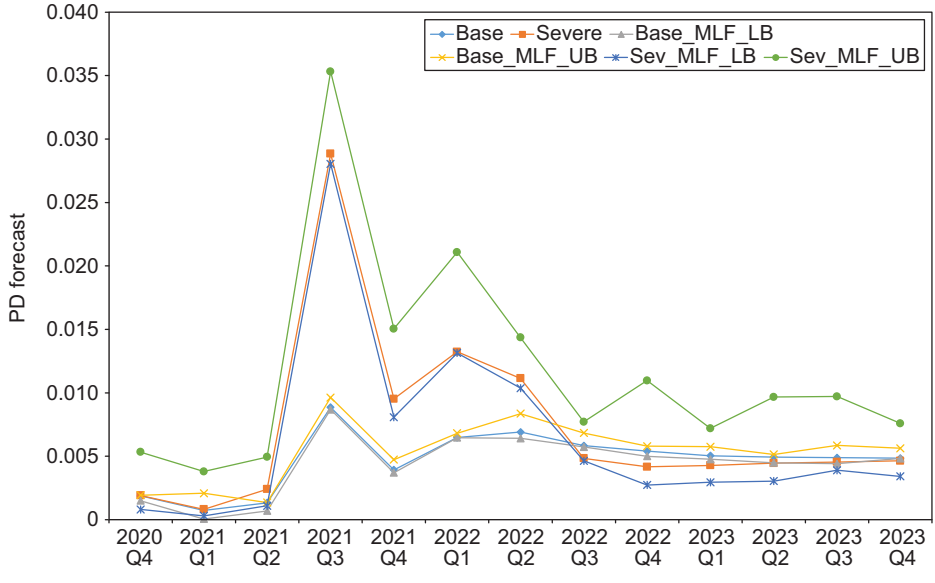
FIGURE 9 Quantification of model risk according the principle of relative entropy forecast upper and lower bounds for neglected interaction terms: Moody's large corporate financial, macroeconomic credit quality, duration and Merton DTD explanatory variables for one-quarter default model 1.



allows for the inclusion of a large number of variables and the utilization of a long historical data set.

The models that we developed are particularly suitable to this experiment in measuring model risk, as they feature rich risk modeling data and variable structures that allow for the investigation of varied modeling assumptions. Our base data were a lengthy borrower-level history of corporate ratings and defaults sourced from Moody's in the period 1990–2015. The data were enhanced by attaching an extensive set of financial, macroeconomic and equity market variables to form the basis of candidate explanatory variables. The obligor-level hazard rate models developed had a one-quarter default horizon and, further, featured time decay and duration effects. Based upon the relevant literature, we also considered an alternative structural risk factor, which is: the structural modeling DTD measure constructed from the market value of equity and accounting leverage measures. We then compared these hybrid structural–reduced-form models with the financial ratio and macroeconomic variable only models. It has been shown that adding the DTD measures to our leading mod-

FIGURE 10 Quantification of model risk according the principle of relative entropy forecast upper and lower bounds for misspecified link function: Moody's large corporate financial, macroeconomic credit quality, duration and Merton DTD explanatory variables for one-quarter default model 1.



els did not invalidate the other variables chosen but significantly augmented model performance and resulted in comparable scenario forecasts.

We have addressed an overarching conceptual validation question regarding stress testing in demonstrating the model's fitness-for-purpose for use as inputs to downstream models. This is a particular concern in the case of wholesale portfolios, since in this setting it is the more predominant practice to have this disconnect, which leads to challenges in demonstrating the conceptual soundness of both credit risk and stress testing models to model validators and supervisors (Global Credit Data 2019). The type of PD models we investigated in this paper addressed this issue as they are directly applicable to stress testing. There is no known academic or practitioner literature addressing this issue that is particular to the wholesale credit asset class, which is where our research presents its main contribution.

Our conclusion is that validation methods chosen in the stress testing context should be capable of testing model assumptions, given the sensitive regulatory uses of these models and concerns raised in the industry about the effect of model misspecification on capital and reserves. Our research adds to the literature by offer-

TABLE 6 Quantification of model risk according to the principle of relative entropy forecast upper and lower bounds summary statistics: Moody's large corporate financial, macroeconomic credit quality, duration and Merton DTD explanatory variables for one-quarter default model 1. [Table continues on next two pages.]

(a) Omitted-variable bias

	Model 1						Model 2						Model 3					
	Base	LB	UB	Sev.	LB	UB	Base	LB	UB	Sev.	LB	UB	Base	LB	UB	Sev.	LB	UB
Mean	0.47	0.44	0.56	0.73	0.65	1.29	0.47	0.44	0.55	0.55	0.44	1.10	0.28	0.24	0.38	0.34	0.18	0.76
SD	0.23	0.24	0.24	0.74	0.74	0.88	0.26	0.27	0.25	0.36	0.32	0.45	0.08	0.10	0.10	0.14	0.16	0.39
Minimum	0.07	0.00	0.08	0.08	−0.06	0.23	0.02	0.00	0.09	0.02	0.04	0.04	0.10	0.03	0.25	0.20	0.00	0.28
25th prc.	0.39	0.39	0.52	0.42	0.26	0.73	0.38	0.37	0.53	0.49	0.34	0.92	0.21	0.18	0.29	0.26	0.06	0.46
Median	0.49	0.45	0.60	0.45	0.40	0.87	0.56	0.52	0.67	0.52	0.43	1.07	0.29	0.28	0.38	0.28	0.13	0.59
75th prc.	0.58	0.57	0.66	0.95	0.84	1.69	0.63	0.59	0.72	0.78	0.56	1.37	0.33	0.33	0.45	0.42	0.30	0.97
Maximum	0.89	0.88	1.01	2.88	2.80	3.61	0.76	0.75	0.77	1.10	0.98	1.81	0.41	0.37	0.60	0.60	0.52	1.41

TABLE 6 Continued.

(b) Neglected interaction terms

	Model 1						Model 2						Model 3					
	Base	**LB**	**UB**	**Sev.**	**LB**	**UB**	**Base**	**LB**	**UB**	**Sev.**	**LB**	**UB**	**Base**	**LB**	**UB**	**Sev.**	**LB**	**UB**
Mean	0.47	0.43	0.55	0.73	0.65	1.30	0.47	0.43	0.55	0.55	0.44	0.98	0.28	0.24	0.37	0.34	0.22	0.76
SD	0.23	0.23	0.22	0.74	0.75	0.73	0.26	0.25	0.26	0.36	0.34	0.41	0.08	0.08	0.08	0.14	0.13	0.27
Minimum	0.07	0.00	0.16	0.08	0.00	0.46	0.02	0.00	0.08	0.02	0.00	0.25	0.10	0.07	0.20	0.20	0.08	0.30
25th prc.	0.39	0.34	0.52	0.42	0.40	0.80	0.38	0.36	0.45	0.49	0.38	0.76	0.21	0.20	0.34	0.26	0.20	0.52
Median	0.49	0.47	0.59	0.45	0.86	1.14	0.56	0.52	0.67	0.52	0.74	0.99	0.29	0.27	0.40	0.28	0.44	0.71
75th prc.	0.58	0.56	0.62	0.95	1.20	1.76	0.63	0.57	0.71	0.78	0.99	1.27	0.33	0.29	0.43	0.42	0.71	0.95
Maximum	0.89	0.81	0.95	2.88	2.83	3.08	0.76	0.71	0.84	1.10	0.97	1.64	0.41	0.39	0.47	0.60	0.44	1.17

TABLE 6 Continued.

(c) Misspecified link function

	Model 1						Model 2						Model 3					
	Base	**LB**	**UB**	**Sev.**	**LB**	**UB**	**Base**	**LB**	**UB**	**Sev.**	**LB**	**UB**	**Base**	**LB**	**UB**	**Sev.**	**LB**	**UB**
Mean	0.47	0.43	0.44	0.73	0.63	1.17	0.47	0.43	0.44	0.55	0.46	0.96	0.28	0.24	0.24	0.34	0.26	0.68
SD	0.23	0.23	0.24	0.74	0.75	0.85	0.26	0.25	0.26	0.36	0.34	0.46	0.08	0.08	0.07	0.14	0.15	0.33
Minimum	0.07	0.00	0.01	0.08	0.03	0.38	0.02	0.00	0.01	0.02	0.00	0.13	0.10	0.07	0.08	0.20	0.08	0.27
25th prc.	0.39	0.34	0.37	0.42	0.27	0.72	0.38	0.36	0.33	0.49	0.30	0.57	0.21	0.20	0.20	0.26	0.13	0.43
Median	0.49	0.47	0.48	0.45	0.34	0.97	0.56	0.52	0.54	0.52	0.44	1.09	0.29	0.27	0.27	0.28	0.24	0.62
75th prc.	0.58	0.56	0.57	0.95	0.81	1.44	0.63	0.57	0.57	0.78	0.75	1.28	0.33	0.29	0.28	0.42	0.37	0.96
Maximum	0.89	0.81	0.87	2.88	2.80	3.53	0.76	0.71	0.73	1.10	1.06	1.71	0.41	0.39	0.34	0.60	0.52	0.89

Units: percentages. SD denotes standard deviation.

ing state-of-the-art techniques as viable options in the arsenal of model validators, developers and supervisors seeking to manage model risk.

Given the wide relevance and scope of the topics addressed in this study, there is no shortage of fruitful avenues along which we could extend this research. Some proposals include, but are not limited to:

- alternative econometric techniques, such as various classes of machine learning models, including nonparametric alternatives;
- asset classes beyond the large corporate segments, such as small business, real estate or even retail;
- the consideration of industry specificity in model specification; and
- data sets in jurisdictions apart from the United States, else pooled data encompassing different countries with a consideration of geographical effects.

DECLARATION OF INTEREST

The author reports no conflicts of interest. The author alone is responsible for the content and writing of the paper.

REFERENCES

Abellán, J., and Mantas, C. J. (2014). Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications* **41**(8), 3825–3830 (https://doi.org/10.1016/j.eswaa.2013.12.003).

Aguais, S. D., Forest, L. R., Jr., King, M., Lennon, M. C, and Lordkipanidze, B. (2008). Designing and implementing a Basel II compliant PIT–TTC ratings framework. Paper 6902, MPRA. URL: https://mpra.ub.uni-muenchen.de/6902/1/MPRA_paper_6902.pdf.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance* **23**(4), 589–609 (https://doi.org/10.2307/2978933).

Altman, E. I., and Narayanan, P. (1997). An international survey of business failure classification models. *Financial Markets, Institutions and Instruments* **6**(2), 1–57 (https://doi.org/10.1111/1468-0416.00010).

Baesens, B., Van Gestel, T., Stepanova, M., Van den Poel, D., and Vanthieinen, J. (2005). Neural network survival analysis for personal loan data. *Journal of the Operational Research Society* **56**(9), 1089–1098 (https://doi.org/10.1057/palgrave.jors.2601990).

Baker, D. (2009). Background on the stress tests: anyone got an extra $120 billion? *American Prospect*, May 8. URL: https://prospect.org/economy/background-stress-tests-anyone-got-extra-120-billion/.

Basel Committee on Banking Supervision (2011). Basel III: a global regulatory framework for more resilient banks and banking systems. Standards Document, June, Bank for International Settlements. URL: https://www.bis.org/publ/bcbs189.htm.

Bellotti, T., and Crook, J. (2013). Forecasting and stress testing credit card default using dynamic models. *International Journal of Forecasting* **29**(4), 563–574 (https://doi.org/10.1016/j.ijforecast.2013.04.003.).

Bellotti, T., and Crook, J. (2014). Retail credit stress testing using a discrete hazard model with macroeconomic factors. *Journal of the Operational Research Society* **65**(3), 340–350 (https://doi.org/10.1057/jors.2013.91).

Board of Governors of the Federal Reserve System (2009). The supervisory capital assessment program: overview of results. White Paper, May, Federal Reserve, Washington, DC. URL: www.federalreserve.gov/bankinforeg/bcreg20090507a1.pdf.

Board of Governors of the Federal Reserve System (2016). Dodd–Frank Act stress test 2016: supervisory stress test methodology and results. Report, June, Federal Reserve, Washington, DC. URL: www.federalreserve.gov/newsevents/pressreleases/files/bcreg20160623a1.pdf.

Breeden, J., and Crook, J. (2022). Multihorizon discrete time survival models. *Journal of the Operational Research Society* **73**(1), 56–69 (https://doi.org/10.1080/01605682.2020.1777907).

Chava, S., and Jarrow, R. (2004). Bankruptcy prediction with industry effects. *Review of Finance* **8**(4), 537–569 (https://doi.org/10.1093/rof/8.4.537).

Cheng, K. F., Chu, C. K., and Hwang, R.-C. (2010). Predicting bankruptcy using the discrete-time semiparametric hazard model. *Quantitative Finance* **10**(9), 1055–1066 (https://doi.org/10.1080/14697680902814274).

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society* B **34**(2), 187–220 (https://doi.org/10.1111/j.2517-6161.1972.tb00899.x).

Cox, D. R., and Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall/CRC, Boca Raton, FL (https://doi.org/10.1201/9781315137438).

Dirick, L., Claeskens, G., and Baesens, B. (2015). An Akaike information criterion for multiple event mixture cure models. *European Journal of Operational Research* **241**(2), 449–457 (https://doi.org/10.1016/j.ejor.2014.08.038).

Duffie, D., and Singleton, K. J. (1999). Modeling term structures of defaultable bonds. *Review of Financial Studies* **12**(4), 687–720 (https://doi.org/10.1093/rfs/12.4.687).

Dwyer, D. W., Kogacil, A. E., and Stein, R. M. (2004). Moody's KMV RiskCalcTM v2.1 model. White Paper, April, Moody's Analytics. URL: https://bit.ly/3L6C5uR.

Financial Accounting Standards Board (2016). Financial instruments – credit losses (topic 326): measurement of credit losses on financial instruments. Accounting Standards Update 2016-13, June, FASB, Norwalk, CT. URL: https://bit.ly/3U1gpEA.

Financial Services Authority (2008). Stress and scenario testing. Consultation Paper 8/24, December, FSA, London. URL: www.fasb.org/jsp/FASB/Document_C/DocumentPage?cid=1176168232528.

García, S., Fernández, A., Luengo, J., and Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power. *Information Sciences* **180**(10), 2044–2064 (https://doi.org/10.1016/j.ins.2009.12.010).

Glasserman, P., and Xu, X. (2014). Robust risk measurement and model risk. *Quantitative Finance* **14**(1), 29–58 (https://doi.org/10.1080/14697688.2013.822989).

Global Credit Data (2019). Current expected credit losses (CECL) benchmarking survey. Survey, January, Global Credit Data. URL: https://globalcreditdata.org/gcd_library/cecl-survey-2018-public-version/?seq_no=3.

Gross, M. W., Leika, M., and Lukyantsau, P. (2020). Expected credit loss modeling from a top-down stress testing perspective. Working Paper WP/20/111, July, International Monetary Fund (https://doi.org/10.5089/9781513549088.001).

Haldane, A. G. (2009). Why banks failed the stress test. Speech presented at Marcus-Evans Conference on Stress-Testing, London, UK, February 9. URL: www.bis.org/review/r090219d.pdf.

Hansen, L. P., and Sargent, T. (2007). *Robustness*. Princeton University Press (https://doi.org/10.1515/9781400829385).

Hao, J., Jiang, H., Ma, J., and Wang, G. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications* **38**(1), 223–230 (https://doi.org/10.1016/j.eswa.2010.06.048).

Hofer, V. (2015). Adapting a classification rule to local and global shift when only unlabelled data are available. *European Journal of Operational Research* **243**(1), 177–189 (https://doi.org/10.1016/j.ejor.2014.11.022).

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. 2013. *Applied Logistic Regression*, 3rd edn. Wiley (https://doi.org/10.1002/9781118548387).

Houwelingen, H. C., and Putter, H. (2008). Dynamic predicting by landmarking as an alternative for multi-state modeling: an application to acute lymphoid leukemia data. *Lifetime Data Analysis* **14**(1), 447–463 (https://doi.org/10.1007/s10985-008-9099-8).

Jacobs, M., Jr. (2015). The quantification and aggregation of model risk: perspectives on potential approaches. *International Journal of Financial Engineering and Risk Management* **2**(2), 124–154 (https://doi.org/10.1504/IJFERM.2015.074045).

Jacobs, M., Jr. (2019). The accuracy of alternative supervisory methodologies for the stress testing of credit risk. *International Journal of Financial Engineering and Risk Management* **3**(3), 254–296 (https://doi.org/10.1504/IJFERM.2019.10029778).

Jacobs, M., Jr. (2020). A holistic model validation framework for current expected credit loss (CECL) model development and implementation. *International Journal of Financial Studies* **8**(27), 1–36 (https://doi.org/10.3390/ijfs8020027).

Jarrow, R. A., and Turnbull, S. M. (1995). Pricing derivatives on financial securities subject to credit risk. *Journal of Finance* **50**(1), 53–85 (https://doi.org/10.1111/j.1540-6261.1995.tb05167.x).

Jenkins, S. P. (1995). Easy estimation methods for discrete-time duration models. *Oxford Bulletin of Economics and Statistics* **57**(1), 129–138 (https://doi.org/10.1111/j.1468-0084.1995.tb00031.x).

Kalbfleisch, J. D., and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edn. Wiley (https://doi.org/10.1002/9781118032985).

Lessmann, S., Baesens, B., Seow, H.-V., and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research. *European Journal of Operational Research* **247**(1), 124–136 (https://doi.org/10.1016/j.ejor.2015.05.030).

Lin, D. Y., and Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association* **84**(408), 1074–1078 (https://doi.org/10.2307/2290085).

Merton, R. C. (1974). On the pricing of corporate debt: the risk structure of interest rates. *Journal of Finance* **29**(2), 449–470 (https://doi.org/10.1111/j.1540-6261.1974.tb03058.x).

Shumway, T. (2001). Forecasting bankruptcy more accurately: a simple hazard model. *Journal of Business* **74**(1), 101–124 (https://doi.org/10.1086/209665).

Singer, J. D., and Willett, J. M. (1993). It's about time: using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational and Behavioral Statistics* **18**(2), 155–195 (https://doi.org/10.3102/10769986018002155).

Skoglund, J. (2019). Quantification of model risk in stress testing and scenario analysis. *The Journal of Risk Model Validation* **13**(1), 1–23 (https://doi.org/10.21314/JRMV.2019.201).

Small, K. A., and Rosen, H. S. (1981). Applied welfare economics with discrete choice models. *Econometrica* **49**(1), 105–130 (https://doi.org/10.2307/1911129).

Stepanova, M., and Thomas, L. (2002). Survival analysis methods for personal loan data. *Operations Research* **50**(2), 277–289 (https://doi.org/10.1287/opre.50.2.277.426).

Thomas, L. C. (2010). Consumer finance: challenges for operational research. *Journal of the Operational Research Society* **61**(1), 41–52 (https://doi.org/10.1057/jors.2009.104).