Journal of Futures Markets

WILEY

RESEARCH ARTICLE OPEN ACCESS

Appraising Model Complexity in Option Pricing

Mark Cummins[1] | Francesco Esposito[2]

[1]University of Strathclyde, Glasgow, Scotland | [2]Dublin City University, Dublin, Ireland

Correspondence: Mark Cummins (mark.cummins@strath.ac.uk)

Received: 22 April 2024 | **Revised:** 4 December 2024 | **Accepted:** 6 February 2025

Keywords: affine and nonaffine jump-diffusion model specifications | model complexity | model confidence set | option pricing models | stochastic volatility

ABSTRACT

The research question we consider is whether incremental complexity in option pricing models is justified by incremental model performance. We apply the model confidence set as a formal model comparison approach in appraising stochastic volatility jump-diffusion option pricing models, spanning affine and nonaffine specifications. Jumps in price with stochastic (constant) arrival intensity produce superior (inferior) outcomes. A parsimonious negative exponential price jump distribution outperforms the popular normal distribution. Jumps in volatility (synchronized or not) worsen model performance. A parsimonious nonlinear hyperbolic drift extension of the Heston model performs particularly well. Nonlinear CEV models generally do not produce appreciable model performance.

JEL Classification: C12, C52, C58, G13

1 | Introduction

[a] perfectly specified option pricing model is bound to be too complex for applications.

Bakshi et al. (1997, p. 2004)

Development work on option pricing models has seen, over recent decades, ever more complex model specifications being proposed to, purportedly, better capture observed market dynamics. We reflect on this pursuit of model complexity. The contribution of our study is to take stock of the extant academic literature on option pricing models and assess the observed trend of ever-increasing model complexity, across the affine and nonaffine classes of models. Our research question is set as follows: Is incremental model complexity justified by incremental model performance? In answering this question, we provide insights into what forms of model complexity provide superior performance. To proceed, we (i) provide a workable definition of model complexity, (ii) consider a large suite of option pricing models of various classes proposed in the literature, (iii) identify a range of measures of model performance that span practical requirements pertaining to pricing, hedging and volatility modeling, and (iv), in an important departure from the existing literature, propose a formal statistical device that allows us to identify the set of superior models from a suite of competing option pricing models and to rank these superior models therein. An important distinction from current work is the integrated approach to model selection and model ranking. The approach also explicitly tackles the issue of multiple comparisons bias (Romano et al. 2010) that results from testing multiple competing models simultaneously. We uniquely adopt the model confidence set (MCS) approach of Hansen et al. (2011) that allows for a statistically robust means to select equivalently performing superior option pricing models.

Our study is motivated by the work of Ignatieva et al. (2015), who conduct a comprehensive model specification analysis on time-series equity index data. Ignatieva et al. (2015) extend the analysis of Eraker et al. (2003) and, in contrast to previous studies, contribute through the analysis of a large model set, comprising a range of affine and nonaffine model specifications that incorporate stochastic volatility and jump components. The model comparison exercise that the authors conduct is limited to the use of the deviance information criteria (DIC) to obtain a model ranking, while Bayes factors are employed for pairwise

comparisons of several model specifications. Similar samples and techniques, leading to similar conclusions, are employed in Kaeck et al. (2017), whereby models are ranked by DIC. Ignatieva et al. (2015) provide some important insights, concluding that pure stochastic volatility models are inferior in performance to their jump-augmented counterparts, while the authors also provide evidence that nonaffine specifications outperform affine models, even after including jumps.

We differ in this study through an analysis of a large suite of affine and nonaffine stochastic volatility jump-diffusion models on a cross-section of option pricing data. In theory, consistency between the physical and risk-neutral distributions is only achieved in the case of perfect model specification, such that differences in cross-section option implied and time-series estimated parameters are an indication of model misspecification (Bakshi et al. 1997). While Broadie et al. (2007) make the case that a good option pricing model must be able to fit both cross-sectional and time-series properties, Bates (2003) points out that such consistency is necessary for estimating risk premia (not a focus of our paper) and is valid under a hypothesis of correct model specification plus measurement error. Our study is premised on an assumption of model misspecification. Bates (2003) further makes a couple of arguments against mixing data sources. First, the approach mixes statistical approaches and, in effect, imposes a two-stage procedure of implied calibration followed by historical estimation of the respective parameters. Second, attempting to capture implied and historical market dynamics jointly in a single (misspecified) model makes the process of model selection and explaining model rejection more difficult. Indeed, Bates (2003) advocates structuring tests in the form of cross-sectional derivatives price patterns versus time-series properties. In this sense, our study complements the time-series equity index work of Ignatieva et al. (2015) with our cross-sectional equity options analysis.

We review the option pricing literature, focusing on both affine and nonaffine model specifications, and on the methods used for estimation, testing, and model comparison. While some prominent reviews and commentaries on option pricing models exist (Bates 1996b, 2003; Broadie and Detemple 2004), none perform an extensive model comparison exercise as we do here. Seminal writings in the study of option pricing model performance are the works of Bates (1996a) and Bakshi et al. (1997), which both set standards in model specification analysis by examining (i) measures of in-sample and out-of-sample mispricing, (ii) the distributional characteristics of the option implied stochastic process and their consistency with the underlying returns time-series distribution, and (iii) single-instrument and delta-neutral hedging analysis. Following these influential studies on the behavior of stochastic volatility option pricing models, the literature has particularly focused on the affine jump-diffusion model framework, as a consequence of the stylized approach originated in Heston (1993), Duffie and Kan (1996), and Duffie et al. (2000), and the dimensional flexibility and computational efficiency that integral transform methods offer in terms of option pricing and hedge measure calculation. Nonetheless, there have been a number of studies dealing with pricing models employing alternative volatility specifications. For instance, Benzoni (2002) provides an empirical analysis of affine and log-normal volatility models, while Christoffersen et al. (2010) consider alternative volatility characterizations to the square root variance process, including fixed constant elasticity of variance (CEV) parameters, such as the "three-halves" and the continuous-time autoregressive conditional heteroskedasticity (ARCH) model. Other nonlinear models originate from the interest rate literature, for example, Chan et al. (1992), whereas Chernov et al. (2003) and Christoffersen et al. (2009) deal with multifactor extensions of the affine, CEV, and log-normal models. CEV-based equity price models appear in Beckers (1980) and Macbeth and Merville (1980), with alternative CEV stochastic volatility models proposed by Jones (2003) and Aït-Sahalia and Kimmel (2007). Further alternatives are represented by the infinite-activity time-change Lévy model class of Carr and Wu (2004).

Our work adds to the discourse on option pricing model specification by means of drawing on the above-referenced and related literature and comparing the largest set of option pricing models considered to date in a single study. This contrasts with the majority of studies in this field to date that provide limited benchmarking of model performance when such studies propose new models. The model set we examine is obtained as augmented and extended versions of the baseline affine stochastic volatility diffusion model of Heston (1993), on which we incrementally and selectively combine model components to increase model complexity. This gives us our working definition of model complexity, which we define as model constructs derived from introducing jumps in the returns and/or volatility processes, imposing state dependency on jump intensity, imposing distributional forms on the jump processes, increasing the dimension of the parameter vector,[1] and increasing the dimension of the system state. Further elements of model complexity are assumed to derive from introducing nonlinearity in the drift and/or diffusion components.

The option pricing model literature generally lacks a formal statistical testing approach. We, in contrast, exploit the rigor of the MCS (Hansen et al. 2011) in our model selection exercise. The MCS is a general model comparison test that, starting with an initial model set, permits the automatic selection of the subset of best-performing models based on statistical equivalence and generates a model ranking across the model set by means of a p value function. Building on the Reality Check (RC) of White (2000) and the Superior Predictive Ability (SPA) of Hansen (2005), the MCS explicitly accounts for the multiple comparisons bias typical of multiple hypothesis test settings (White 2000; Romano and Wolf 2005b), such as the model selection exercise we pursue. Multiple comparison bias implies that, when comparing a family of competing models simultaneously, the superiority of one model over another model may, in fact, be a random artifact. This source of bias is an issue that is overlooked in the option pricing model literature but is particularly important in our setting, given the large suite of option pricing models we consider.

We design several model comparison tests targeting different performance measures and several meaningful sample clusters, an approach that provides a multifaceted and comprehensive view of model behavior. We investigate measures of model performance concerning (i) option model mispricing, (ii) robust delta-hedging, and (iii) implied volatility (IV) consistency. Our

analysis delivers some important messages. We conclude that stochastic intensity jumps in price, preferably with a negative exponential distribution, are a necessary extension to the affine diffusion model of Heston (1993), whereas we generally find no evidence to justify the flexible CEV model class. Constant intensity jumps are found to be detrimental to model performance, while jumps in volatility, whether synchronized or not with jumps in price, are redundant. Normal or infinite variance jumps, while showing equivalent performance for some performance measures, are perhaps unnecessary complications. Two-factor volatility models sometimes produce superior performance, but not consistently across all performance measures. While we do find the simple stochastic volatility model of Heston (1993) performs well from a hedging perspective, its nonlinear "quasi-affine" augmentation with hyperbolic drift provides an excellent alternative to the affine paradigm.

The remainder of the paper is organized as follows. In Section 2 we introduce the technical specifications of the option pricing model set we examine. In Section 3 we introduce our testing framework, where we define our implementation of the MCS methodology of Hansen et al. (2011). Section 4 sets out the experimental design, including the alternative model performance measures we use. Section 5 provides a detailed discussion of the empirical results. Section 6 sets out concluding remarks.

2 | The Option Pricing Model Set

2.1 | Option Pricing Model Literature

The cornerstone of the initial model set we consider is the seminal stochastic volatility diffusion model of Heston (1993) (hereafter, the Heston model), which first formally embedded a stochastic volatility specification in an option pricing setting. The success of this model and its widespread use in academia and industry is mainly due to the quasianalytic pricing formula that it offers and the efficiency of the computational methods required for implementation. The power of the Heston model lies in its tractable representation of a mean-reverting volatility process and its ability to capture the well-established leverage effect in equity markets through the correlation between asset returns and stochastic volatility. Nonetheless, several studies have highlighted the inadequacy of this model to explain several features of market prices, see, for example, Bakshi et al. (1997), Eraker (2004), and Broadie et al. (2007). While the Heston model allows for a skewed leptokurtic distribution of asset returns, this higher-order moment information is fully loaded onto just two parameters: the correlation parameter and the volatility of variance parameter. This leads to restricted levels of skewness and kurtosis compared with implied market levels, see, for instance, Das and Sundaram (1999).

Many extensions to the Heston model have been proposed, the most popular of which fall into the general affine jump-diffusion framework of Duffie et al. (2000). Such extensions include several forms of jump augmentation. Bates (1996a) and Bakshi et al. (1997) allow the price process to jump with deterministic intensity. Duffie et al. (2000) propose a model specification that allows for jumps in volatility, along with joint and correlated jumps in returns and volatility. Eraker et al. (2003) similarly allows for both coupled and decoupled jumps in return and volatility. Bates (2000) considers price jumps with stochastic intensity, which allow for clustered discontinuities that correspond to periods of high volatility, while Eraker (2004) considers several similar jump combinations. In our study, we consider these model specifications, which are all based on a single-factor volatility process.

Affine multifactor volatility extensions have also been proposed in the option pricing literature. Duffie et al. (2000) and Bates (2000) both consider a stochastic mean volatility process, allowing for a time-varying long-run factor to which the volatility process reverts. Other multifactor extensions include stochastic interest rates, such as Scott (1997) and Bakshi et al. (1997), which have also been combined with stochastic dividends, such as in Jones (2003). These specifications are motivated by the remarks of Christoffersen et al. (2009) that the level and the slope of the volatility smile fluctuate independently, requiring the use of two factors to adequately capture this dynamic. We include stochastic mean volatility model specifications in the initial model set, but we exclude other multifactor extensions (such as stochastic interest rates and stochastic dividends) on the basis that such model specifications are more appealing from a financial economics perspective.

With respect to jump size characteristics, the jump in volatility has usually been modeled as a positive exponential, whereas the jump in returns has taken the form of a normal distribution. We include both in our model specifications. In respect of the price jump size, we also consider several alternative specifications. In particular, we build on the jump specification of Kou (2002) and allow for negative exponential jumps, as this specification provides wider skewness and kurtosis. Moreover, we expand the initial model set with price jump specifications characterized by a negative tail with extreme value distribution (EVD), commonly referred to as the Lomax distribution,[2] which has similar behavior to the finite moment log stable class, see Carr and Wu (2003). This jump specification allows us to explore model behavior without leaving the affine framework and trespassing into the time-change infinitely active Lévy model class of Carr and Wu (2004).[3] The fat-tail distribution is constrained to produce infinite variance jumps. The interest in this jump setup lies in the argument that long maturity IV should exhibit flat behavior, as a consequence of the central limit theorem (CLT) acting upon the conditional distribution of returns (Carr and Wu 2003), whereas a fat-tailed random variable contradicts the CLT, allowing for persistent asymmetric long-term volatility curves.

With this rich affine jump-diffusion model set, we provide evidence that contributes to the debate about what features and combinations of features are associated with a significant increase in performance over the Heston model. Despite the flexibility of these jump-augmentations, such models still exhibit signs of misspecification, see, for example, Bates (2000). Towards addressing this misspecification, more elaborate model constructs have been proposed in the literature. Such extensions have, for example, introduced nonlinearity into the drift and/or the diffusion component. The most popular of the nonlinear

diffusion models is the CEV model, employed in econometric studies, such as Jones (2003), Chernov et al. (2003), and Aït-Sahalia and Kimmel (2007). The CEV model relaxes the square root diffusion assumption, allowing for a general exponent power function for the variance process. Special cases of the CEV model are the exponential ARCH diffusion (Nelson 1990) and the so-called "three-halves" model (Lewis 2000), which have been employed in empirical studies of option pricing, such as Christoffersen et al. (2010). We also consider some price jump extensions of the CEV model class to investigate whether nonlinearity in the diffusion component combined with jumps is able to produce significant improvements over affine model performance. We limit the jump specifications to stochastic intensity price jumps because, as we will show later, other jump components do not appear to contribute appreciably.

An interesting extension of the CEV model that involves the drift specification has been suggested in short-term interest rate studies, such as Chan et al. (1992), Conley et al. (1997), and Aït-Sahalia (1999), with analysis of equity returns and IV found in Bakshi et al. (2006), Chourdakis and Dotsis (2011), Ignatieva et al. (2015), and Kaeck et al. (2017). The model includes a Laurent polynomial in the drift whose exponents span from −1 to +2, allowing for a richer dynamic to the mean-reversion component. We include this model in our initial model set. We also consider a two-factor CEV model, for comparability purposes with the two-factor extension of the Heston model. Finally, the initial model set is completed with a further nonlinear model, represented by the mean-reverting log-normal volatility model introduced by Scott (1987) and, concurrently, by Hull and White (1987) and Wiggins (1987). This model is appealing because, as noted in Christoffersen et al. (2010), the logarithm of the realized variance tends to produce absolute changes that are uncorrelated from the log-volatility level, whereas the plain data exhibit at least linear dependency between changes and levels. As shown in the cited article, the quantile–quantile plot of the logarithmic realized variance appears to flatten. Thus, this model is expected to provide an improved representation of the market volatility. Nonetheless, the performance of this particular model in the context of option pricing has not been thoroughly explored, to date, although applications of this model can be found in Benzoni (2002) and Chernov et al. (2003). We, therefore, provide incremental insights in this respect.

2.2 | Model Specifications

We converge on a set of 28 alternative model specifications in total. We formalize the model specifications through the following overarching model:

$$ds = s[\tilde{r}dt + \sqrt{v}\,dW_0 + (e^{z_0} - 1)dN_0], \tag{1a}$$

$$dv = \mu(v)dt + \sigma(v)dW_1 + \zeta(z_1)dN_1. \tag{1b}$$

The process s represents the underlying log-price, which is a geometric motion with risk-neutral drift $\tilde{r}$.[4] The Brownian drivers are in general correlated, that is, $d\left[W_0, W_1\right] = \rho dt$, with $\rho \leq 0$ constant. The jump in returns has a random size determined by the variable z_0, which can have a normal, a negative exponential[5] or a negative Lomax EVD distribution. The parameter η_0 is introduced to indicate the expected size for the normal and the negative exponential as well as the location parameter in the EVD distribution, whereas the parameter ν_0 indicates the standard deviation and the scale parameter, respectively, in the case of the normal and the Lomax EVD distribution.

We define $\zeta(z_1)$ to be the volatility jump function corresponding to the volatility jump size variable z_1. Whenever the volatility jump is present, its size distribution is exponential, with parameter η_1, representing the expected jump displacement. The Poisson jumps N_0 and N_1 are driven by the jump intensity processes

$$d\Lambda_0 = \lambda_0 \tilde{v} dt, \tag{2a}$$

$$d\Lambda_1 = \lambda_1 dt \tag{2b}$$

with either $\tilde{v} = 1$ or v, depending on whether the model has constant or stochastic intensity associated with the jump in the returns process. In the latter instance, the jump intensity of the returns process is proportional to the volatility factor v. In respect of the jump intensity embedded in the stochastic volatility process, we assume this to be constant, with the exception of synchronized jumps ($N_0 = N_1$), whereby (2b) is discarded.[6]

Finally, to complete the initial model set, we define the auxiliary long-run stochastic mean factor, a, that defines the multifactor volatility models with appropriate adjustment of the volatility process drift:

$$da = (\alpha - \beta a)dt + \delta\sqrt{a}\,dW_2, \tag{3}$$

where W_2 is assumed to be independent from the other stochastic drivers W_0 and W_1. We do not consider multifactor log-normal volatility models.

Table 1 summarizes the main models that we use, mapping these to the relevant literature. In Section 2.3, we introduce a labeling scheme to identify the various model specifications that will ease the reader's navigation of the empirical testing discussion.

2.3 | Naming Convention

To assist the readability of the experimental results, we introduce a labeling convention to identify the suite of models presented in Section 2.2. This labeling convention departs from the typical notation used in the affine option pricing model literature but is introduced as this is the first study to consider such a range of affine and nonaffine model specifications in a unified study. Each model is identified by a symbolic alphanumeric code of three elements. The model codification is designed to provide an easy mnemonic to identify each model. The first element is either the letter A, C, or L, respectively, indicating an affine, a CEV or a log-normal diffusion type model. A superscript to this letter indicates auxiliary features of the model.

TABLE 1 | Model specification literature mapping.

Model	$\mu(\cdot)$	$\sigma(\cdot)$	$\zeta(\cdot)$	Jumps characteristics
Heston (1993)			0	(a)
Bates (1996a)		$\sigma\sqrt{v}$		(a), (b)
Affine model (Duffie et al. 2000; Eraker et al. 2003)			z_1	(a), (b), (c), (d)
Eraker (2004)	$a - bv$			(a), (b), (d), (f)
Extended-Kou (Kou 2002) (jump structure only)		σ		(g)
GARCH-SV (Christoffersen et al. 2010)		σv	0	(a)
3/2 model (Lewis 2000; Christoffersen et al. 2010)		$\sigma v^{3/2}$		(a)
CEV model (Jones 2003; Aït-Sahalia and Kimmel 2007)		σv^{γ}		(a)
Polynomial drift and CEV (Bakshi et al. 2006)	$a_{-1}v^{-1} + a_0 + a_1 v + a_2 v^2$			(a)
Log-normal model (Scott 1987)	$v(a - b\log v)$	σv		(a)

Note: This table summarizes the characteristics of a selection of models nested within the overarching model specification in Equations (1a, 1b) with a mapping to the relevant literature. Note that not all of the models are included in our initial model set $\mathcal{M}_0$. Jump characteristics are labeled as follows:
(a) no jumps,
(b) norm jumps in rets with const intensity,
(c) async norm jumps in rets and exp jumps in vol with const intensity,
(d) correlated sync norm jumps in rets and exp jumps in vol with const intensity,
(e) async norm jumps in rets and exp jumps in vol with stoch intensity,
(f) correlated sync norm jumps in rets and exp jumps in vol with stoch intensity,
(g) double-tail negative Lomax and positive exp jumps in rets with stoch intensity.
Abbreviations: CEV, constant elasticity of variance; GARCH, generalized autoregressive conditional heteroskedasticity; SV, stochastic volatility.

Specifically, in the CEV (C) case, a g or h indicates, respectively, a CEV parameter fixed at one or three-halves. Moreover, an a signals the presence of a hyperbolic drift component adjoined to the linear drift, while a b is associated with a similar drift specification extended with a parabolic component, completing the full Laurent polynomial drift. A subscript to the first letter signals the dimension of the volatility process, which can have either one (1) or two (2) volatility factors.

The next couple of letters of the alphanumeric code then describe the jump specification. In this jump coding, the first letter indicates the volatility jump information, while the second describes the underlying price jump information. If an empty sign is present, then there is no jump in the corresponding model element. Therefore, $\varnothing\varnothing$ indicates no jumps in either volatility or the underlying. Otherwise, if a jump is present, then either a C or an S is used to indicate a jump with, respectively, a constant or stochastic intensity. A bar above the two-letter combination indicates a correlated synchronized jump in volatility and returns. The type of jump size distribution is indicated by a subscript that can either be e, n or p, corresponding, respectively, to a one-sided exponential, a normal or a negative-sided Lomax distribution. The jump in volatility can only be positive exponential, whereas, in the case of the underlying index, an exponential jump is negative.

To bring this naming convention to life, we provide a few examples. The Heston model, as our baseline model, is an affine (A) stochastic volatility diffusion only ($\varnothing\varnothing$) model with a single volatility process (subscript 1). The labeling for the Heston model is therefore A_1-$\varnothing\varnothing$. In the affine (A) extension by Bates (1996b), the Heston model is augmented with a jump component in the asset price process only that has a constant jump intensity and a normal distribution ($\varnothing C_n$). With the model of Bates (1996b) retaining a single volatility process (subscript 1), we therefore label this model as A_1-$\varnothing C_n$. The variant of this stochastic volatility jump-diffusion model that replaces the constant jump intensity assumption with that of stochastic jump intensity (S) is labeled A_1-$\varnothing S_n$. The double-jump model of Duffie et al. (2000), augments the affine (A) Heston model with (i) a constant jump intensity asset price jump component with normal distribution (C_n) and (ii) a constant jump intensity volatility jump component with exponential distribution (C_e), where it is further assumed that jumps are correlated and occur synchronously. In this case, the notation is A_1-$C_e\bar{C}_n$, where the overbar notation denotes the synchronicity of the jumps. The asynchronous jumps version of this double-jump drops the overbar and is therefore A_1-C_eC_n. Where considered, CEV and log-normal counterparts drop the A labeling and replace it with C and L, respectively. Finally, the modification of the Heston model that includes a hyperbolic drift component adjoined to the linear drift is labeled A_1^a-$\varnothing\varnothing$, while the Heston variant associated with a similar drift specification extended with a parabolic component is labeled A_1^b-$\varnothing\varnothing$. With these examples provided, the naming convention referenced in the remained of the paper should be more accessible.

3 | Model Selection in Option Pricing Models

Option pricing model selection necessarily involves a choice among misspecified models and, hence, requires procedures to achieve prudent conclusions with regard to the lowest amount of misspecification across the models. Appendix A (Supporting Information) reviews how the literature has dealt with this model selection challenge, which provides the context for our positioning of the MCS framework that we propose in Section 3.1.

3.1 | The Model Confidence Set

We present the formal MCS framework we use, introduced in the seminal work of Hansen et al. (2011), which offers advantages over previous testing approaches. The statistical testing procedure involves a sequence of multivariate tests, which terminates with (i) the identification of the superior model set based on a defined performance criterion and (ii) a ranking of the superior models therein. The joint nature of the model comparison test is an important feature given the simultaneous comparison of multiple models. When performing multidimensional statistical testing, the concept of *confidence level*, that is, the probability of committing a false rejection, becomes inadequate. When many dimensions are involved, controlling for a confidence region that covers multiple chances of false rejections is a necessity. This is demonstrated by Romano and Wolf (2010). It is possible, however, to construct joint tests that allow one to control the overall probability of committing *at least* one false rejection, namely, the *familywise error rate* (FWER) (Romano and Wolf 2010). The MCS test is capable by design of assessing model comparisons jointly, controlling for the overall FWER. It provides a distinct partitioning of the initial model set into a set of superior models, that is, the MCS by definition, and a set of inferior models. The MCS includes all models deemed to be statistically equivalent, according to the assumed measure of performance, with different levels of confidence allowing for a model ranking to be determined within the MCS. The MCS thus allows for model selection and model ranking to be performed in a single implementation step.

We define the initial model set as $\mathcal{M}_0 \equiv \{\mathbf{M}_0, \mathbf{M}_1, ..., \mathbf{M}_m\}$, where the $\mathbf{M}_j$ are stochastic models describing market behavior; in our case, price and volatility. The initial model set, $\mathcal{M}_0$, is indexed as the testing procedure iteratively extracts model subsets using an elimination rule until the MCS, $\mathcal{M}^\star$, of equivalently superior models is found. The model comparison requires a measure of model performance through which to establish the MCS. We assume that the performance of model $\mathbf{M}_i \in \mathcal{M}_0$ is measured by a defined loss function L_i, whereby the lower this figure, the better the model performance. In the notation to follow, we omit the dependence on the data sample X_t, $t = 1, ..., T$, and possibly on the model parameter vector θ. In general, however, the model performance is a, possibly parametric, function of the data or functional of the data distribution. The metric of model comparison we use is the relative performance measure $d_{ij} \coloneqq L_i - L_j$, where the ordering matters. The testing procedure presented targets the quantity $\mu_{ij} \coloneqq \mathbb{E}[d_{ij}]$, such that $\mathbf{M}_i \succ \mathbf{M}_j$ ($\mathbf{M}_i$ is preferred to $\mathbf{M}_j$), if $\mu_{ij} < 0$, or $\mathbf{M}_i \sim \mathbf{M}_j$ ($\mathbf{M}_i$ is equivalent to $\mathbf{M}_j$), if $\mu_{ij} = 0$.

With the model comparison metric established, the MCS is thus defined as the subgroup of the initial model set such that each of its elements is either superior or equivalent to any $\mathbf{M}_j$, that is,

$$\mathcal{M}^\star \equiv \{\mathbf{M}_i \in \mathcal{M}_0 : \mu_{ij} \leq 0, \forall\, \mathbf{M}_j \in \mathcal{M}_0\}. \tag{4}$$

Note that (4) defines the collection of models from the initial model set that are deemed to be preferred to all other competing models. With (4) established, we present our test of model comparison along the lines of Hansen et al. (2011), which we coin the max-MCS test.

3.1.1 | The max-MCS Test

The implementation of the max-MCS test deviates from the well-established RC test of White (2000) and the SPA test of Hansen (2005). These tests are designed to examine the null hypothesis of superior performance of a nominated benchmark model against a collection of alternative competitor models. With the max-MCS test, the null is inverted in that it examines the hypothesis of equivalence of the constituent members of the initial model set. A further innovation in the max-MCS test design lies in how it arranges the model comparison. In particular, the max-MCS test circumvents the need to nominate a benchmark model, referring in effect to each model as a 'target' model.

The design of the max-MCS test relies on a strategy already employed in the RC and the SPA, whereby the multiple comparison is reduced to a multiplicity of scalar tests, exploiting the max function. The test sequence produces a stack of rejected models, whose size is unknown at the start of the procedure. In practice, the max-MCS test is given by the sequence of simultaneous *equivalence tests* defined by the following null hypothesis:

$$H_{0,\mathcal{M}_k} : \mu_{i_1 i_2} = \mu_{i_1 i_3} = \cdots = \mu_{i_{m-k-1} i_{m-k}} = 0, \quad \mathbf{M}_{i_u} \\ , \mathbf{M}_{i_v} \in \mathcal{M}_k, i_u < i_v, k = 0, 1, ..., \tag{5}$$

where k is the test sequence index and $i.$ are the model indices at each iteration. If at step k the null, $H_{0,\mathcal{M}_k}$, is rejected then the *elimination rule* is enforced and the model with the worst target statistic is expelled from the model set $\mathcal{M}_k$. The test sequence, therefore, generates the inclusion chain $\mathcal{M}_0 \supset \mathcal{M}_1 \supset \cdots \supset \mathcal{M}_k$, which terminates the first time that the null hypothesis is accepted. The algorithm continues until either the equivalence hypothesis is accepted or the model set becomes a singleton. For the implementation of each equivalence test, Hansen et al. (2011) exploit a mapping of the many hypotheses in (5) to some function, reducing the simultaneous tests to a scalar test. Among several transformations, the authors propose to use the fact that if $\max|\mu_{i,j}| = 0$, then each $\mu_{i,j}$ will be equal to zero. Therefore, the max-MCS test we adopt in the empirical application of this work is defined by the statistic[7]

$$T_{R,\mathcal{M}_k} = \max_{\mathbf{M}_i, \mathbf{M}_j \in \mathcal{M}_k} |t_{ij}|, \tag{6}$$

which involves calculating the sample statistics $\bar{d}_{ij} = \frac{1}{T}\sum_t d_{ij,t}$—where we introduce the time index t, such that $d_{ij,t} \coloneqq L_{i,t} - L_{j,t}$—and their standardized values $t_{ij} = \bar{d}_{ij}/\widehat{\text{var}}(\bar{d}_{ij})$. The basis of the test is the CLT result $\sqrt{T}(\bar{Z} - \mathbb{E}\bar{Z}) \xrightarrow{d} N(0, \Omega)$, as $T \to \infty$, with $\bar{Z}_s^{(k)} = \bar{d}_{ij}$, $\mathbf{M}_i, \mathbf{M}_j \in \mathcal{M}_k$ and $i \neq j$, and with the index for the model pairings $s = 1, ..., \frac{(m-k)(m-k-1)}{2}$. This results from appropriate regularity conditions concerning the time series of the loss function L_i (Hansen et al. 2011).

As the distribution of the max of an n-dimensional normal random variable is not known in closed form, we need to simulate the equivalence hypothesis among the test comparisons as the max of a multivariate normal $\xi^{(k)}$ and hence generate the asymptotic distribution for $T_{R,\mathcal{M}_k}$, which corresponds to the distribution of the $\max_s|\xi_s^{(k)}|$. The variance–covariance matrix Ω_ξ is preemptively estimated with a bootstrap technique. To define a p value for this MCS, we denote by $\hat{p}_{\mathcal{M}_k}$ the p value of the max test at step k, allowing us to then define the sequence (up to a given step k') $\hat{p}_{k'} \equiv \max_{k\le k'}\hat{p}_{\mathcal{M}_k}$. At the last step $k^\star$, when the equivalence hypothesis is accepted and the MCS is identified, we denote by $\hat{p}_i$, for $k^\star \le i \le m$, the p values associated with the best models determined from the max distribution of $\hat{\mathcal{M}}^\star$. With this sequence, we are able to associate a corresponding p value to each model, which measures the probability of that model belonging to the final MCS. These p values provide an index capable of identifying and ranking clusters of equivalent models. See Hansen et al. (2011) for further discussion and illustration of the MCS p value concept.

For the purposes of computing the test, we notice that, when we consider the initial model set at $k = 0$, the multiple hypothesis condition (5) requires only the combinations and not the permutations of paired model comparisons. That is, we avoid the need to include symmetrical hypotheses, as the rejection of one hypothesis would imply the rejection of its corresponding reordered[8] hypothesis. Dismissing the complementary hypothesis in this way is necessary because, as the absolute value function comes into play, the position of the statistics with respect to the zero is irrelevant for acceptance or rejection of the null. That is, $|L_i - L_j|$ or $|L_j - L_i|$ are identical and provide the same statistics. Testing with such redundant hypotheses included would have the effect of artificially inflating the dimension of the problem, leading to a distortion of the correlation structure of the asymptotic test. This represents an important difference between our max-MCS test and the procedure of Hansen et al. (2011).

When considering model performance comparison, the max-MCS test offers an appealing approach that allows one to automatically select the subset of best-performing models as a result of the elimination sequence, while at the same time providing a model ranking based on the associated p values. This capability comes with an important property, which can be summarized as follows. Considering the asymptotic limit of the sample statistic, $\forall\, \mathbf{M} \in \mathcal{M}_0$, *if* the probability of rejecting the equivalence hypothesis when it is true is less than or equal to α, that is, a given significance level, and *if* the probability of rejecting the equivalence when it is false is one, and *if* the probability of rejecting a model that is in $\mathcal{M}^\star$ is zero *then*, indicating with $\hat{\mathcal{M}}^\star_{1-\alpha}$ the estimate of the MCS at the confidence level $(1-\alpha)$, we have that the $P\left(\mathcal{M}^\star \subset \hat{\mathcal{M}}^\star_{1-\alpha}\right) \ge 1-\alpha$ and the probability that an inferior model belongs to the MCS estimate is zero. This is the main property that characterizes the MCS. In fact, the claim that for any model subset, the probability of committing at least one false rejection of the equivalence hypothesis is set to α, corresponds to requiring that the max-MCS provide *strong* control of the FWER. A sufficient condition is the critical values sequence to be monotonic, see Romano and Wolf (2005a).

4 | Experiment Design

Building on the testing framework introduced in Section 3, we are now in a position to design our model comparison exercise for the collection of option pricing models described in Section 2. The model comparison contest is organized as joint tests targeting several alternative loss functions, defining alternative model performance measures. The output of each model comparison contest is the MCS with respect to a given model performance measure, allowing us to draw conclusions about the set of preferred models and the complexity required to execute effective option pricing. To begin, we discuss the option pricing procedure used, followed by details of the parameter estimation procedure. We then present the model performance measures constructed and the techniques used to evaluate these measures.

4.1 | Option Pricing Engine

Towards choosing an option pricing engine to use in our study, we appraise the existing literature. For the pricing of derivatives under the affine model framework, integral transform techniques are commonly used, offering a flexible and computationally efficient approach. Fourier inversion was first introduced by Heston (1993) and then refined for computational efficiency by Carr and Madan (1999) with the use of the fast Fourier transform. A multidimensional Fourier inversion technique is presented in Shephard (1991a, 1991b). With respect to the log-normal model, a quasianalytic solution exploiting the Laplace transform is found in Perelló et al. (2008), whereas alternative models for derivative pricing can in general be treated with simulation, see, for instance, Platen and Bruti-Liberati (2010) who construct a multidimensional jump-diffusion simulation with a predetermined order of convergence. Another technique traditionally used for pricing alternative models is the recombining tree method, introduced in finance by Cox and Ross (1976). Further numerical applications for system dynamics, such as those of the nonlinear drift models, are represented, for instance, by the quasianalytic approximation of the price density constructed in Aït-Sahalia (1999, 2008).

In our study, we decide on an option pricing method that can be consistently applied across the entire suite of affine and non-affine models. We employ partial integro-differential equation (PIDE) solution techniques for the generality and flexibility they bring to the derivative pricing problem. A finite difference method combined with numerical integration is used to construct the solution of the pricing equation for the reference model set. See, for example, Tavella and Randall (2000) and Duffy (2006) for reviews of the method and some applications.

To obtain a pricing function, we construct a numerical solution of the backward equation defining the price of a European call/put option price written on an underlying S_t, with maturity T and strike price K:

$$\mathbb{E}^{\mathbb{Q}}[(S_T - K)^{\pm}|\mathcal{F}_t]\mathrm{e}^{-r(T-t)} \tag{7}$$

with the interest rate r assumed nonstochastic and $\mathcal{F}_t$ a filtration.[9] We employ the martingale approach to achieve the option price and construct the risk-neutral dynamics accordingly In the literature, it is acknowledged that when the market is not complete, that is, when some risk factors are not traded, the pricing equation obtained via the hedging argument is not unique, see, for instance, Scott (1987). This is the case for the models we study. Thus, the market equilibrium argument is invoked and a market risk premium is introduced. The introduction of the risk premium usually has the effect of altering the parameters of some risk factors, without changing the overall structure of the main stochastic differential equation (SDE), see Bates (2000). Nonetheless, the analysis of the effect of market risk premia on the main equation is beyond the scope of this article. For an interesting analysis pertaining to jump-diffusion models, see, for instance, Pan (2002). For all practical purposes, we assume that the risk-premia are absorbed into the appropriate model parameters.

A simplification introduced in the analysis that reduces considerably the cost of computation, is as follows. The strike price K in (7) is extracted from the pay-off function, and the (terminal) underlying price is redefined as $s_T = S_T/K$. Hence, the new factor s evolves with the same dynamics as in (1a) and (1b) with initial condition s_t. The choice of standardizing by the strike gives s_t the interpretation of a dimensionless measure of "moneyness." As a result, the solution domain is centered over an appropriately chosen range and is initiated by a unique terminal condition. The computational cost is therefore reduced because, normalizing the underlying price to an option moneyness range and making the strike price identical for all the options, allow us to produce one single set of solutions for all the available data. This option price standardization contrasts with the use of dollar prices, as in Bakshi et al. (1997), which induces dependence on the index level and, therefore, the time of the price record, and to the alternative option-to-underlying price standardization proposed in Bates (1996a, 2000).

Therefore, for the jth option in the sample array on day t, we define $O_{t,j}$ to be the theoretical price, which corresponds to the present value of the strike times, a stochastic factor depending on $s_{t,j}$ and the model parameters as follows[10]:

$$O_{t,j}(k, K_{t,j}, T_{t,j}, s_{t,j}; \Theta, v_t) = \mathbb{E}^{\mathbb{Q}}[(s_{T,j} - 1)^{\pm}|\mathcal{F}_t]K_{t,j}e^{-r(T_{t,j}-t)}, \tag{8}$$

where the strike price is always unitary and the binomial variable k indicates the type of options contract. Hence, we conduct the analysis in terms of the relative theoretical option prices $\omega_{t,j} = O_{t,j}/K_{t,j}$ compared with the relative market option prices $\pi_{t,j} = P_{t,j}/K_{t,j}$, where $P_{t,j}$ is the market price for the jth option on date t. In the following, unless otherwise explicitly stated, when we write about option prices we will be referring to relative option prices.

4.2 | Parameter Estimation

Before the parameter optimization exercise, the model parameters are assumed to be uniformly distributed across the parametric region. The allowed stationary volatility is between 5% and 70%, and the expected jump size can vary between 0.5 and 3 times the stationary volatility. The total volatility variance explained by the jump component is at most 50% of its asymptotic value. The correlation parameter can change across the negative domain, while, if a CEV factor is present, it ranges between 0.5 and 1.5. Finally, a jump in the level of prices can have an intensity factor between 0.01 and 5, an expected jump size in price between 0% and −30%, with a standard deviation between 1% and 50%, which means that rare jumps are allowed with a wide range of variation. The target function is assumed to be the log-likelihood of the pricing residuals. Despite evidence that calls for the adoption of an autocorrelated model of pricing residuals, see, for instance, Bates (1996a, 2000) or Lindström et al. (2008), we use an independent Gaussian hypothesis for the residuals, as it reduces the computational burden.[11] The likelihood is characterized by assuming that the pricing error is null and the daily variance is estimated by the cross-section mean squared error (MSE). The full sample experiment determines the posterior of the model parameters.

The estimation of the parameter posterior is conducted using a Bayesian procedure. The parametric likelihood ascertained from the experimental results H_1 but before the evaluation of the hypothesis concerning a zero-sum pricing error A is

$$\mathbf{P}(\Theta|H_1, A) \propto \mathbf{P}(\Theta|H_0, A)\mathbf{P}(Y|H_0, A, \Theta), \tag{9}$$

where H_0 represents the prior hypothesis about the parameter distribution and Y is the sample option pricing residuals. As a consequence, the experimental results $H_1 = \{Y, H_0\}$ are the events represented by the mispricing values obtained traversing the parametric space. The residual likelihood is obtained through the program

$$\mathbf{P}(Y|H_0, A, \Theta) \equiv \max_{V} \mathbf{P}(Y|H_0, A, \Theta, V), \tag{10}$$

where $V = \{v_k\}_{k=1}^{T}$ is the IV time series. We remark that, although the volatility probability structure defines the option pricing function, for the construction of the likelihood in (10), no filtering technique has been applied, for example, Bates (2000), which implies that no coherency between the cross-section of the option prices and the time-series dimension of the latent factor has been imposed. Therefore, each daily volatility observation is the result of an independent optimization procedure. The rationale for this methodological choice is twofold. First, we achieve a simplification of the optimization problem, leading to computational efficiency that is important, given the large option data panel that we use in our study. Second, we obtain volatility paths that are completely price implied and v is free to adapt to the shape of the market price surface, according to the model specification. In this regard, the solutions obtained represent an optimum for the sake of pricing, whereas imposing IV path coherency might achieve the same mispricing results only in the best possible scenario, all else being equal. Although the estimated volatility might not necessarily match its transitional probabilities, to a certain extent, the parametric structure of the pricing function does influence the volatility path, as its variations are required to optimally match the $t + 1$ prices. This procedure disregards the evaluation of the transitional

probability density of the implied factor, instead the IV represents a parametric array and the problem assumes the nature of a multistage optimization. In the following, however, we will introduce a model performance measure that allows us to explicitly test the coherency between the implied probability structure and the volatility trajectory produced by each model via (10).

The density in (9) is then combined with the mispricing posterior to obtain

$$\mathbf{P}(\Theta, A|H_1) = \mathbf{P}(\Theta|H_1, A)\mathbf{P}(A|H_1). \tag{11}$$

The model parameters and the estimation errors are finally obtained as expectations under the posterior distribution in (11). For an application of parameter learning, see West (1993), and for a more general introduction, see Carvalho et al. (2010).

4.3 | Model Performance Measures

With the model set appropriately parametrized, our investigation of model complexity can begin. According to the model set testing procedure established in Section 3.1, we need to define a model performance measure. For our purposes, we construct several model performance measures commonly used in the literature,[12] to establish the MCS under different criteria. This allows for a comprehensive picture as to what models are deemed superior to others and as to whether model complexity, as we have defined it, is strictly necessary to achieve effective option pricing.

We employ three categories of model performance measures. For the first category, we focus on mispricing and define two related measures. We first construct a measure of likelihood to test the goodness-of-fit to market prices that are produced by the model set. We use the actual estimation log-likelihood:

$$L_{t,i}^{(1)} := N_t \frac{(\Sigma_j \varepsilon_{t,ij})^2}{\Sigma_j \varepsilon_{t,ij}^2} + \log\left(\frac{\Sigma_j \varepsilon_{t,ij}^2}{N_t}\right) + \frac{m_i^2}{s_i^2} + \log\left(s_i^2\right), \tag{12}$$

where i refers to a given model and j refers to the jth option observed on day t. The variable N_t indicates the number of securities on the observation day. The variables $\varepsilon_{t,ij}$ represent the individual pricing errors and are given by

$$\varepsilon_{t,ij} := \omega_{t,ij} - \pi_{t,j},$$

that is, the difference between the model theoretical option price and the observed market price. The overall sample mispricing for model i is thus

$$m_i = \frac{\sum_t \sum_{j=1}^{N_t} \varepsilon_{t,ij}}{\Sigma_t N_t}$$

and s_i^2 corresponds to its variance. This likelihood $L^{(1)}$ also corresponds to the objective function of the parameter estimation. We proceed with a discrete-time model across a discrete parametric space, optimizing the volatility trajectory for each time unit at each parameter point on a fairly large parametric grid, which consists of the intersection of parameter intervals. Conditioning on the parameter vector value, the option pricing function is monotonic with respect to the volatility value at each time point, thus a grid search is performed at each cross-section. Formally, the procedure is described as follows:

$$\min_{\theta} \{\min_{v} L^{(1)}(\theta, v)\},$$

where θ is the parameter vector value and

$$v = \{v_1, ..., v_T\}$$

is the volatility sequence. The result can be improved by progressively updating the parameter prior distribution.

As a second measure, we seek to capture the variability of the mispricing. For this, we consider the root mean squared error (RMSE), which has been used in previous research both for parameterization and model performance measurement, see, for example, Bakshi et al. (1997) and Duffie et al. (2000) for an application of the MSE. Formally, we define the RMSE as follows:

$$L_{t,i}^{(2)} := \sqrt{\frac{\Sigma_j \varepsilon_{t,ij}^2}{N_t}}. \tag{13}$$

Given the full likelihood specification in (12), we expect similar behavior between these two measures, as the full likelihood and the MSE exhibit high sample correlation.[13] Estimation under $L^{(2)}$ follows similarly to that described for $L^{(1)}$.

The second category of model performance measure we adopt seeks to appraise model hedging performance. We follow Bakshi et al. (1997) in designing the hedging strategy to support the construction of our measure of model hedging performance. As the presence of nontraded risk factors dictates, we construct a measure of min-variance hedging performance. The intuition is to measure the sample variance of the discrepancies between the variations of the replicating portfolio, determined by the delta exposure times of the underlying variations, compared with the actual variations of the corresponding option prices. We start by determining the individual option per unit time underlying exposure as the min-variance delta $X_{t,ij}$, that is,

$$X_{t,ij} = \frac{d[s_{t,j}, \omega_{t,ij}]}{d[s_{t,j}, s_{t,j}]},$$

where $d\left[u_t, w_t\right]$ represents the instantaneous cross-variation between the stochastic processes u_t and w_t. Hence, we construct the loss function as the variance of the hedging errors, that is, the difference between the replicating portfolio and the corresponding option variations. The error is defined as follows[14]:

$$\varepsilon_{t,ij} = X_{t-\Delta,ij}(s_{t,j} - s_{t-\Delta,j}) - (\pi_{t,j} - \pi_{t-\Delta,j}), \quad j \in U_{t-\Delta},$$

where $U_{t-\Delta}$ contains the intersection of the $t - \Delta$ and t day price array. Besides the fact that the hedging error represents another

measure of model ability to correctly reproduce market prices, the replicating portfolio discrepancy is also an out-of-sample measure, that is, it measures the ability of the model to project forward market prices. We depart somewhat from the measure constructed in Bakshi et al. (1997), where instead of tracking an individual replicating portfolio performance, we measure the hedging error across the entire sample. We track all of the possible replicating portfolios and our main target is the daily hedging error variability. Therefore, the loss function is represented by

$$L_{t,i}^{(3)} := \frac{\Sigma_j \varepsilon_{t,ij}^2}{N_t} - \left(\frac{\Sigma_j \varepsilon_{t,ij}}{N_t}\right)^2, \tag{14}$$

which is the hedging error variance. This measure allows for model comparison of economic relevance in an option pricing context, also giving insights into model forecasting ability.

Finally, the third loss function category we employ is devised to target the internal consistency of each option pricing model in respect of the IV trajectories that maximize (10) at the termination of the calibration procedure. With this loss function, we wish to measure the coherency of the model implied probability structure and the output behavior of the volatility trajectory. Although methodologically different, similar test procedures have been employed in, for instance, Eraker et al. (2003). We follow this study in constructing the standardized residuals:

$$\varepsilon_{t,i} = \frac{(v_{t,i} - v_{t-\Delta,i}) - [\mu(v_{t-\Delta,i}) - \lambda_1 \bar{z}_1]\Delta}{\sigma(v_{t-\Delta,i})\sqrt{\Delta}} \sim N(0,1),$$

which corresponds to the standardization of the Euler scheme rendered variable v. Therefore, it is straightforward to define the model performance measure in terms of the Cramer–Smirnov type statistic, see Darling (1957). That is,

$$\begin{aligned} L_i^{(4)} := \tfrac{1}{N}\Big[&\Big(\Phi_{(-\infty,-5)} - \tfrac{1}{N}\Sigma_t \mathbf{I}_{\{\varepsilon_{t,i}<-5\}}\Big)^2 \\ &+ \Big(\Phi_{(5,\infty)} - \tfrac{1}{N}\Sigma_t \mathbf{I}_{\{\varepsilon_{t,i}\geq 5\}}\Big)^2 \\ &+ \Sigma_{h=1}^{K}\Big(\Phi_{(-5+(h-1)M,-5+hM)} - \tfrac{1}{N} \\ &\Sigma_t \mathbf{I}_{\{-5+(h-1)M\leq\varepsilon_{t,i}<-5+hM\}}\Big)\Big]^2, \end{aligned} \tag{15}$$

where $\Phi_{(a,b)}$ is assumed to indicate the probability mass of the standard normal distribution within the referenced interval, $M = 10/K$, where K is an integer of choice,[15] N is the number of observed points and $\mathbf{I}_{\{\cdot\}}$ is the indicator function. The function in (15) measures the distance between the empirical distribution of the volatility standardized residuals and the standard normal and provides a measure of how coherent model i has been in modeling IV.

With the model performance measures now established, the experimental design is complete and ready for implementation in the forthcoming section. In particular, we are in a position to run the MCS-based model comparison exercise.

5 | Empirical Testing

In this section, we run several MCS tests constructed on the mispricing, hedging, and IV coherency measures of model performance (Section 4.3). We seek to answer our research question of whether incremental model complexity is justified by incremental model performance. The objective is to isolate from the initial model set (Section 2.2), the subset of superior-performing models deemed to be statistically equivalent based on a given performance measure. The procedure we employ allows us to identify the MCS, comprising the best-performing models, and to produce a model ranking defined by p value measures that give the individual probability of each model belonging to the estimated MCS (Section 3.1). Appendix B (Supporting Information) provides an overview of the S&P500 index options data set that we use (U.S. Options Price Reporting Authority 2015),[16] along with the model parameter estimates obtained from the estimation procedure, with comparisons made to the past literature. A discussion of the absolute model performance of the candidate models is also provided in the Supporting Information Appendix as context for the relative model performance presented in this section.

5.1 | Absolute Model Performance

We first provide insights into the absolute performance of each model. In Table 2, we observe the sample average of several key measures of absolute model performance based on the out-of-the-money (OTM) sample. Despite the large number of options traded daily, good model performance is widespread across the model set, with the exception of one model class. From the likelihood measure, we get an informal model ranking from which we note that models with constant jump intensity in the underlying tend to have considerably inferior performance, being unable to beat the Heston A_1-ØØ model, while models with jumps in volatility generally worsen the performance of the corresponding model specifications without jumps. These models also achieve relatively high RMSEs. In general, the mispricing among the top-performing models ranges between 0 and 5 basis points, with an RMSE between 40 and 50 bps. If we limit our screening to models with a minimal sample RMSE between 40 and 45 bps, we find (i) all of the affine models with stochastic intensity jumps in the underlying process, including the double-jump model specifications augmented with jumps in volatility (except for the double-jump model with correlated jumps), (ii) all of the CEV and ARCH diffusion models with stochastic intensity jumps, (iii) the hyperbolic drift augmented A_1^a-ØØ model, (iv) the log-normal volatility L_1-ØØ model, and (v) the two-factor volatility models, A_2-ØØ and C_2-ØØ.

Building on the above mispricing evidence, the rightmost columns of Table 2 show the sample average of the min-variance hedging and the (square root) IV coherency measures. In the former case, we notice relatively homogeneous behavior for the top-performing models, with the worst scores evidenced for the two-factor volatility models. The IV coherency measure provides the distance between the empirical distribution of the standardized residuals and that of a normal variable, therefore indicating the probability gap between the

TABLE 2 | Model absolute performance.

	Likelihood	Mispricing (bps)	RMSE (bps)	Mispricing *p* values	Mispricing autocorrelation *t* stat	RMSE autocorrelation *t* stat	Mispricing skewness	Mispricing kurtosis	Min-var hedge (bps)	IV coherency (sqrt)
A_1-∅∅	47,992	1.17	47.39	0.911	39.15	25.96	−1.64	5.96	17.77	0.043
A_1-C_e∅	44,667	1.01	48.92	0.905	39.20	26.28	−1.66	5.66	17.78	0.054
A_1-∅C_n	8373	−59.09	86.94	0.073	39.77	37.86	−0.47	2.45	18.68	0.099
A_1-∅S_n	60,247	−0.50	42.55	0.894	38.93	24.58	−1.59	6.99	17.70	0.038
A_1-$C_e C_n$	7413	−63.91	90.79	0.053	39.76	38.12	−0.45	2.43	18.77	0.098
A_1-$C_e S_n$	55,735	−0.76	43.97	0.885	38.99	24.91	−1.69	6.40	17.71	0.050
A_1-$C_e \bar{C}_n$	27,397	−9.49	60.21	0.694	39.53	28.46	−1.35	4.21	17.95	0.069
A_1-$S_e \bar{S}_n$	47,024	1.65	47.97	0.925	39.14	25.83	−1.79	5.88	17.83	0.054
A_1-∅C_e	12,329	−47.00	75.10	0.150	39.78	37.15	−0.53	2.35	19.03	0.092
A_1-∅S_e	65,580	−1.26	40.61	0.880	38.93	23.72	−1.49	7.35	17.75	0.039
A_1-$C_e C_e$	12,571	−46.11	74.84	0.157	39.79	37.10	−0.55	2.35	19.01	0.087
A_1-$C_e S_e$	60,367	−1.63	42.04	0.869	39.00	24.14	−1.64	6.54	17.76	0.052
A_1-∅C_p	16,673	−41.92	63.62	0.223	39.68	36.26	−0.57	2.50	18.56	0.082
A_1-∅S_p	55,615	−0.16	44.03	0.893	39.04	24.81	−1.60	6.49	17.76	0.040
C_1-∅∅	47,968	1.31	47.50	0.916	38.99	24.50	−1.88	6.21	17.86	0.060
C_1-∅S_n	61,306	−0.59	42.33	0.897	38.61	22.80	−1.94	6.74	17.77	0.053
C_1-∅S_e	66,858	−1.40	40.33	0.883	38.70	22.13	−1.97	6.90	17.82	0.054
C_1^g-∅∅	47,693	1.56	47.72	0.923	38.92	24.52	−1.88	6.32	17.87	0.059
C_1^g-∅S_n	60,995	−0.45	42.48	0.901	38.54	22.75	−1.93	6.80	17.77	0.052
C_1^g-∅S_e	66,152	−1.16	40.62	0.889	38.56	22.13	−1.96	7.02	17.83	0.053
C_1^h-∅∅	40,363	2.49	51.60	0.941	38.59	25.21	−1.66	6.66	17.95	0.068
A_1^a-∅∅	64,908	1.14	42.33	0.941	38.30	22.42	−1.83	6.41	17.75	0.033
C_1^a-∅∅	48,881	2.41	48.86	0.986	35.58	23.06	−1.05	4.61	17.86	0.037
A_1^b-∅∅	36,258	5.54	56.52	0.899	24.56	28.54	0.52	4.40	17.87	0.071
C_1^b-∅∅	35,161	5.59	57.29	0.898	24.70	28.58	0.50	4.31	17.88	0.066
L_1-∅∅	68,232	4.59	40.74	0.920	32.70	22.78	−1.12	7.35	17.71	0.024

(Continues)

TABLE 2 | (Continued)

	Likelihood	Mispricing (bps)	RMSE (bps)	Mispricing *p* values	Mispricing autocorrelation *t* stat	RMSE autocorrelation *t* stat	Mispricing skewness	Mispricing kurtosis	Min-var hedge (bps)	IV coherency (sqrt)
A_2-ØØ	65,910	4.09	41.43	0.986	38.47	23.98	−1.94	7.15	19.54	0.040
C_2-ØØ	64,928	5.02	42.38	0.930	29.65	22.91	−0.84	4.81	22.44	0.050

Note: This table shows several relevant statistics for each model in the reference model set. From the leftmost column, we find: the average likelihood of the $L^{(1)}$ type function; the mispricing expressed in basis points, that is, the average difference between theoretical and observed market prices; the root mean squared error (RMSE) expressed in basis points, that is, the standard deviation of the latter individual quantities; the *t* stat *p* value from testing the significance of the mispricing value (model bias); the *t* stat from testing the significance of the lag 1 autocorrelation for the mispricing time series; the *t* stat from testing the significance of the *lag* 1 autocorrelation for the RMSE time series; the mispricing sample skewness; the mispricing sample kurtosis; the hedging error expressed in basis points; and the squared root of the implied coherency measure. The model specifications are described in Section 2.2, with the model labels described in Section 2.3.
Abbreviation: IV, implied volatility.

hypothetical and the realized model behavior. This absolute measure highlights the optimal performance of the L_1-ØØ model, followed closely by A_1^a-ØØ, C_1^a-ØØ and then A_1-ØS_e. The model performance under the IV coherency measure is varied.

As motivated earlier, while the measures discussed thus far provide an indication of the model ranking, the analysis is informal and the measures tell us nothing about how significant are the differences between models and, hence, to what extent one model should be preferred to another. Furthermore, making formal statements through distributional hypotheses of the pricing residuals is complicated by the levels of skewness, kurtosis, and autocorrelation evidenced in the mispricings (Table 2). The RMSE also exhibits significant autocorrelation, signaling the presence of heteroskedasticity in the residuals. These are indicators that the option pricing models, although performing well along several dimensions, fail to completely explain the observed market evolution.

5.2 | Relative Model Performance: MCS Testing

We move now to our relative analysis based on the MCS approach. Section 3.1.1 sets out the details of the max-MCS test we use. We consider an 85% confidence level for the MCS estimate; hence, we set the significance level $\alpha = 15\%$. In Table 3 we present the results of the max-MCS test outcomes, performed on the OTM sample defined in Appendix D.1 (Supporting Information), while Figures 1–4 provide an accessible visualization of the MCS's reported in Table 3. Note that the MCS is defined by the models listed *above* the MCS cut-off line in the table. For comparative purposes, we list the nearest three models that failed to enter the MCS, which appear *below* the MCS cut-off line. Note that the models are ordered by the *p* value concept defined in Section 3.1.1.

We note that the affine jump-diffusion models with constant intensity jumps in price are systematically rejected under all metrics. In contrast, several models with stochastic intensity jumps in price appear across the MCSs we report, which aligns with the evidence of Bates (2000) and Pan (2002). In previous findings, however, Eraker et al. (2003) provide evidence for the importance of jumps in price, assuming model specifications with constant jump intensity, while, in contrast, Eraker (2004) concludes that jumps, irrespective of specification (including stochastic jump intensity), add little explanatory power to option pricing models. Subsequent analysis, see Yun (2011), shows that the results of Eraker et al. (2003) and Eraker (2004) are related to the low-volatility sample employed. Similar to the regression analysis of Bakshi et al. (1997), Yun (2011) shows that time-varying jump premia are correlated to volatility, providing indirect empirical support for the superiority of stochastic intensity jump model specifications. With our use of the statistically robust MCS testing procedure, we provide corroborating evidence that models with a lack of memory in the timing of price jumps produce inferior performance, while those with memory produce superior performance. In Section 5.3, we consider both low- and high-volatility regimes to check the robustness of this contention.

TABLE 3 | max-MCS test: Out-of-the-money (OTM) sample.

$L^{(1)}$ Likelihood A		$L^{(2)}$ RMSE		$L^{(3)}$ Min-var hedging		$L^{(4)}$ IV coherency	
L_1-$\varnothing\varnothing$	1	L_1-$\varnothing\varnothing$	1	A_1-$\varnothing S_n$	1	L_1-$\varnothing\varnothing$	1
C_1-$\varnothing S_e$	0.998	C_2-$\varnothing\varnothing$	1	L_1-$\varnothing\varnothing$	1	A_1^a-$\varnothing\varnothing$	0.665
A_1-$\varnothing S_e$	0.752	C_1-$\varnothing S_e$	0.999	A_1^a-$\varnothing\varnothing$	0.893	A_1-$\varnothing S_n$	0.331
A_1^a-$\varnothing\varnothing$	0.565	A_1^a-$\varnothing\varnothing$	0.998	A_1-$C_e S_n$	0.783	A_1-$\varnothing S_e$	0.289
A_2-$\varnothing\varnothing$	0.562	A_2-$\varnothing\varnothing$	0.943	A_1-$C_e\varnothing$	0.665	C_1^a-$\varnothing\varnothing$	0.182
C_2-$\varnothing\varnothing$	0.475	A_1-$\varnothing S_e$	0.867	A_1-$\varnothing\varnothing$	0.614	A_1-$\varnothing S_p$	0.173
C_1^g-$\varnothing S_e$	0.124	C_1^g-$\varnothing S_e$	0.557	A_1-$\varnothing S_p$	0.550	A_1-$\varnothing\varnothing$	0.140
C_1^g-$\varnothing S_n$	0	A_1-$\varnothing C_p$	0.080	A_1-$C_e S_e$	0.449	A_2-$\varnothing\varnothing$	0.078
C_1-$\varnothing S_n$	0	C_1^g-$\varnothing S_n$	0.006	A_1-$\varnothing S_e$	0.357	A_1-$C_e S_n$	0.036
		C_1-$\varnothing S_n$	0.001	C_1^g-$\varnothing S_n$	0.300		
				C_1-$\varnothing S_n$	0.292		
				A_1^b-$\varnothing\varnothing$	0.121		
				A_2-$\varnothing\varnothing$	0.121		
				C_1-$\varnothing S_e$	0.103		
				C_1^g-$\varnothing S_e$	0.085		

Note: This table shows the max-MCS test results for the OTM sample as defined in Supporting Information Appendix B.1 and for each model performance measure $L^{(i)}$ as defined in Section 4.3. The MCS is defined by the models listed above the MCS cut-off line. For comparative purposes, we list the nearest three models that failed to enter the MCS, which appear below the MCS cut-off line. The models are ordered by p value. The confidence level for the MCS test is set to 15%. All the models are sorted by their MCS p value. The max-MCS test is described in Section 3.1.1. The model specifications are described in Section 2.2, with the model labels described in Section 2.3. Abbreviations: IV, implied volatility; MCS, model confidence set; RMSE, root mean squared error.

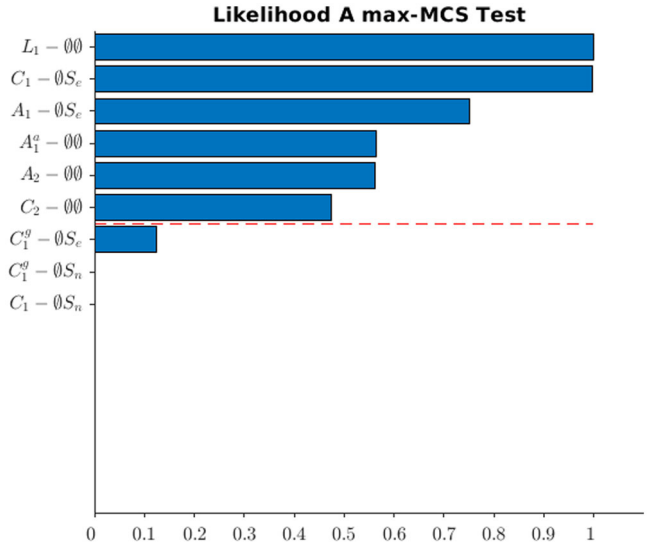

FIGURE 1 | max-MCS test visualization (likelihood A): Out-of-the-money (OTM) sample. *Note:* This figure visualizes the max-MCS test results for the OTM sample as defined in Supporting Information Appendix B.1 and for the likelihood A performance measure as defined in Section 4.3. This is a visualization of the first column in Table 3. The MCS is defined by the models listed to the left of the MCS cut-off line (dashed line). For comparative purposes, we list the nearest three models that failed to enter the MCS, which appear to the right of the MCS cut-off line. MCS, model confidence set. [Color figure can be viewed at wileyonlinelibrary.com]

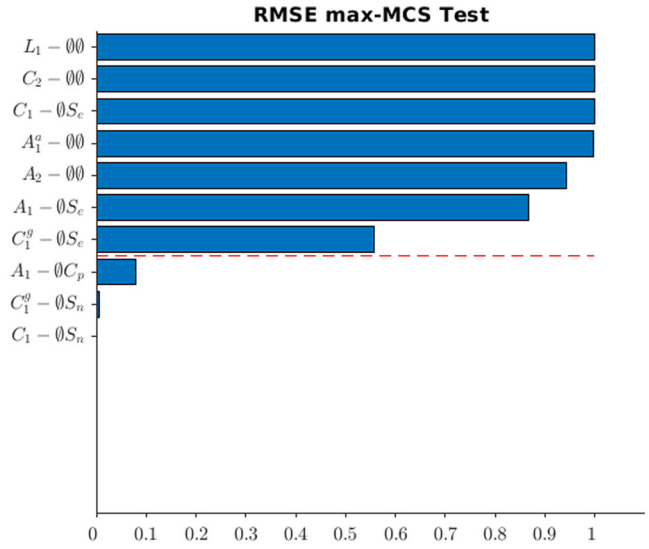

FIGURE 2 | max-MCS test visualization (RMSE): Out-of-the-money (OTM) sample. *Note:* This figure visualizes the max-MCS test results for the OTM sample as defined in Supporting Information Appendix B.1 and for the RMSE performance measure as defined in Section 4.3. This is a visualization of the second column in Table 3. The MCS is defined by the models listed to the left of the MCS cut-off line (dashed line). For comparative purposes, we list the nearest three models that failed to enter the MCS, which appear to the right of the MCS cut-off line. MCS, model confidence set; RMSE, root mean squared error. [Color figure can be viewed at wileyonlinelibrary.com]

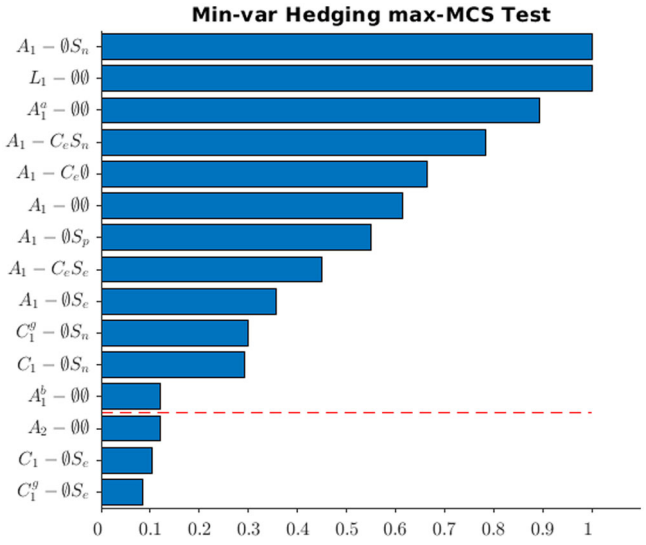

FIGURE 3 | max-MCS test visualization (Min-var Hedging): Out-of-the-money (OTM) sample. *Note:* This figure visualizes the max-MCS test results for the OTM sample as defined in Supporting Information Appendix B.1 and for the Min-var Hedging performance measure as defined in Section 4.3. This is a visualization of the third column in Table 3. The MCS is defined by the models listed to the left of the MCS cut-off line (dashed line). For comparative purposes, we list the nearest three models that failed to enter the MCS, which appear to the right of the MCS cut-off line. MCS, model confidence set. [Color figure can be viewed at wileyonlinelibrary.com]

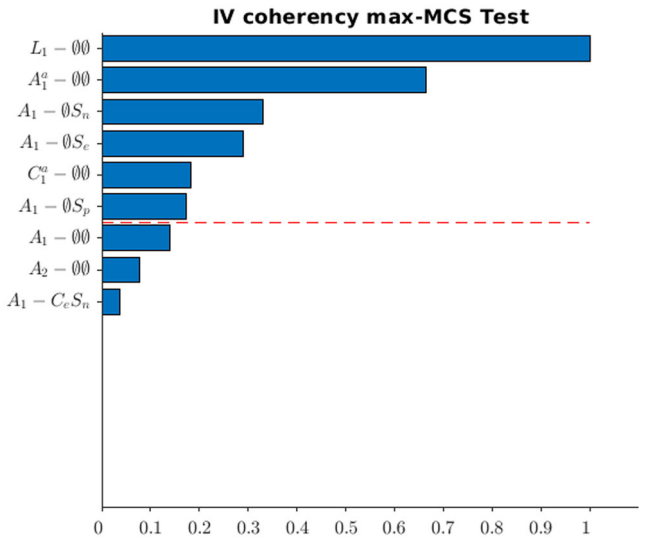

FIGURE 4 | max-MCS test visualization (IV coherency): Out-of-the-money (OTM) sample. *Note:* This figure visualizes the max-MCS test results for the OTM sample as defined in Supporting Information Appendix B.1 and for the IV coherency performance measure as defined in Section 4.3. This is a visualization of the fourth column in Table 3. The MCS is defined by the models listed to the left of the MCS cut-off line (dashed line). For comparative purposes, we list the nearest three models that failed to enter the MCS, which appear to the right of the MCS cut-off line. IV, implied volatility; MCS, model confidence set. [Color figure can be viewed at wileyonlinelibrary.com]

With respect to the form of the price jump size distribution, negative exponential jumps appear to produce higher performance across the greatest number of performance measures, although other jump specifications do appear in several MCSs, including the normal distribution. This is notable given that much of the existing literature advocates a normal distribution for price jumps, while the negative exponential distribution is defined by a single parameter rather than the two parameters of the normal distribution.

We also find little evidence that jumps in volatility are of relevance. Moreover, double-jump models with either synchronous or asynchronous jumps in volatility, whenever present in the MCS, are generally outperformed by the model counterpart without jumps in volatility, signaling that jumps in volatility generally hurt model performance. This is reflected, for instance, in the higher rankings for the A_1-$\emptyset S_e$ and A_1-$\emptyset S_n$ models within the reported MCSs for all performance measures.

Another constituent model of note within the reported MCSs is the CEV C_1-$\emptyset S_e$, with stochastic intensity jumps in price of negative exponential distribution—the analog of the affine A_1-$\emptyset S_e$ model. The increased complexity of the elasticity of variance extension above its affine counterpart leads to superior performance across the mispricing-based likelihood $L^{(1)}$ and RMSE $L^{(2)}$ measures. This broadly aligns, for instance, with Jones (2003), who concludes that CEV models produce superior performance than the Heston model. However, we show that the plain CEV model, C_1-$\emptyset\emptyset$, does not feature in the MCS. Therefore, it is necessary to augment the CEV model with a parsimonious jump in returns with stochastic intensity to achieve superior pricing performance. However, while this is true from a pricing perspective, the same model produces inferior performance relative to the A_1-$\emptyset S_e$ model when assessed on the basis of the hedging $L^{(3)}$ and the IV coherency $L^{(4)}$ tests.

Other significant models of note are the quasiaffine A_1^a-$\emptyset\emptyset$ and the log-normal L_1-$\emptyset\emptyset$ models, which show superiority across all performance measures, while the two-factor CEV C_2-$\emptyset\emptyset$ model is consistently a member of the mispricing-derived MCSs. The latter model, however, does not produce superior hedging or IV coherency. The quasiaffine and log-normal models are among the most successful model specifications. Both the L_1-$\emptyset\emptyset$ model and A_1^a-$\emptyset\emptyset$ models are highly coherent in terms of IV, but the former is more highly ranked. This evidence pertaining to the L_1-$\emptyset\emptyset$ model contrasts with Benzoni (2002), who concludes that the Heston and log-normal models have similar performance.

If considered just in terms of parsimony of parameters, it is striking that the quasiaffine A_1^a-$\emptyset\emptyset$ model manages to provide top-range performance, throughout the OTM sample, at the cost of just one auxiliary component in the volatility drift relative to the Heston A_1-$\emptyset\emptyset$ model. As has been suggested by studies, such as G. Li and Zhang (2013), the affine drift of the state variable contributes to model misspecification. By inserting the hyperbolic term in the affine drift, we are able to achieve high performance for this augmented model. We conjecture that there is a technical explanation for this behavior. We notice that the asymptotic variance of the Heston model can be written as

the product of the model constraint $\frac{\sigma^2}{2a}$ times the squared long-run mean.[17] This implies that a plain affine model cannot generate a volatility variance that is larger than its own squared mean, which strongly constrains the peaks achievable by the model trajectories and, therefore, the kurtosis of the underlying returns. On the contrary, a model such as A_1^a-ØØ, which prevents the volatility from hitting the zero boundary, allows the volatility of volatility to grow without limitation, giving the model more flexibility to reproduce actual IV behavior.

While the quasiaffine A_1^a-ØØ, as a drift-based augmentation of the Heston A_1-ØØ model, performs particularly well, it is notable that the Heston model produces superior, though midrange, hedging variance performance. We further notice that the Heston model sits marginally outside the MCS when assessed on IV coherency. This consideration suggests that if one is able to accept some loss of performance with respect to pricing, the least complex model presented in our model suite can be exploited for hedging and volatility modeling purposes.

In respect of our research question pertaining to model complexity, it is worth summarizing our findings thus far. We provide evidence that moving from a constant price jump intensity to a more complex stochastic jump intensity specification is justified. We find no real evidence, though, to support increased complexity through the integration of jumps in volatility, whether individually or jointly with price jumps. We see that a more parsimonious negative exponential price jump distribution performs relatively better than other distributions, although the normal and Lomax distributions are generally acceptable. We further find that the more parsimonious quasiaffine A_1^a-ØØ and log-normal L_1-ØØ models perform particularly well relative to more complex jump specifications. In respect of the models we reject, we get further insights into our research question. We see no evidence to support the increased model complexity induced by assuming CEV specifications, with the exceptions being the price jump-augmented version of the single-factor CEV model and the two-factor CEV model. We also find no evidence for parabolic nonlinearity in the drift.

5.3 | Segmented Analysis

Building on the results to date, we extend our study with a segmented analysis. Specifically, we perform the model selection exercise across several alternative options samples to assess whether there are specific moneyness and maturity effects, while we also consider model performance across alternative volatility regimes. To check the robustness of our findings, we perform the model selection exercise across several alternative options samples to assess whether there are specific moneyness and maturity effects, while we also consider model performance across alternative volatility regimes. Specifically, we consider the following samples: a deep-out-the-money sample, that is, call options with log-moneyness lower than −7.5% and put options with log-moneyness greater than 7.5%; a quasi-at-the-money (QATM) sample, that is, call options with log-moneyness between −2.5% and 0 and put options with log-moneyness between 0 and 2.5%; a long-term-to-maturity (LTTM) sample, that is, OTM options with tenors between 6 and 12 months; a short-term-to-maturity (STTM) sample, that is, OTM options with tenors between 1 and 3 months; a low-volatility environment sample, that is, OTM options selected with a measure of IV below the sample median; and a high-volatility environment (HVOTM) sample, that is, OTM options selected with a measure of IV above the sample median.

To conserve space, we defer the full discussion of the maturity and volatility regime analysis to Supporting Information Appendix C (Supporting Information). In summary, we find that results broadly align with the full OTM sample, but with some variations, particularly in the QATM and LTTM sample. The RMSE measure is notably lower for the STTM sample, while higher in the HVOTM sample. Stochastic price jump intensity models outperform, with jumps in volatility playing a minimal role, which supports our main findings. Additionally, EVD models show promise for long maturities, especially in the LTTM case. The volatility-segmented analysis shows that model performance differs significantly in high-volatility conditions, with hedging models performing similarly across regimes, yet showing some unexpected behavior in the HVOTM sample.

6 | Conclusion

In this study, we tackle a research question of relevance to academics and practitioners. We investigate if increasing complexity in option pricing modeling is justified by commensurate improvements in model performance. We consider an index option market context and propose an overarching modeling framework that captures many of the affine and nonaffine jump-diffusion model specifications deployed to date in the literature to jointly describe the underlying equity index and the associated stochastic volatility dynamics. We define model complexity as departures from the seminal stochastic volatility diffusion model of Heston (1993). We consider several specifications of jumps in price and volatility, across the affine and CEV model classes, while we consider nonlinearity in the drift and diffusion, in addition to two-factor volatility model specifications. We construct an option pricing model comparison exercise involving a large data set of traded index options. The performance of the model set is measured and tested for model superiority using a range of common model comparison indicators, providing information regarding mispricing behavior, hedging performance, and internal coherency with respect to modeling the volatility path. Differing from the existing literature, we use a statistically rigorous model selection approach that is premised on the MCS methodology of Hansen et al. (2011). We provide insights into the trade-off between model complexity and model performance.

Our empirical evidence suggests that there is a payoff to some forms of model complexity. In particular, jumps in the price component are a necessary extension to produce superior outcomes. Moreover, we find that the type of jump is important. In fact, constant intensity jumps in price drastically reduce performance, whereas stochastic intensity jumps seem to be the preferable choice. We find that a negative exponential appears to be the best choice for the price jump distribution form, although normal and EVD price jumps generally produce

equivalent performance. Jumps in volatility, either synchronized or not, worsen model performance, suggesting it is not worth the augmented size of the parametric vector. Furthermore, we provide evidence that extending to a two-factor model incorporating a stochastic mean volatility factor improves model performance without the need for a jump. We also gather evidence that the affine drift is a significant source of misspecification, as a simple extension of the Heston model with a hyperbolic factor drastically improves performance. Most notably, from the models rejected, we find, for the most part, that the CEV model class, which is the most popular nonlinear extension to the affine model class, does not seem to produce appreciable improvements in performance.

It is important to note that the testing framework we adopt—namely, the max-MCS—supports multiple comparisons across complex models, providing a general methodology that can potentially target any type of model performance measure. Nonetheless, care should be taken during the testing design stage as excessive simplification of the underlying hypotheses might weaken the testing results. However, the flexibility of the MCS approach facilitates the construction of extensive suits of tests that can provide different points of observation for the overall model performance significance. The intersection of the battery of tests can provide insightful information with respect to model superiority and area where model refinement is required.

We believe our work will encourage the use of the MCS as a minimum statistical standard for researchers when proposing new option pricing models and benchmarking against existing models, particularly when done on a large scale. The MCS is conservative, however, in its control of the FWER defined as the probability of making at least one false discovery. Controlling for the FWER in this way, however, lacks power, where power is loosely defined as the ability to reject false null hypotheses, that is, to identify true discoveries. Our study should motivate research in the direction of *generalized* approaches that offer greater statistical power when correcting for multiple comparisons bias in such multimodel testing, see Beran (1988), Lehmann and Romano (2005), Romano and Wolf (2005a, 2007, 2010). This would be particularly important, for instance, where an even larger suite of option pricing models is analyzed.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Our database comprises S&P500 index (SPX) options traded on the Chicago Board Options Exchange (CBOE) and made available by the US Option Pricing Reporting Authority (OPRA) (https://www.opraplan.com/). Restrictions apply to the availability of these data, which were used under license for this study.

Endnotes

[1] Note that model parameters do not enter the sample statistic constructed under the MCS framework. In this sense, the number of parameters of a given model is not a factor in the calculation of the MCS. In option pricing model specification analysis, it is well established that an option pricing model with more parameters is not necessarily going to lead to better performance (Bakshi et al. 1997). Hence, the MCS is a suitable framework for us to compare different models with different parameter dimensions.

[2] The Lomax distribution corresponds to a Pareto Type II distribution that has been shifted to initiate the range of variation at zero. The EVD negative tail distribution has been chosen to test an implicit hypothesis on the LTTM behavior of the pricing function.

[3] See also other studies in the context of option pricing models, such as Huang and Wu (2004), H. Li et al. (2008), and Yang and Kanniainen (2017).

[4] The risk-neutral drift is the process such that the present value of the underlying price results in a martingale. The $\tilde{r}$ process corresponds to the interest rate process r in the absence of jumps, and it is otherwise adjusted appropriately to compensate for the bias introduced by the discontinuous price component. The coefficient r is the constant interest rate corresponding to the 3-month Treasury bill. For pricing modeling purposes, the interest rate parameter is assumed to be constant, though during the optimization it is observed daily.

[5] The articles Kou (2002) and Kou and Wang (2004) deal with a double-exponential jump size distribution. We consider a single-sided negative distribution as a hypothesis on the jump-induced skewness.

[6] From unreported analysis, we acknowledge that the impact of stochastic intensity jumps in volatility is marginal, w.r.t. the likelihood measure, whereas the self-exciting feature of the jump in this model renders it more troublesome to be estimated.

[7] To aid the reader, we make the Matlab code that implements the max-MCS test in (6) available through Matlab Exchange here (https://uk.mathworks.com/matlabcentral/fileexchange/175383-mht_mcs_max). The Matlab code is fully downloadable, and instructions for the function usage are included in the function help section.

[8] When testing for (5), we need to keep track of which model determines the negative value of $t_{i,j}$ to reject the correct model.

[9] Although the interest rate is assumed nonstochastic, daily interest rate dynamics may be captured by the use of 3 month T-Bill quotations as a proxy for the risk-free rate. In this article, we do not incorporate the effect of stochastic interest rate into the pricing function. This impact is assumed to be small on shorter maturity options and progressively increasing with the tenor. See, for instance, Scott (1997) for a detailed analysis.

[10] The indices t, j attached to the maturity date T and the strike K of the jth option in the t day sample, do not change the constant nature of these variables. They indicate instead, respectively, the reference panel date and the position of the corresponding option within the panel, which is not necessarily constant.

[11] During the estimation, the sample is represented by 1445 observation days containing on average 845 prices, ranging from a minimum of 34 to a maximum of 5594. A grand total of 2,967,861 observations are used. We discuss the data in detail in Supporting Information Appendix B.1.

[12] In Section 3.1, we defined the MCS test with respect to a loss function. It is straightforward that a model performance measure is the opposite of a loss function, that is, the lower the loss then the better the performance. Formally, they are the same function, though when we refer to a superior model performance, the actual test result will show an inferior loss. We will use the two terms interchangeably, whenever no confusion arises.

[13] In the exercise produced in this article, the likelihood (12) and the MSE show a −0.95 correlation.

[14] Notice that across the several loss function definitions, we deliberately use the letter ε to indicate model errors. This is for the purposes

of methodological consistency. Any definition is to be considered confined to the performance measure to which it is associated.

[15] In the empirical exercise, we aggregate the residuals in 35 bins between -5 and 5 plus two bins at $\pm\infty$.

[16] U.S. Options Price Reporting Authority 2015. *S&P 500 Index Options (SPX) Data. Proprietary Data Accessed Under License*. https://www.opraplan.com/.

[17] The presence of a jump in volatility is irrelevant as, asymptotically, the model behaves as a jump-less model with modified parameters.

References

Aït-Sahalia, Y. 1999. "Transition Densities for Interest Rate and Other Nonlinear Diffusions." *Review of Financial Studies* 54: 1361–1395.

Aït-Sahalia, Y. 2008. "Closed-Form Likelihood Expansions for Multivariate Diffusions." *Annals of Statistics* 36, no. 2: 906–937.

Aït-Sahalia, Y., and R. Kimmel. 2007. "Maximum Likelihood Estimation of Stochastic Volatility Models." *Journal of Financial Econometrics* 83: 413–452.

Bakshi, G., C. Cao, and Z. Chen. 1997. "Empirical Performance of Alternative Option Pricing Models." *Journal of Finance* 52, no. 5: 2003–2049.

Bakshi, G., N. Ju, and H. Ou-Yang. 2006. "Estimation of Continuous-Time Models With an Application to Equity Volatility Dynamics." *Journal of Financial Economics* 82, no. 1: 227–249.

Bates, D. 1996a. "Jumps and Stochastic Volatility: Exchange Rate Processes Implied in Deutsche Mark Options." *Review of Financial Studies* 9: 69–107.

Bates, D. 1996b. "Testing Option Pricing Models." In *Statistical Methods in Finance, Handbook of Statistics*, edited by G. S. Maddala and C. R. Rao, Vol. 14, 567–611. Elsevier.

Bates, D. 2000. "Post-'87 Crash Fears in the S&P 500 Futures Option Market." *Journal of Econometrics* 94, no. 1–2: 181–238.

Bates, D. 2003. "Empirical Option Pricing: A Retrospection." *Journal of Econometrics* 116, no. 1–2: 387–404.

Beckers, S. 1980. "The Constant Elasticity of Variance Model and Its Implications for Option Pricing." *Journal of Finance* 35, no. 3: 661–673.

Benzoni, L. 2002. "Pricing Options Under Stochastic Volatility: An Empirical Investigation." Carlson School of Management Working Paper.

Beran, R. 1988. "Balanced Simultaneous Confidence Sets." *Journal of the American Statistical Association* 83, no. 403: 679–686.

Broadie, M., M. Chernov, and M. Johannes. 2007. "Model Specification and Risk Premia: Evidence From Futures Options." *Journal of Finance* 62, no. 3: 1453–1490.

Broadie, M., and J. Detemple. 2004. "Option Pricing: Valuation Models and Applications." *Management Science* 50, no. 9: 1145–1177.

Carr, P., and D. Madan. 1999. "Option Valuation Using the Fast Fourier Transform." *Journal of Computational Finance* 2: 61–73.

Carr, P., and L. Wu. 2003. "The Finite Moment Log-Stable Process and Option Pricing." *Journal of Finance* 58, no. 2: 753–777.

Carr, P., and L. Wu. 2004. "Time-Changed Lévy Processes and Option Pricing." *Journal of Financial Economics* 71, no. 1: 113–141.

Carvalho, C. M., M. S. Johannes, H. F. Lopes, and N. G. Polson. 2010. "Particle Learning and Smoothing." *Statistical Science* 25, no. 1: 88–106.

Chan, K., G. Karolyi, F. Longstaff, and A. Sanders. 1992. "An Empirical Comparison of Alternative Models of the Short-Term Interest Rate." *Journal of Finance* 47, no. 3: 1209–1227.

Chernov, M., A. Gallant, E. Ghysels, and G. Tauchen. 2003. "Alternative Models of Stock Price Dynamics." *Journal of Econometrics* 116: 225–257.

Chourdakis, K., and G. Dotsis. 2011. "Maximum Likelihood Estimation of Non-Affine Volatility Processes." *Journal of Empirical Finance* 18, no. 3: 533–545.

Christoffersen, P., S. Heston, and K. Jacobs. 2009. "The Shape and Term Structure of the Index Option Smirk: Why Multifactor Stochastic Volatility Models Work so Well." *Management Science* 55, no. 12: 1914–1932.

Christoffersen, P., K. Jacobs, and K. Mimouni. 2010. "Volatility Dynamics for the S&P500: Evidence From Realized Volatility, Daily Returns, and Option Prices." *Review of Financial Studies* 23, no. 8: 3141–3189.

Conley, T., L. Hansen, E. Luttmer, and J. Scheinkman. 1997. "Short-Term Interest Rates as Subordinated Diffusions." *Review of Financial Studies* 10: 525–577.

Cox, J., and S. Ross. 1976. "The Valuation of Options for Alternative Stochastic Processes." *Journal of Financial Economics* 3, no. 1–2: 145–166.

Darling, D. 1957. "The Kolmogorov–Smirnov, Cramer–Von Mises Tests." *Annals of Mathematical Statistics* 28, no. 4: 823–838.

Das, S. R., and R. K. Sundaram. 1999. "Of Smiles and Smirks: A Term Structure Perspective." *Journal of Financial and Quantitative Analysis* 34, no. 2: 211–239.

Duffie, D., and R. Kan. 1996. "A Yield Factor Model of Interest Rates." *Mathematical Finance* 6, no. 4: 379–406.

Duffie, D., J. Pan, and K. Singleton. 2000. "Trasform Analysis and Asset Pricing for Affine Jump-Diffusions." *Econometrica* 68, no. 6: 1343–1376.

Duffy, D. 2006. *Finite Difference Methods in Financial Engineering: A Partial Differential Equation Approach*. Wiley.

Eraker, B. 2004. "Do Stock Prices and Volatility Jump? Reconciling Evidence From Spot and Option Prices." *Journal of Finance* 59, no. 3: 1367–1403.

Eraker, B., M. Johannes, and N. Polson. 2003. "The Impact of Jumps in Volatility and Returns." *Journal of Finance* 58, no. 3: 1269–1300.

Hansen, P. 2005. "A Test for Superior Predictive Ability." *Journal of Business & Economic Statistics* 23, no. 4: 365–380.

Hansen, P., A. Lunde, and J. Nason. 2011. "The Model Confidence Set." *Econometrica* 79, no. 2: 453–497.

Heston, S. 1993. "A Closed-Form Solution for Options With Stochastic Volatility With Applications to Bond and Currency Options." *Review of Financial Studies* 6, no. 2: 327–343.

Huang, J., and L. Wu. 2004. "Specification Analysis of Option Pricing Models Based on Time-Changed Lévy Processes." *Journal of Finance* 59, no. 3: 1405–1439.

Hull, J., and A. White. 1987. "The Pricing of Options on Assets With Stochastic Volatilities." *Journal of Finance* 42, no. 2: 281–300.

Ignatieva, K., P. Rodrigues, and N. Seeger. 2015. "Empirical Analysis of Affine Versus Nonaffine Variance Specifications in Jump-Diffusion Models for Equity Indices." *Journal of Business & Economic Statistics* 33, no. 1: 68–75.

Jones, C. 2003. "The Dynamics of Stochastic Volatility: Evidence From Underlying and Options Markets." *Journal of Econometrics* 116, no. 1: 181–224.

Kaeck, A., P. Rodrigues, and N. Seeger. 2017. "Equity Index Variance: Evidence From Flexible Parametric Jump-Diffusion Models." *Journal of Banking & Finance* 83, no. C: 85–103.

Kou, S. 2002. "A Jump-Diffusion Model for Option Pricing." *Management Science* 48, no. 8: 1086–1011.

Kou, S., and H. Wang. 2004. "Option Pricing Under a Double Exponential Jump Diffusion Model." *Management Science* 50, no. 9: 1178–1192.

Lehmann, E., and J. Romano. 2005. "Generalisation of the Family-Wise Error Rate." *Annals of Statistics* 33, no. 3: 1138–1154.

Lewis, A. 2000. *Option Valuation Under Stochastic Volatility*. Finance Press.

Li, G., and C. Zhang. 2013. "Diagnosing Affine Models of Options Pricing: Evidence From VIX." *Journal of Financial Economics* 107, no. 1: 199–219.

Li, H., M. T. Wells, and C. L. Yu. 2008. "A Bayesian Analysis of Return Dynamics With Lévy Jumps." *Review of Financial Studies* 21, no. 5: 2345–2378.

Lindström, E., J. Ströjby, M. Brodén, M. Wiktorsson, and J. Holst. 2008. "Sequential Calibration of Options." *Computational Statistics & Data Analysis* 52, no. 6: 2877–2891.

Macbeth, J., and L. Merville. 1980. "Tests of Black–Scholes and Cox Call Option Valuation Models." *Journal of Finance* 35, no. 2: 285–301.

Nelson, D. 1990. "ARCH Models as Diffusion Approximations." *Journal of Econometrics* 45, no. 1: 7–38.

Pan, J. 2002. "The Jump-Risk Premia Implicit in Options: Evidence From an Integrated Time-Series Study." *Journal of Financial Economics* 63, no. 1: 3–50.

Perelló, J., R. Sircar, and J. Masoliver. 2008. "Option Pricing Under Stochastic Volatility: The Exponential Ornstein–Uhlenbeck Model." *Journal of Statistical Mechanics: Theory and Experiment* 6: 1–22.

Platen, E., and N. Bruti-Liberati. 2010. *Numerical Solution of Stochastic Differential Equations With Jumps in Finance*. Springer.

Romano, J., M. Azeem, and M. Wolf. 2010. "Hypothesis Testing in Econometrics." *Annual Review of Economics* 2, no. 1: 75–104.

Romano, J., and M. Wolf. 2005a. "Exact and Approximate Step-Down Methods for Multiple Hypothesis Testing." *Journal of the American Statistical Association* 100, no. 469: 94–108.

Romano, J., and M. Wolf. 2005b. "Stepwise Multiple Testing as Formalized Data Snooping." *Econometrica* 73, no. 4: 1237–1282.

Romano, J., and M. Wolf. 2007. "Control of Generalized Error Rates in Multiple Testing." *Annals of Statistics* 35, no. 4: 1378–1408.

Romano, J., and M. Wolf. 2010. "Balanced Control of Generalized Error Rates." *Annals of Statistics* 38, no. 1: 598–633.

Scott, L. O. 1987. "Option Pricing When the Variance Changes Randomly: Theory, Estimation, and an Application." *Journal of Financial and Quantitative Analysis* 22: 419–438.

Scott, L. O. 1997. "Pricing Stock Options in a Jump-Diffusion Model With Stochastic Volatility and Interest Rates: Applications of Fourier Inversion Methods." *Mathematical Finance* 7, no. 4: 413–426.

Shephard, N. 1991a. "From Characteristic Function to Distribution Function: A Simple Framework for the Theory." *Econometric Theory* 7, no. 4: 519–529.

Shephard, N. 1991b. "Numerical Integration Rules for Multivariate Inversions." *Journal of Statistical Computation and Simulation* 39: 37–46.

Tavella, D., and C. Randall. 2000. *Pricing Financial Instruments: The Finite Difference Method*. Wiley.

U.S. Options Price Reporting Authority. 2015. *S&P 500 Index Options (SPX) Data. Proprietary Data Accessed Under License*. https://www.opraplan.com/.

West, M. 1993. "Approximating Posterior Distributions by Mixture." *Journal of the Royal Statistical Society. Series B (Methodological)* 55, no. 2: 409–422.

White, H. 2000. "A Reality Check for Data Snooping." *Econometrica* 68, no. 5: 1097–1126.

Wiggins, J. 1987. "Option Values Under Stochastic Volatility: Theory and Empirical Estimates." *Journal of Financial Economics* 19: 351–372.

Yang, H., and J. Kanniainen. 2017. "Jump and Volatility Dynamics for the S&P 500: Evidence for Infinite-Activity Jumps With Non-Affine Volatility Dynamics From Stock and Option Markets." *Review of Finance* 21, no. 2: 811–844.

Yun, J. 2011. "The Role of Time-Varying Jump Risk Premia in Pricing Stock Index Options." *Journal of Empirical Finance* 18, no. 5: 833–846.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.