HEC MONTRÉAL

École affliée à l'Université de Montréal

Three essays on volatility and extreme events in financial and electricity markets

par

Rémi Galarneau-Vincent

Thèse présentée en vue de l'obtention du grade de Ph. D.
en administration (Spécialisation Ingénierie Financière)

Décembre 2022

HEC MONTRÉAL

École affliée à l'Université de Montréal

Cette thèse intitulée:

Three essays on volatility and extreme events in financial and electricity markets

Présentée par:

Rémi Galarneau-Vincent

a été évaluée par un jury composé des personnes suivantes :

David Ardia
HEC Montréal
Président-rapporteur

Geneviève Gauthier
HEC Montréal
Directrice de recherche

Pascal François
HEC Montréal
Membre

Lars Stentoft
Western University
Membre externe du comité conjoint

Michel Denault
HEC Montréal
Représentant du directeur de HEC Montréal

Résumé

Cette thèse se concentre sur des applications reliées à la volatilité et aux évènements extrêmes dans les marchés des produits dérivés et de l'électricité.

Le premier essai propose une nouvelle représentation factorielle de la surface de volatilité implicite dont le titre sous-jacent est le S&P 500. Les cinq facteurs proposés capturent adéquatement le niveau, la pente de degré d'exercice et de l'échéance, de l'atténuation du «smile» ainsi que du «smirk». De plus, leur comportement asymptotique permet d'effectuer une extrapolation de la surface bien au-delà des échéances et des degrés d'exercice observés. Pour chaque échéance fixée, notre modèle de volatilité implicite garantit l'existence d'une fonction de densité risque neutre pour le prix de l'actif sous-jacent. La performance du modèle ajusté sur les options du S&P 500 se compare favorablement aux méthodes de références existantes. Les avantages d'une surface de volatilité implicite lissée sont illustrés lors de l'évaluation de dérivés d'indices non liquides, de l'extraction de la densité neutre au risque et des moments neutres au risque ainsi que du calcul des sensibilités du prix des options. Cet article est paru dans *Journal of Futures Markets, 2022, 42(10), 1912–1940*.

Dans le second essai, nous modélisons la dynamique conjointe de l'indice S&P 500 et de sa surface de volatilité implicite. En effet, l'ensemble des caractéristiques de la surface de volatilité actuelle est pris en compte pour modéliser les déformations futures. Deux exercices sont effectués pour démontrer la capacité du modèle «joint implied volatility and return» (JIVR) à générer avec précision des scénarios pour la surface de volatilité implicite future conjointement avec le rendement du sous-jacent. Le premier exercice consiste à évaluer le risque des positions «straddle» et «strangle». Le deuxième exercice analyse la

performance prédictive du modèle JIVR à prédire la distribution de l'indice VIX. Le modèle JIVR s'avère efficace pour une gestion du risque liée aux options de l'indice S&P 500.

Le dernier essai se concentre sur la prévision des pics de prix d'électricité DART observé dans la zone de Long Island du NYISO. L'écart DART est d'une grande importance économique pour les négociants d'énergie qui s'y exposent. Un ensemble de variables comprenant des caractéristiques prospectives, rétrospectives et saisonnières est proposé. Quatre algorithmes d'apprentissage automatiques sont utilisés : la régression logistique, la forêt aléatoire, l'arbre de «gradient boosting» et les réseaux de neurones artificiels. Les mesures de performance statistiques attestent que tous les modèles présentent un pouvoir de prédiction, tant sur l'échantillon d'entraînement que sur l'échantillon test. Un exercice d'évaluation des variables illustre la valeur ajoutée de plusieurs variables construites. Les avantages de la prévision des pics de prix sont illustrés par un exercice de négociation. Cet article est paru dans *Energy Economics, 2023, p. 106521*.

Mots clés: Volatilité implicite, Modèle factoriel, Grecs, Gestion des risques, VIX, Marché de l'électricité, Écarts DART, NYISO, Analyse prédictive, Apprentissage statistique.

Méthodes de recherche: Économétrie, Analyse multivariée, Apprentissage automatique.

Abstract

This thesis concentrates on applications related to the volatility and extreme events present in the derivatives and electricity markets.

The first essay proposes a new factorial representation of the implied volatility surface with the underlying security being the S&P 500. The five proposed factors adequately capture the level, the moneyness and maturity slopes the smile attenuation, and the smirk. In addition, their asymptotic behaviour allows for an extrapolation of the surface well beyond the range of observed maturities and moneyness. For each fixed maturity, our implied volatility model guarantees the existence of a risk-neutral density function for the underlying asset price. The performance of the adjusted model on the S&P 500 options compares favourably with existing benchmarks. The benefits of a smoothed implied volatility surface are illustrated through the valuation of illiquid index derivatives, the extraction of the risk-neutral density and risk-neutral moments, and the calculation of option price sensitivities. This article has been published in *Journal of Futures Markets, 2022, 42(10), 1912–1940.*

In the second essay, we model the joint dynamics of the S&P 500 index and of its associated implied volatility surface. Indeed, all the characteristics of the current volatility surface are taken into account to model future deformations. Two exercises are conducted to demonstrate the ability of the joint implied volatility and return (JIVR) model to accurately generate scenarios for the future implied volatility surface jointly with the return of the underlying asset. The first exercise consists of evaluating the risk of straddle and strangle positions. The second exercise analyses the predictive performance of the JIVR model in predicting the distribution of the VIX index. The JIVR model proves to be effective for

managing the risk associated with S&P 500 index options.

The last essay concentrates on forecasting the electric day-ahead price minus the real-time price (DART) spikes observed in the Long Island zone of the NYISO. DART spread is of paramount importance to virtual bidders. A tailored feature set encompassing forward-looking, backward-looking, and seasonal features, including novel engineered features, is proposed. Four machine learning algorithms: logistic regression, random forest, gradient boosting tree, and artificial neural networks are trained. Statistical performance measures attest that all models exhibit prediction power, both in-sample and out-of-sample. A feature assessment exercise illustrates the value added of multiple engineered features. The benefits of forecasting day-ahead spikes are illustrated through a trading exercise. This article has been published in *Energy Economics, 2023, p. 106521*.

Keywords: Implied volatility, Factor models, Greeks, Risk management, VIX, Power markets, Spikes prediction, DART spreads, NYISO, Predictive analytics, Statistical learning.

Research methods: Econometrics, Multivariate analysis, Statistical learning.

Contents

List of Figures

List of Tables

Abbreviations

ALR	**A**verage **L**ikelihood **R**atio
ARMSE	**A**verage **R**oot **M**ean **S**quare **E**rror
ATM	**A**t-**T**he-**M**oney
AUC	**A**rea **U**nder the receiver Operating **C**urve
BIC	**B**ayesian **I**nformation **C**riterion
BMS	**B**lack-**M**erton-**S**choles
CBOE	**C**hicago **B**oard of **E**xchange
CDD	**C**ooling **D**egree **D**ay
CT	**C**halamandaris and **T**sekrekos
CTHI	**C**umulative **T**emperature and **H**umidity **I**ndex
DA	**D**ay-**A**head
DART	**D**ay-**A**head price minus the **R**eal-**T**ime Price
DNN	**D**eep **N**eural **N**etworks
DOTM	**D**eep-**O**ut-of-**T**he-**M**oney
GARCH	**G**eneralized **A**utoRegressive **C**onditional **H**eteroskedasticity
GG	**G**oncalves and **G**uidolin
HDD	**H**eating **D**egree **D**ay
ITM	**I**n-**T**he-**M**oney
IV	**I**mplied **V**olatility
IVRMSE	**I**mplied **V**olatility **R**oot **M**ean **S**quare **E**rror
JIVR	**J**oint **I**mplied **V**olatility and **R**eturns
LBMP	**L**ocational-**B**ased **M**arginal **P**ricing

NGARCH	**N**onlinear **A**symmetric **G**eneralized **A**utoRegressive **C**onditional **H**eteroskedasticity
NIG	**N**ormal-**I**nverse **G**aussian
NYISO	**N**ew **Y**ork **I**ndependent **S**ystem **O**perator
OLS	**O**rdinary **L**east **S**quare
OTM	**O**ut-of-**T**he-**M**oney
P&L	**P**rofit and **L**osses
PCA	**P**rincipal **C**omponent **A**nalysis
RH	**R**elative **H**umidity
RIX	**R**are **D**isaster **I**ndex
RMSE	**R**oot **M**ean **S**quare **E**rror
ROC	**R**eceiver **O**perating **C**urve
RT	**R**eal-**T**ime
SHAP	**SH**apley **A**dditive ex**P**lanations
SOPA	**S**tandard **O**ption **P**ricing **A**pproach
THI	**T**emperature and **H**umidity **I**ndex
VaR	**V**alue-**a**t-**R**isk
VOLVOL	**VOL**atility of the implied **VOL**atility

Remerciements

J'aimerais tout d'abord remercier ma superviseure de doctorat, Geneviève Gauthier, sans qui l'achèvement de mon doctorat n'aurait pas pu être possible. En plus de m'avoir constamment amené à approfondir mon raisonnement, elle m'a offert durant les quatre dernières années d'excellents conseils, un mentorat constant, de la bonne compagnie et sans oublier les nombreuses phrases culte qui auront marqué ses quatre dernières années.

Je voudrais exprimer mes sincères remerciements aux Professeurs Frédéric Godin et Pascal François pour leurs conseils, leur précieuse aide ainsi que d'excellentes séances discussion qui m'ont permis de cheminer dans mon raisonnement.

Je souhaite aussi remercier deux de mes collègues, Gabrielle Trudeau et Samuel Léveillé, avec qui j'ai pu avoir de très belles conversations sur la finance quantitative.

Finalement, je voudrais remercier les membres de ma famille qui ont été présent pour moi tout au long de mon parcours universitaire. Merci!

Chapter 1

Introduction

Risk management is a core priority of financial institutions. The survival and profitability of an institution are linked to the sound management and quantification of its portfolios. Events such as the financial crisis of 2008 or the COVID-19 pandemic remind us of the importance of a sound and strong risk management system to ensure that financial institutions remain solvent during distressing periods. After the catastrophic financial crisis of 2008, regulators accelerated the implementation of stricter regulations on financial institutions. The regulations particularly concentrate on better quantifying the risks of the assets such that institutions can better estimate the amount of capital to hold to meet their obligations even during the worst economic downturns. This thesis share has a common theme the quantification, management and assessment of risk over two markets: the derivatives market and the electricity market.

The first two essays concentrate on the derivatives market. Derivatives are financial contracts whose payoff is linked to an underlying asset. Derivatives are often used by financial institutions as insurance contracts to reduce or limit the risk associated with some of their positions. Even though derivatives are priced, market practitioners do not work with option prices directly, but with the option's implied volatility inferred from the Black and Scholes, 1973 formula. Implied volatility is a measure of the expected future volatility of an underlying asset based on the price of its options. It represents the market's perception of the underlying asset's volatility and future price movements. In contrast, option prices

alone do not provide clear information about the underlying asset's future price movements or volatility. By transforming option prices into implied volatility, market participants can form views on options more easily. They can compare the implied volatility of different options with varying moneyness and maturities to gain insights into the market's expectations about future price movements. Despite the limitations of the Black and Scholes, 1973 model, which is used to calculate implied volatility, it remains a widely accepted way to calculate implied volatility and a crucial tool for options pricing and risk management. The implied volatility surface obtained by combining the implied volatility of all quoted options embeds a rich source of information about market participants' forward-looking view of the market dynamics. Leveraging this information to manage and estimate the risk of derivatives is a core objective of the first two essays. The first essay takes a static perspective, by proposing a factor model to complete the implied volatility surface. A completed implied volatility surface generates multiple applications related to risk management and the mark-to-market of illiquid or complex derivatives that are not quoted. The second essay concentrates on the dynamics of the implied volatility surface to quantify the risk associated with these derivatives.

The first essay develops a model to complete the implied volatility surface since the implied volatility surface is incomplete, i.e. only a few observations with limited moneyness and maturity ranges are observed. Completing the implied volatility surface is of great importance for financial institutions that trade illiquid or complex over-the-counter (OTC) derivatives and have to manage their counterparty risk. A completed implied volatility surface can be used to mark-to-market these derivatives throughout their lives, without the need to observe a quoted derivative with the same characteristics. The implied volatility surface shape is captured by five economically interpretable factors. Each factor is selected to capture a particular component of the implied volatility surface. When choosing the functional factors, special attention is paid to ensure that the fitted surface is smooth, twice-differentiable and well-behaved asymptotically. These properties are necessary for the extraction of the risk-neutral density and the limitation of arbitrage opportunities generated by the model. The resulting 5-factor model is able to interpolate and extrapolate the implied volatility surface while being coherent with the observable surface. Complet-

ing the implied volatility surface stems many applications related to risk management and the extraction of market information from the implied volatility surface. In terms of risk management, the 5-factor model can mark-to-market illiquid derivatives. The interpolation and extrapolation capabilities of the model are paramount for extracting the complete risk-neutral density, or when pricing derivatives from a replicating option portfolio based on the Carr and Madan, 2001 that heavily relies on often scarce DOTM options. The 5-factor model can also compute option's Greeks by taking into account the shape of the implied volatility surface, which is useful for hedging illiquid options.

The second essay builds on the first by introducing a joint model that captures the dynamics of the implied volatility surface and the underlying asset's log-returns (JIVR). A model that captures these joint dynamics is able to forecast the distribution of a wide range of derivatives. Therefore, the JIVR model is a powerful tool to effectively quantify the risk of a derivative. Unlike the traditional approach, the JIVR model leverages information from the implied volatility surface to forecast returns over short-term horizons of 1 and 5 days. Integrating the implied volatility surface as an input to the JIVR model allows for a novel characterization of the S&P 500 log-returns. The JIVR model is used to jointly simulate the forecasted distributions of implied volatility and log-returns. The derivative return distribution can, thus, be computed from the forecasted distribution of the implied volatility and the log-returns. This application can be paramount to large investors required to compute the VaR of their positions over the standard 1-day and 5-day horizon.

Finally, the last chapter focuses on the electricity market while keeping risk management as a central objective of the research paper. In this chapter, the view of virtual bidders taking positions in the electricity market of the Long Island zone of New York State is adopted. The objective of virtual bidders is to exploit the inefficiencies present in the electricity market. One of their main strategies consists in selling or buying electricity on the day-ahead market and reversing their position on the real-time market. However, electricity prices are well known for their extreme volatility due to the fact that consumption must equate with generation continuously. Therefore, the positions taken by virtual bidders are exposed to large downside events known as price spikes. To reduce the risk of their positions and improve the risk-reward profile of their trading strategies, we use

statistical learning approaches to forecast the likelihood of such extreme events based on a tailored feature set. Trading strategies that integrate the models' forecasts are shown to be substantially less risky and far more profitable than base case approaches.

Chapter 2

Venturing into Uncharted Territory: An Extensible Parametric Implied Volatility Surface Model

Abstract*

A new factor-based representation of implied volatility surfaces is proposed. The factors adequately capture the moneyness and maturity slopes, the smile attenuation, and the smirk. Furthermore, the implied volatility specification is twice continuously differentiable and well behaved asymptotically, allowing for clean interpolation and extrapolation over a wide range of moneyness and maturity. Fitting performance on S&P 500 options compares favorably with existing benchmarks. The benefits of a smoothed implied volatility surface are illustrated through the valuation of illiquid index derivatives, the extraction of the risk-neutral density and risk-neutral moments, and the calculation of option price sensitivities.

Keywords: Implied volatility surfaces, Derivatives pricing, Factor models, Greeks.

*Joint work with Pascal François, Geneviève Gauthier, and Frédéric Godin. Fraçois and Gauthier are affiliated with HEC Montréal and Godin is affiliated with Concordia University.

2.1 Introduction

Since their introduction on exchanges, derivatives have become a central part in modern asset pricing theory. Beyond their fundamental role as risk management tools, options are contingent claims whose market prices convey all the information needed to determine state prices (Cox and Ross, 1976). In conjunction with the information embedded in the underlying asset returns, these state prices can further be disentangled into physical probabilities and the pricing kernel (Jackwerth, 2000, Ross, 2015). Breeden and Litzenberger, 1978 show that the risk-neutral density of the underlying asset price can be retrieved from the continuum of options across strikes. From this seminal property, a rich set of extended results has emerged in the financial economics literature, including but not limited to: the spanning of a terminal payoff with a portfolio of discount bonds and options (Bakshi and Madan, 2000, Carr and Madan, 2001), the inference of risk-neutral moments (Bakshi and Kapadia, 2003, Conrad et al., 2013, Neumann and Skiadopoulos, 2013, Ammann and Feser, 2019), the construction of the risk-neutral density (Birru and Figlewski, 2012, Figlewski, 2018), the static hedging of options (Carr and Wu, 2014), and the calculation of risk metrics such as the VIX index (Neuberger, 1994), the SVIX ((Martin, 2017; Martin & Wagner, 2019)), or the rare disaster index (RIX) (Gao et al., 2018, Gao et al., 2019).

The implementation of these applications is hampered, however, by the limited availability of traded options. Given the one-to-one correspondence between the European option premium and the implied volatility (IV) established by Black and Scholes, 1973, the continuum of options across strikes and maturities can be represented by the IV surface. In practice, observable IVs form a cloud of points only, and the construction of a smoothed IV surface is an empirical challenge for academics and a mandatory daily exercise for industry practitioners who trade options. Observed implied volatilities are particularly scarce far from the money, which is precisely where information about the tails of the risk-neutral distribution can be retrieved. They are also scarce for medium and long maturities. Thus, an accurate extrapolation of the IV surface can help better estimate the high-order risk-neutral moments and how they aggregate over the time horizon.

This paper introduces a new functional form for the IV with three major benefits. First, it is designed to accommodate the well-documented, stylized features of the IV surface

for S&P 500 index options. Chalamandaris and Tsekrekos, 2011 (hereafter CT) explicitly introduce the level, slope and curvature factors, in the spirit of how Nelson and Siegel, 1987 and Diebold and Li, 2006 model the yield curve. The IV specification presented in this paper takes a step further by assigning precise roles to the factors. Specifically, the factors are explicitly designed to capture the slopes in the moneyness and the maturity dimensions, the smile attenuation, (i.e., the fact that the smile convexity decreases with the maturity), and the smirk (i.e., the fact that the IVs for short-term, deep out-of-the-money calls are higher than those at-the-money). In particular, the specific account of the slope attenuation and the smirk greatly enhances the calibration performance on the SPX option IV surface over the January 1996 – December 2019 period compared to the Heston, 1993, the Goncalves and Guidolin, 2006 (hereafter GG), and the Chalamandaris and Tsekrekos, 2011 benchmarks.

Second, the factors are tailored to admit a stable asymptotic behaviour, which makes it possible to extrapolate beyond quoted moneyness levels and maturities. This feature proves to be particularly valuable when it comes to valuing long-term, illiquid, index-option-related contracts (such as structured notes). Third, the factors are constructed so that the surface is twice continuously differentiable, which produces a well-behaved risk-neutral density function for each time horizon. These last two features represent a significant improvement over factor-based benchmarks as extrapolating the IV surface has remained an empirical challenge in the literature.[2]

Several smoothing methods have been proposed in the literature. They can be broadly classified into parametric and non-parametric approaches.[3]

Parametric smoothing allows seamless interpolation, parsimony, interpretability, and limited computational requirements for option pricing. However, the parametric specifications proposed in earlier works are chosen for their simplicity and convenience. Dumas

[2] Our numerical experiments show that CT and GG models often induce anomalous risk-neutral densities when our method does not.

[3] Among non-parametric smoothing methods are the Gaussian kernel (Cont and Da Fonseca, 2002), the principal component analysis (Israelov and Kelly, 2017), and the neural networks (Ackerer et al., 2019). One major limitation of non-parametric approaches is that the extrapolation of the smoothed IV surface is a non-trivial issue.

et al., 1998 examine the IV as a polynomial function of strike and maturity. Goncalves and Guidolin, 2006 apply a similar specification to the log-IV, replacing the strike with a moneyness factor. Jackwerth and Rubinstein, 1996 and Bliss and Panigirtzoglou, 2002 use cubic splines.

The paper adds to three strands of the literature. First, it is related to the mark-to-market of thinly traded securities (Skiadopoulos, 2001, Henderson and Pearson, 2011, Célérier and Vallée, 2017). A wide variety of derivatives are not publicly traded on exchanges, e.g. over-the-counter derivatives, structured products, and options embedded in corporate securities. Nevertheless, they can be priced in a consistent manner when the valuation models are anchored to the IV surface of the vanilla options with same underlying asset (Daglish et al., 2007, Bayraktar and Yang, 2011). The pricing consistency requires an arbitrage-free smoothed IV surface (Fengler, 2009). The IV surface specification presented in this paper successfully passes the arbitrage detection assessment using butterfly and calendar spreads (Davis and Hobson, 2007) which represents sufficient conditions for the absence of static arbitrage. A numerical application shows how a complete, smoothed IV surface streamlines the mark-to-market of an equity-index-linked note, making it less reliant on the entry and exit of available strikes and maturities. For the seller of the note, this, in turn, reduces unnecessary liquidity stress on the management of the position.

Second, the paper contributes to the extraction of the risk-neutral density and risk-neutral moments. As argued by Jackwerth, 2004, the challenge in constructing the risk-neutral density does not reside in the centre of the distribution but in the tails, where few options are observable. Parametric smoothing methods, therefore, complete the shape of the density either with composite distributions or with mixture models (Figlewski, 2018). The IV specification of this paper circumvents this issue by using factors that are twice continuously differentiable and well-behaved asymptotically. The generated risk-neutral densities are shown to be smooth and regular, and the mass of probability over the full spectrum of moneyness levels adds up to exactly one. Furthermore, numerical experiments show the substantial correction in the computation of risk-neutral skewness and kurtosis when the Carr and Madan, 2001 formula is discretized.

Third, the specification for the implied volatility surface proposed in this paper can

be applied to the dynamic hedging of options. Delta and gamma are derived analytically and are shown to be smile-implied Greeks (Bates, 2005, Alexander and Nogueira, 2007, François and Stentoft, 2021). That is, these Greeks are consistent with the observed shape of the volatility smile and they do not depend on any assumption regarding the underlying asset dynamics. Furthermore, the specification for the implied volatility surface allows for the enhanced management of volatility risk, not only through the traditional definition of the option vega but also through the sensitivity of the option value with respect to the long-term volatility level and to the IV maturity slope.

The paper is structured as follows. Section 2.2 presents the data. Section 2.3 introduces the IV specification, compares its fitting performance with selected benchmarks, and checks for the presence of arbitrage opportunities in the smoothed IV surface. Section 2.4 details the applications of the IV surface model to derivatives pricing and risk management. Section 2.5 concludes.

2.2 Data

The dataset extracted from the OptionMetrics database consists of European call and put options on the S&P 500 index (SPX options) quoted daily on the CBOE from January 4, 1996, to December 31, 2019.[4] For each option quote, the data includes bid and ask prices, from which mid-prices are calculated to serve as option prices. The data also includes forward prices of the S&P 500 index associated with the same maturity date for each option. For each quote, the associated option moneyness is defined as

$$M = \frac{1}{\sqrt{\tau}} \log\left(\frac{F_{t,\tau}}{K}\right), \tag{2.1}$$

where τ is the annualized time-to-maturity[5] of the option, $F_{t,\tau}$ is the day-t forward price of maturity τ given by OptionMetrics, and K is the strike price. Thus, $M = 0$ corresponds to at-the-money (ATM) options, $M < 0$ to out-of-the-money (OTM) calls and $M > 0$ to OTM puts. As time-to-maturity increases, the range of traded strike prices also widens due

[4]Only spot options are considered, i.e., options clearing on the spot price of the S&P 500 index.

[5]For clarity of exposition, τ is reported in days in tables and figures.

to the scaling property of volatility. Scaling the moneyness by $\frac{1}{\sqrt{\tau}}$ generates a comparable moneyness measure across time-to-maturity.

Data exclusion procedures are in line with the filters of Bakshi et al., 1997. Options with the following characteristics are removed: (i) a time-to-maturity shorter than 6 days, (ii) a price lower than $3/8\$$, (iii) a zero bid price, and (iv) options with a bid-ask spread larger than 175% of the option's mid-price.[6] Furthermore, all in-the-money options are excluded (i.e., puts with $M < 0$ and calls with $M \geq 0$). The resulting sample contains 3,468,515 option quotes spanning on 6,039 days.[7]

	Calls		Puts			
	$M \leq -0.2$	$-0.2 < M \leq 0$	$0 < M \leq 0.2$	$0.2 < M \leq 0.8$	$M \geq 0.8$	All
Average IV(%)	17.90	15.23	19.95	28.43	44.12	23.94
Standard deviation IV(%)	7.19	5.46	5.53	6.40	9.60	10.30
Number of contracts	274,252	774,321	772,735	1,303,428	279,447	3,404,183
	$\tau \leq 30$	$30 < \tau \leq 90$	$90 < \tau \leq 180$	$180 < \tau \leq 365$	$\tau \geq 365$	All
Average IV(%)	24.68	23.93	23.88	24.18	23.55	23.94
Standard deviation IV(%)	13.08	11.27	10.32	9.50	8.42	10.30
Number of contracts	277,521	1,008,508	610,650	703,008	804,496	3,404,183

Descriptive statistics of the SPX options implied volatility (IV) daily data from January 4, 1996, to June 26, 2019, across multiple times-to-maturity and moneyness buckets. M represents the moneyness defined in Equation (2.1) and τ is the time-to-maturity of the option in days.

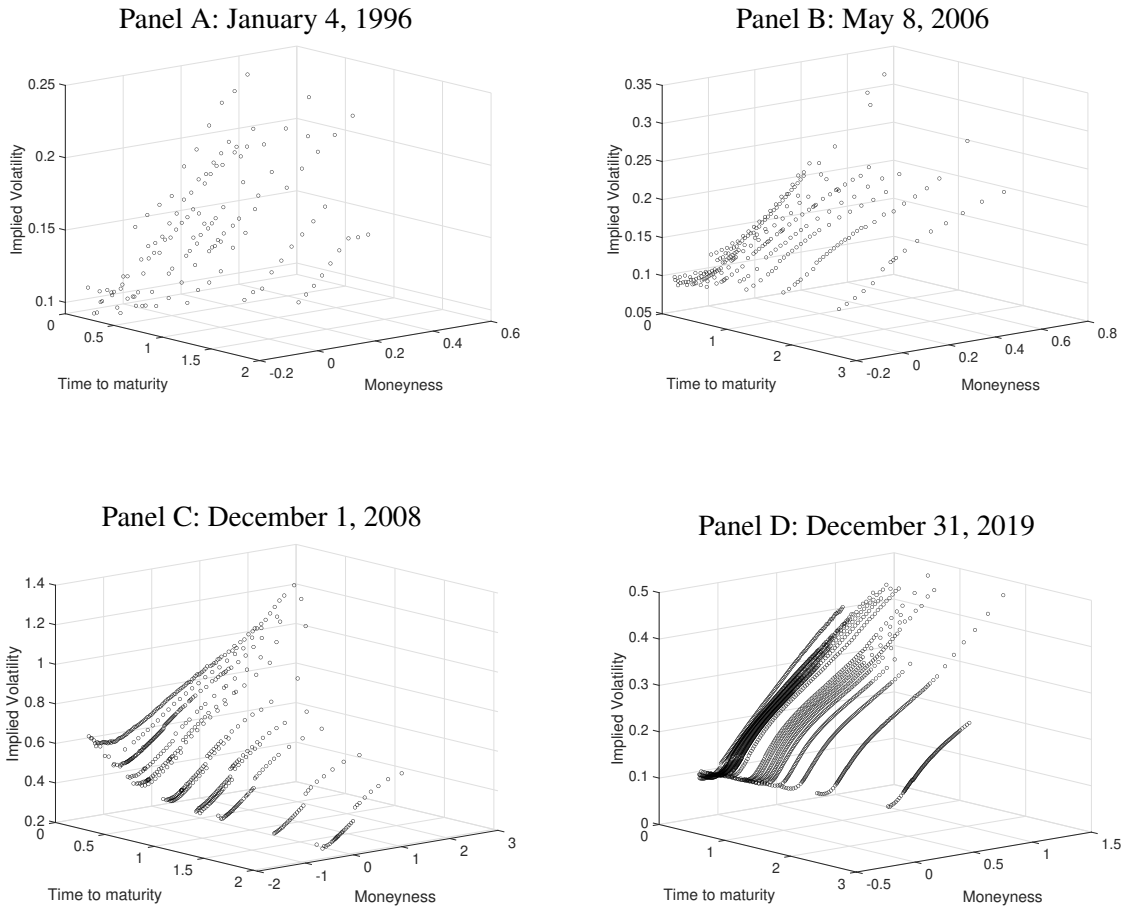
Table 2.1: Descriptive statistics of the SPX options data

For each option in the sample, its implied volatility is computed with the Black, 1976 formula. Table 2.1 presents descriptive statistics about option implied volatilities (hereafter IV) associated with all quotes from the dataset. The IV surfaces from the various days covered by the data typically display asymmetry and often the well-known smirk (i.e., calls with a moneyness $M < -0.2$ having higher IV than calls with a moneyness $-0.2 < M \leq$

[6] This filter is inspired by Azzone and Baviera, 2022. Options with a large ratio of bid-ask spread over price yield inaccurate mid-prices, which then induce anomalies in the IV surfaces and artificial spikes in the factors' time series. Options excluded by this filter represent a minuscule proportion of the sample (0.3%).

[7] The October 9, 2006, IV surface is removed from the dataset since the IV surface displays a broken shape which most likely indicates that the data is erroneous for that specific day.

0). Moreover, as seen in Table 2.1, there are slightly fewer options with short maturities (less than or equal to 90 days) than there are options with long maturities (more than 90 days) options.



The various panels display the set of implied volatilities associated with retained option quotes from the OptionMetrics dataset on four selected days. January 4, 1996, is the first day in the sample. May 8, 2006, is a low volatility day. December 1, 2008, represents the peak of the 2008 financial crisis. December 31, 2019, is the last day of the sample. The moneyness is defined in Equation (2.1).

Figure 2.1: Observed IV surfaces on four selected dates

Figure 2.1 shows the set of all option IV considered on four selected days. These days are representative of stylized features of the IV surface. As observed when comparing the

first day of the sample (Panel A) to others, the number of quoted options has significantly gone up with time. The increased number of quoted options mainly results in much lower strike intervals between the quotes of a given maturity, but also in additional traded maturities. Panel B clearly displays the smile attenuation, where the moneyness slope flattens as the time-to-maturity increases. The IV surface from Panel C observed during the subprime crisis exhibits a time-to-maturity slope, i.e., a downward trend of the IV with respect to the option's time-to-maturity. The volatility smirk is particularly visible in Panel D, OTM calls having a higher IV than ATM options. Such characteristics of IV surfaces are well-known in the literature and are mentioned, for instance, in Cont and Da Fonseca, 2002 and Rebonato, 2005. Furthermore, all four panels highlight the positive impact of scaling the moneyness by $\frac{1}{\sqrt{\tau}}$ where the range of moneyness remains similar for all time-to-maturities and the implied volatility levels become comparable for options with similar moneyness across time-to-maturity.

2.3 Model specification and performance

The following three requirements are considered for the parametric function: (i) factor interpretability, (ii) twice differentiability, and (iii) extrapolation ability.

First, our factors are carefully selected functions of the strike and time-to-maturity. Their specification has been designed to match patterns commonly observed on IV surfaces such as the smile attenuation or the volatility smirk. Assigning a specific role to each factor allows for their clear interpretation. It also contributes to an accurate calibration to the observed IV surfaces, as shown below. Second, the functional forms for our factors are twice continuously differentiable, which ensures the existence of a continuous risk-neutral density function for the price of the underlying asset (Breeden and Litzenberger, 1978). This, in turn, limits the presence of arbitrage opportunities. Third, our factors are asymptotically stable and can therefore be easily extrapolated beyond the observed moneyness levels and maturities. As shown in Section 4, this property is helpful for various applications including option pricing, risk management and asset pricing.

In the next subsections, we introduce and discuss our implied volatility model. We then

report how the calibration of our specification strongly outperforms that of the parametric models of Goncalves and Guidolin, 2006 and Chalamandaris and Tsekrekos, 2011. We then proceed to a formal screening procedure detecting the presence of theoretical arbitrage opportunities on both the raw data and the set of model-calibrated IV surfaces.[8] For an agent trading at the bounds of the bid-ask spread, arbitrage opportunities turn out to be rare in the data, and even more so in the model-calibrated IV surfaces.

2.3.1 A parametric implied volatility specification

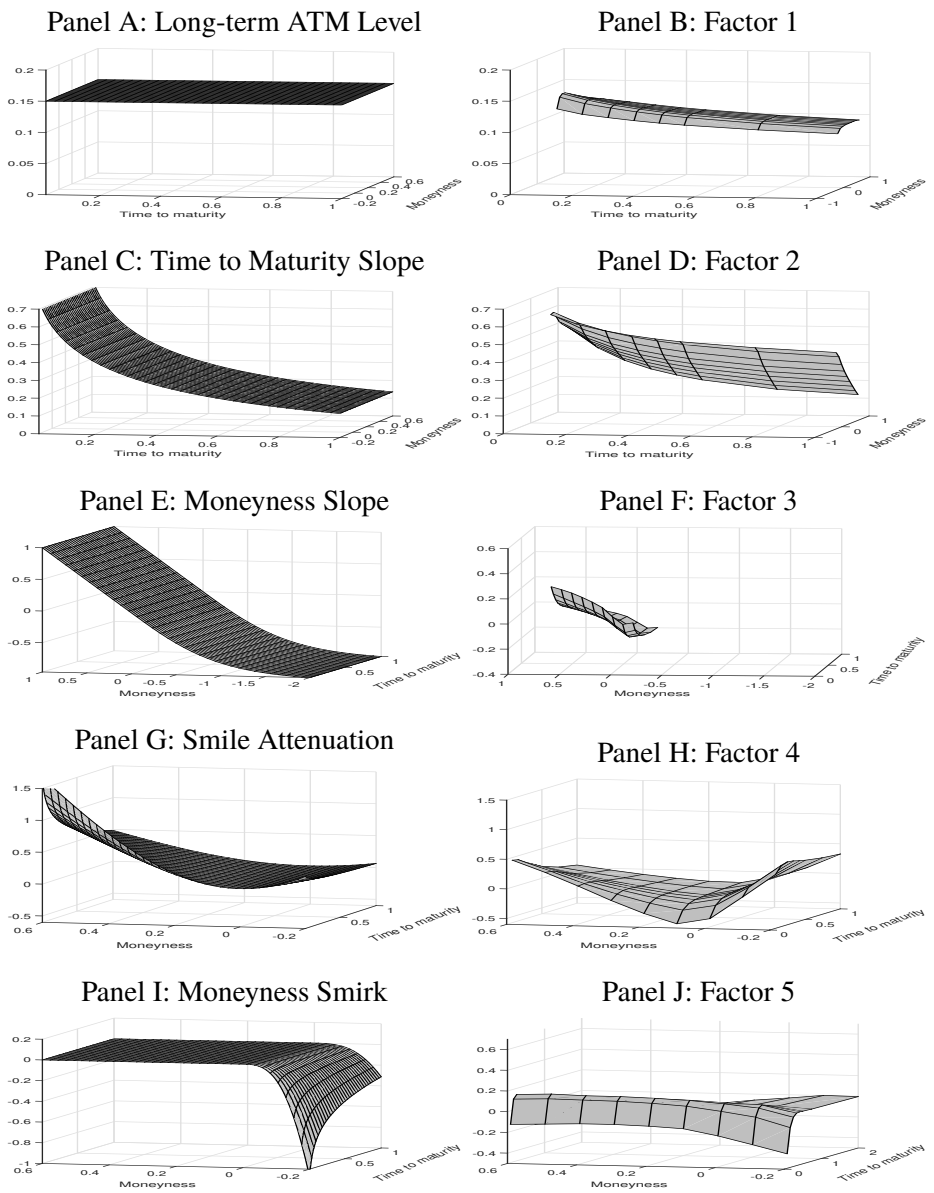
The implied volatility $\sigma(M,\tau)$ observed on a given day for an option with moneyness M defined in Equation (2.1) and time to maturity τ is modeled as

$$\begin{aligned}\sigma\left(M,\tau\right) = & \underbrace{\beta_1}_{\text{Long-term ATM Level}} + \beta_2 \underbrace{\exp\left(-\sqrt{\tau/T_{\text{conv}}}\right)}_{\text{Time to maturity slope}} + \beta_3 \underbrace{\left(M\mathbb{1}_{\{M\geq 0\}} + \frac{e^{2M}-1}{e^{2M}+1}\mathbb{1}_{\{M<0\}}\right)}_{\text{Moneyness slope}} \\ & + \beta_4 \underbrace{\left(1-\exp\left(-M^2\right)\right)\log(\tau/T_{\text{max}})}_{\text{Smile attenuation}} + \beta_5 \underbrace{\left(1-\exp\left((3M)^3\right)\right)\log(\tau/T_{\text{max}})\mathbb{1}_{\{M<0\}}}_{\text{Smirk}}.\end{aligned} \tag{2.2}$$

Fixed values T_{max} and T_{conv} are parameters selected based on empirical observation. T_{max} is the maximal maturity represented by the model. Although the longest time to maturity for options in the sample is three years, the value $T_{\text{max}} = 5$ years is considered herein to allow for extrapolation beyond the longest time to maturity. T_{conv} represents the location of a fast convexity change in the IV term structure with respect to time to maturity. It is set to $T_{\text{conv}} = 0.25$ for reasons explained below.

The five-factor representation in Equation (2.2) is parsimoniously designed to capture the stylized facts of the IV surface. The first factor is constant and its coefficient β_1 is a proxy for the long-term ATM implied volatility. Since $\lim_{\tau\to 0}\lim_{M\to 0}\sigma\left(M,\tau\right) = \beta_1 + \beta_2$, the coefficient β_2 measures the time to maturity slope of the ATM implied volatility. Given that the observed IV curvature over time to maturity is more pronounced for short-term options, the convexity correction is increased for horizons less than three months by relating β_2 with a non-linear function of $\tau/0.25$. The term $\tau/0.25$ creates substantially more convexity for the short-term (under 3 months) part of the IV surface. The third factor picks up the

[8] The method detects static arbitrage opportunities via calendar and butterfly spreads (Davis and Hobson, 2007, Fengler, 2009).



The five factors of Equation (2.2) are represented on the left panels. The right column panels present the first five factors obtained through a PCA applied to the options data as described in Appendix A.1. The proportions of variation explained by each factor are respectively 94.84%, 3.32%, 1.45%, 0.22%, and 0.16%. The top four panels are displayed using an angle highlighting the factor variation with respect to the time to maturity, while the remaining six bottom panels show the variation with respect to the moneyness. Panels E and F have a larger moneyness axis than the other panels to show the entire shape of the third factor.

Figure 2.2: Model factors versus PCA factors

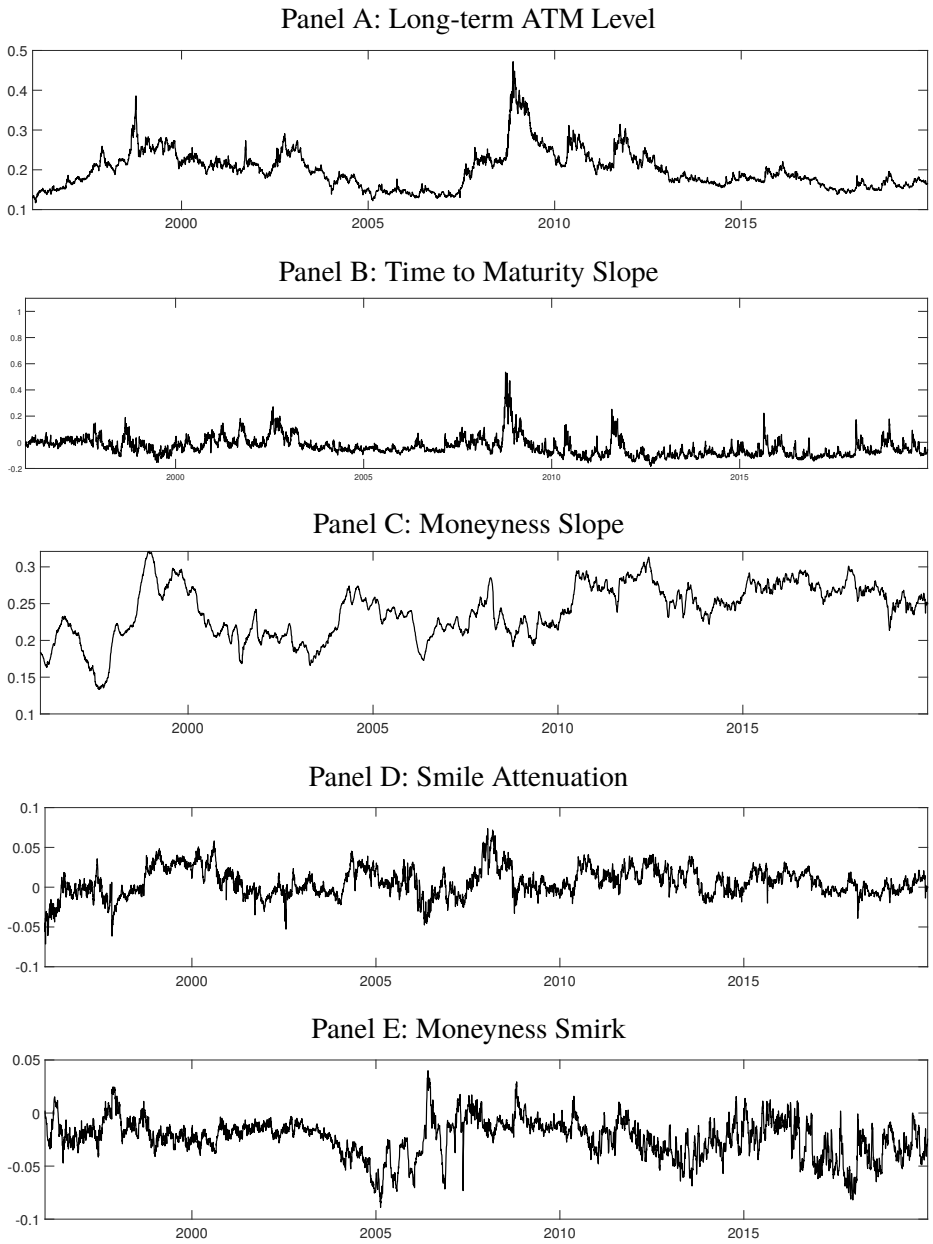
moneyness slope separately for put and call options. In the data, put IV increases almost linearly as the moneyness increases. Conversely, while call IV decreases at first when the moneyness decreases, it then either stabilizes or starts increasing for deeper OTM options as shown in Panels B, C and D of Figure 2.1. The hyperbolic tangent function in the third term generates a shape reminiscent of the blade of a hockey stick. It is strictly decreasing with a negative second derivative equal to 0 at $M = 0$ creating this desirable shape for call IV and generating a smooth moneyness slope function. The fourth factor accounts for the smile attenuation, i.e., it ensures that the smile gets flatter as time to maturity increases. Finally, the fifth factor captures the tilt in the smile for deep OTM calls, referred to as the implied volatility smirk. The smirk factor is designed to fade away as time to maturity increases toward the $T_{\max}$ bound.

Figure 2.2 compares the five factors of Equation (2.2) with those extracted from a Principal Component Analysis (PCA) decomposition of IV surfaces.[9] The use of PCA for the representation of volatility surfaces is motivated by the seminal paper of Cont and Da Fonseca, 2002 proposing a Karhunen-Loève decomposition of the log-surface dynamics generating orthogonal factors. Directly applying PCA to the option sample is not possible since option numbers and characteristics (moneyness and time to maturity) vary from day to day. Because PCA requires a stable sample every day, a grid with respect to moneyness and time to maturity is constructed over the densely populated regions of the surface. When applied to our dataset, the PCA is fitted to a sub-surface inside which only 41% of observed options lie. There are similarities between our five factors (left panels) and the first five PCA factors (right panels). Given that the PCA generates by construction the linear factors fitting the IV surface best in terms of RMSE, such similarity entails that Model (2.2) has the potential to adequately capture the IV surface patterns without requiring the construction of a subgrid of moneyness/time to maturity and discarding an important part of the observations as in the PCA approach.

2.3.2 Daily calibration

As described in Appendix A.2, a daily re-estimation of the set of parameters

[9] The PCA approach is described in Appendix A.1.



Daily evolution of estimates of parameters $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ of Model (2.2), respectively capturing the long-term ATM level, the time to maturity slope, the moneyness slope, the smile attenuation and the moneyness smirk, from January 4, 1996, to June 26, 2019. The estimates are obtained by minimizing the sum of the squared fitting errors while integrating Bayesian information (Appendix A.2).

Figure 2.3: Daily parameter estimates for the IV surface model

$\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ is obtained by minimizing the sum of the squared fitting errors while integrating Bayesian information for regularization purposes. Figure 2.3 plots the time series of estimated parameters. The level of the ATM implied volatility (panel A) remains low during calm market conditions. It spikes during financial turmoil, in particular around the LTCM crisis in September 1998 and, most notably, after the Lehman Brothers collapse by the end of 2008. The time to maturity slope of the IV surface (panel B) is usually slightly positive (i.e., negative β_2), but the 2008 financial crisis is associated with a strong and short-lived slope inversion. The smile asymmetry appears to be more stable in the second half of the sample with a less volatile coefficient β_3 (panel C). The smile attenuation β_4 varies more at the beginning of the sample and during the financial crisis (panel D). Finally, the smirk effect seems more pronounced but also more volatile over the most recent period (panel E).

2.3.3 Benchmarking

The calibration performance of Model (2.2) is compared with that of two benchmarks: the polynomial model of Goncalves and Guidolin, 2006 and the parametric model of Chalamandaris and Tsekrekos, 2011.

Goncalves and Guidolin, 2006, hereafter GG, regress the daily log IV surface on five factors:

$$\log \sigma\left(\tilde{M}, \tau\right) = \delta_1 + \delta_2 \tilde{M} + \delta_3 \tilde{M}^2 + \delta_4 \tau + \delta_5 (\tilde{M}\tau),$$

where their moneyness is defined as $\tilde{M} = \frac{1}{\sqrt{\tau}} \ln\left[\frac{K}{S\exp(r\tau)}\right]$, with S being the underlying asset price and r being the annualized risk-free rate. Working with log-volatilities ensures a positive IV surface. The various factors clearly represent moneyness and maturity slopes along with their interaction, on top of a convexity factor for the moneyness.

Chalamandaris and Tsekrekos, 2011, hereafter CT, fit the daily IV surface of foreign exchange options using seven factors:

$$\begin{aligned} \sigma(m, \tau) =& \theta_1 + \theta_2 1_{m>0} m^2 + \theta_3 1_{m<0} m^2 + \theta_4 \frac{1 - e^{-\lambda\tau}}{\lambda\tau} \\ &+ \theta_5 \left(\frac{1 - e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}\right) + \theta_6 1_{m>0} m\tau + \theta_7 1_{m<0} m\tau, \end{aligned} \tag{2.3}$$

where $m = (\Delta - 0.5) \times 100$ and Δ represents the Black-Scholes option delta.[10]

The first term of Equation (2.3) is the level of the surface. The second and third terms account for the right and left smiles. The curvature of the IV surface at the short and at the medium maturities is picked up by the fourth and fifth terms, respectively. The sixth and seventh terms capture the right and left smile attenuation.

2.3.4 Calibration performance

Figure 2.4 illustrates how Model (2.2) and the CT and GG benchmarks fit the IV data points for the four selected dates of Figure 2.1. The factor design is essential to obtain reliable surfaces across time.

Inspection of Figure 2.4 confirms the well-behaved extrapolation of the IV surface using Model (2.2). In particular, the shape of the IV surface remains consistent with the few implied volatilities observed in extreme regions of time to maturity and moneyness. By contrast, extrapolation of the IV surface using the GG benchmark induces a twist in the maturity slope that does not fit with the data. Likewise, extrapolation of the IV surface using the CT benchmark often induces a cap on far from the money-implied volatilities, which, again, is not in line with market observations.

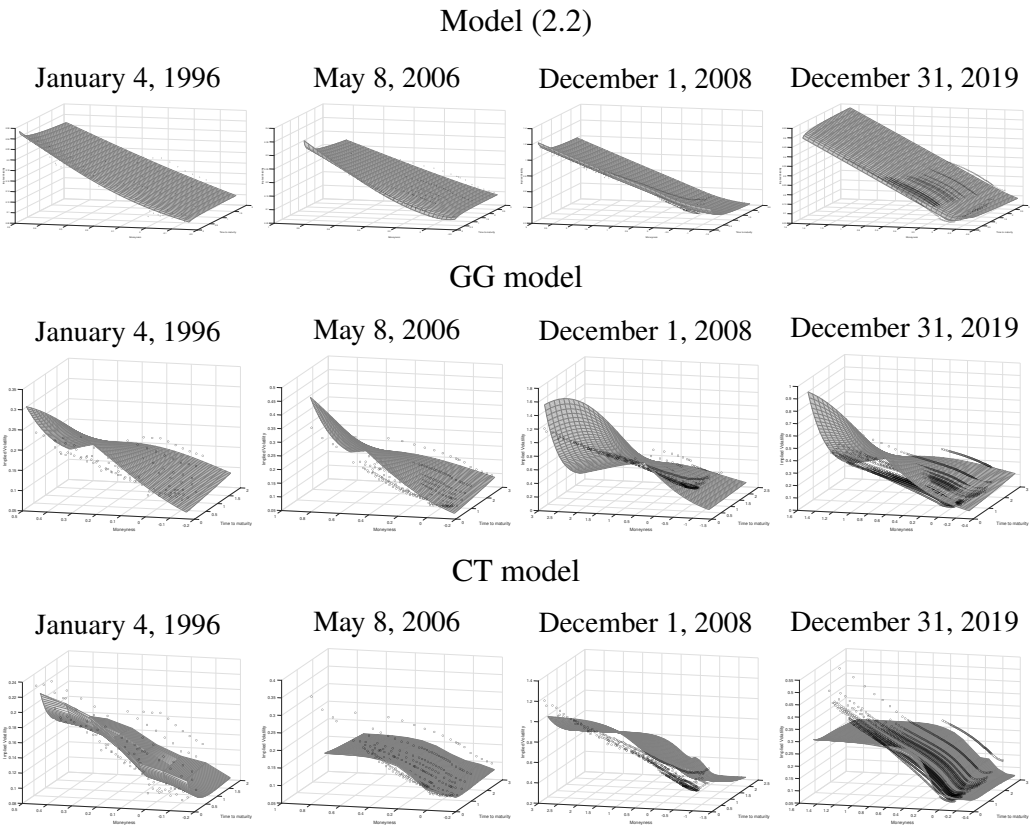
The calibration performance of each model over the entire sample is assessed by computing the daily root mean square error (RMSE) between the observed IV and the corresponding fitted values, that is,

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\sigma(M_i, \tau_i) - \sigma_i^O\right)^2},$$

[10]The CT study uses the Garman and Kohlhagen, 1983 delta defined in the context of the foreign exchange market. The call option Δ is computed as follows:

$$\begin{aligned} \Delta &= \frac{\partial c(K,\tau)}{\partial S} = e^{(d-r)\tau}\phi(d_1(K)) \\ d_1(K) &= \frac{1}{\sigma_{K,\tau}^{obs}\sqrt{\tau}}\log\frac{F_{0,\tau}}{K} + \sigma_{K,\tau}^{obs}\sqrt{\tau}. \end{aligned}$$

where d is the foreign risk-free rate. The $\sigma_{K,\tau}^{obs}$ used when fitting the model to the observed IV surfaces corresponds to the observed option IV.



Observed IVs are plotted against the fitted IV surfaces derived from Model (2.2) –first row– and the two benchmark models in rows 2 and 3. GG refers to Goncalves and Guidolin, 2006 and CT stands for Chalamandaris and Tsekrekos, 2011. January 4, 1996, is the first day in the sample. May 8, 2006, is a low volatility day. December 1, 2008, represents the peak of the 2008 financial crisis. December 31, 2019, is the last day of the sample.

Figure 2.4: Model (2.2) fitted surfaces compared with the benchmark surfaces

where σ_i^O is the observed IV for the i^{th} quote available and M_i, τ_i are its associated moneyness and time to maturity, respectively.

Figure 2.5 shows that the daily RMSE of Model (2.2) is lower than that of the GG and CT models for most of the sample period. Despite the increase in the number of quoted option contracts in more recent periods, the fitting performance of Model (2.2) does not deteriorate over time. This is in sharp contrast with the GG and CT benchmarks. Note that since the GG model is estimated on the log IV surfaces and the measurement error is based

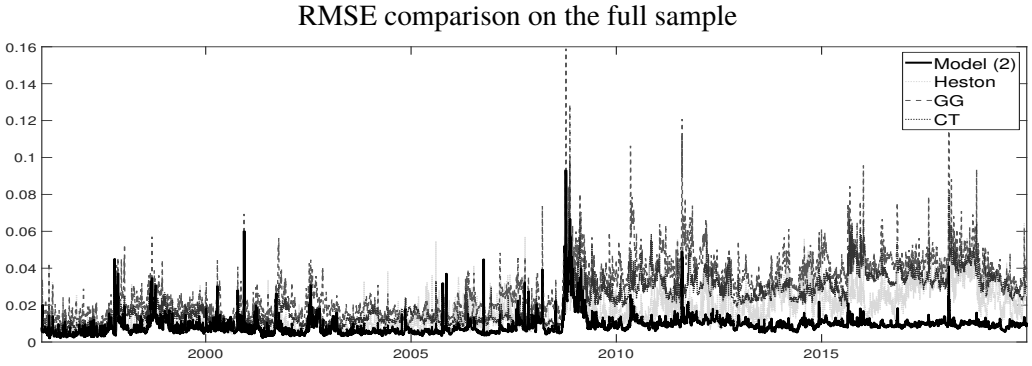


Figure 2.5: RMSE across time

on the IV level, the RMSE does not only capture the variability of the error term, but also the convexity bias caused by Jensen's inequality. The drop in performance of the CT and GT models following the 2008 financial crisis is linked to the wider range of moneyness levels traded after the crisis. Due to the asymptotic behaviour of the models, capturing the surface shape with deeper out-of-the-money options becomes an issue.

Table 2.2 displays the average RMSE (ARMSE) across time over different subregions of the IV surface. The ARMSE across the entire surface for the GG and CT benchmarks is four times larger than that of Model (2.2). The quality of the fit obtained for the CT and GG benchmarks is strongly sensitive to moneyness and time to maturity. In particular, the two models poorly match the OTM put implied volatilities. Interestingly, the ARMSEs associated with the Model (2.2) specification are of similar magnitude across all moneyness and time to maturity buckets.

2.3.5 Arbitrage opportunities

To verify if Model (2.2) generates prices consistent with no-arbitrage principles, a screening procedure inspired by the work of Davis and Hobson, 2007 is applied. The detection of prices violating no-arbitrage restrictions is performed on both the calibrated surfaces and sample observations. The comparison of the number of such violations in both datasets

Model	$M \leq -0.1$ (call)	$-0.1 < M \leq 0.1$	$M > 0.1$ (put)	All
Model (2.2)	0.0141	0.0096	0.0098	0.0107
CT	0.0126	0.0158	0.0387	0.0308
GG	0.0223	0.0148	0.0505	0.0401
Number of options	638,645	846,247	1,983,623	3,468,515
Model	$\tau \leq 60$	$60 < \tau \leq 180$	$\tau > 180$	All
Model (2.2)	0.0116	0.0095	0.0109	0.0107
CT	0.0363	0.0309	0.0271	0.0308
GG	0.0537	0.0249	0.0399	0.0401
Number of options	856,524	1,090,187	1,521,804	3,468,515

The average RMSE over time is reported for each bucket of moneyness (M) and days to maturity (τ). The sample period is January 4, 1996, to June 26, 2019.

Table 2.2: Average RMSE over time from IV surface estimation

serves as a sanity check to assess the propensity of Model (2.2) to either (i) generate prices incompatible with the absence of arbitrage, or (ii) smooth out arbitrage opportunities found in the data.

Davis and Hobson, 2007 study instances of so-called static arbitrage opportunities, which is a convenient relaxation of the general no-arbitrage theory outlined for instance in Delbaen and Schachermayer, 1994. The distinction between both types of arbitrages lies in the difference between information sets (see Carr and Madan, 2005 for a more thorough discussion). Davis and Hobson, 2007 provide sufficient conditions precluding the presence of static arbitrage opportunities within a point-in-time set of European call option prices for several strikes and maturities on a single underlying asset. They show that the absence of no-arbitrage violations within prices of a set of butterfly spread and calendar spread portfolios ensures that the entire set of option prices is arbitrage-free. The approach for the construction of such spread portfolios is described in Appendix A.4.

When applying the screening on observed option quotes, boundaries of the bid-ask interval are used instead of mid-prices. More precisely, whenever a call option is purchased (sold) in the construction of the spread portfolio, the ask (bid) price is considered. Indeed, using mid-prices instead would have led to flagging spurious arbitrage opportunities in the

		Observed prices		Fitted surfaces		
		Arbitrage detected	% Arbitrage detected	Arbitrage detected	% Arbitrage detected	Number of tests
$\tau \leq 60$	$M \leq 0$	662	0.27%	340	0.14%	245,081
	$0 < M \leq 0.3$	791	0.33%	0	0%	243,235
	$M > 0.3$	148	0.04%	109	0.03%	368,312
$60 < \tau \leq 180$	$M \leq 0$	638	0.2%	26	0.01%	325,928
	$0 < M \leq 0.3$	1,263	0.3%	0	0%	426,276
	$M > 0.3$	1,978	0.45%	0	0%	439,667
$\tau > 180$	$M \leq 0$	1,122	0.23%	0	0%	498,182
	$0 < M \leq 0.3$	1,954	0.38%	0	0%	516,449
	$M > 0.3$	11,446	2.31%	0	0%	496,184

Summary statistics on violations per moneyness and time to maturity buckets of no-arbitrage constraints on butterfly spreads and calendar spreads designed as per the methodology outlined in Section 2.3.5 inspired by Davis and Hobson, 2007. The numbers and proportions of violations, which are aggregated across all dates of the sample, are reported for both the data sample and fitted surfaces obtained with Model (2.2).

Table 2.3: Detected static arbitrage opportunities

data which would have been impossible to realize due to the limited ability to trade within the bid-ask range. Conversely, when detecting no-arbitrage violations among calibrated surfaces, prices generated by Model (2.2) are used without any correction for illiquidity considerations. Such discrepancy in testing is a conservative choice as it puts more stringent requirements on Model (2.2) for its prices to satisfy no-arbitrage constraints during the screening. Moreover, the Davis and Hobson, 2007 methodology is based on call option prices, whereas the current sample contains quotes for both call and put options. Thus, to screen for arbitrage opportunities, put option bid and ask quotes in the data are transformed into call ask and bid prices using the put-call parity.

Table 2.3 displays the number of arbitrage opportunities detected in the entire data sample and on fitted surfaces. Numbers provided are aggregates across all dates of the data sample. For each date, one butterfly spread arbitrage test is performed for each option in the dataset. Moreover, there is one calendar spread test for almost all options in the dataset, with the exception of options for which the construction of the calendar spread is impossible

due to the lack of traded options that would have been needed for inclusion into the spread portfolio. For instance, options whose maturity is the last one available on a given day are not tested for calendar spread arbitrage. The counts provided in Table 2.3 thus include both the butterfly and calendar arbitrage tests. Results show that the calibrated surfaces exhibit fewer arbitrage opportunities than do quotes from the data sample in all buckets of moneyness and time to maturity. This result provides reassurance about the suitability of factors designed in Model (2.2). Indeed, the model tends to correct for the arbitrage opportunities found in the data while it generally avoids producing prices violating no-arbitrage constraints.

2.4 Applications of the volatility surface model

Having a complete surface with implied volatilities available on a large range of moneyness and maturity has several practical applications, some of which are presented in this section. Two main categories of applications are outlined: derivatives pricing and risk management.

2.4.1 Derivatives pricing applications

The most direct application of the volatility surface model developed herein is the pricing of financial derivatives. The model can be applied in conjunction with three main approaches, each described subsequently: direct interpolation or extrapolation of the implied volatility, the Carr and Madan, 2001 formula, and the extraction of the underlying asset price risk-neutral density. The three techniques are respectively tailored to different classes of derivatives, which explains why all three are necessary.

Within the Model (2.2) framework, option prices are obtained by substituting the implied volatility $\sigma(M, \tau)$ into the Black-Scholes formula. More precisely, the Black, 1976 formula using the forward price instead of the underlying asset price is considered. Indeed, both the forward-based and underlying-based pricing formulas are equivalent in theory, but the former takes advantage of the OptionMetrics dataset which provides forward prices rather than underlying prices, and it allows circumventing the cumbersome task of performing a daily extraction of implied dividend rates for the various option maturities.

Because the moneyness defined in Equation (2.1) can be inverted to retrieve the strike price through $K = F_{0,\tau}\mathrm{e}^{-\sqrt{\tau}M}$, the call[11] and put prices are

$$c(M,\tau) = \mathrm{e}^{-r\tau}F_{0,\tau}\left[\Phi(\delta_1(M)) - \mathrm{e}^{-\sqrt{\tau}M}\Phi(\delta_2(M))\right], \tag{2.4}$$

$$p(M,\tau) = \mathrm{e}^{-r\tau}F_{0,\tau}\left[-\Phi(-\delta_1(M)) + \mathrm{e}^{-\sqrt{\tau}M}\Phi(-\delta_2(M))\right], \tag{2.5}$$

where Φ is the standard normal cumulative distribution function,

$$\delta_1(M) = \frac{M}{\sigma(M,\tau)} + \frac{1}{2}\sigma(M,\tau)\sqrt{\tau}, \text{ and } \delta_2(M) = \frac{M}{\sigma(M,\tau)} - \frac{1}{2}\sigma(M,\tau)\sqrt{\tau}.$$

Interpolation and extrapolation of the implied volatility

Derivatives traded over-the-counter often lack liquidity. Their pricing through mark-to-market procedures can therefore be a challenging exercise. Determining prices of illiquid derivatives is needed for several reasons: balance sheet assessment and corresponding risk metrics calculation, financial statement reporting, or margin calls determination in the presence of compensation by a clearinghouse.

Figure 2.1 shows that only a few maturities are actively traded on any day. A vanilla option whose strike or maturity is not quoted publicly on an exchange is considered illiquid. For such a contract, the pricing is made completely seamless by Model (2.2) as the implied volatility can be directly obtained by substituting the option moneyness and maturity in the latter formula. The computational effort required is close to nil, which makes the approach extremely convenient for the quick valuation of a large portfolio of derivatives.

[11] Depending on the application, the call option prices are sometimes expressed as a function of the strike price instead of the moneyness. In the Model (2.2) framework, the call price becomes

$$C(K,\tau) = \exp(-r\tau)\left(F_{0,\tau}\Phi(d_1(K)) + K\Phi(d_2(K))\right)$$

with

$$d_1(K) = \frac{1}{\sigma\left(\frac{1}{\sqrt{\tau}}\log\frac{F_{0,\tau}}{K},\tau\right)\sqrt{\tau}}\log\frac{F_{0,\tau}}{K} + \sigma\left(\frac{1}{\sqrt{\tau}}\log\frac{F_{0,\tau}}{K},\tau\right)\sqrt{\tau},$$

$$d_2(K) = d_1(K) - \sigma\left(\frac{1}{\sqrt{\tau}}\log\frac{F_{0,\tau}}{K},\tau\right)\sqrt{\tau}.$$

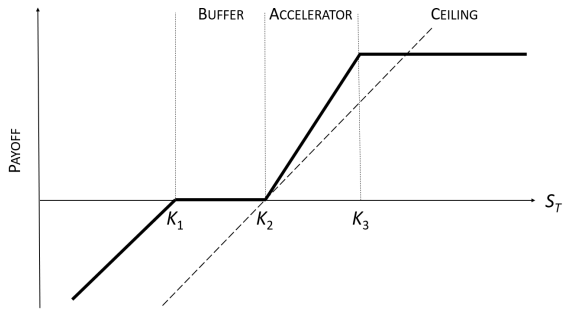


Figure 2.6: Typical Index-linked S&P 500 note payoff function

Consider for instance the mark-to-market of an index-linked S&P 500 note. A typical terminal payoff is displayed in Figure 2.6. There are three thresholds: K_1 for the “buffer" region, K_2 for the “accelerator" region, and K_3 for the “ceiling" region. The terminal payoff X_τ can be replicated with

$$X_\tau = S_\tau - C\left(K_1, 0\right) + \alpha C\left(K_2, 0\right) - \alpha C\left(K_3, 0\right), \tag{2.6}$$

where S_τ is the time-τ underlying asset price and α is the return enhancement factor. The price of calls whose payoff appears in (2.6) are often illiquid and thus can seldom be traded directly, which complicates the valuation of the contract. Nevertheless, the IV surface model (2.2) allows for continued, accurate mark-to-market of the note.

The left panels of Figure 2.7 present the estimated note price using either the quoted options or Model (2.2) implied volatility surfaces, whereas the right panels contain the difference between the two approaches. The rows correspond to the pre-2008 crisis period (first row), the financial turmoil (second and third row), and its aftermath (last row). Because only a limited number of maturities and strikes are quoted, it is often not possible to observe quotes for options whose characteristics match exactly these of options embedded in the note, an issue for the model-free approach. To remedy this issue, two alternative notes are priced for the two closest (smaller and larger) maturities using the closest strikes available for the buffer, the accelerator, and the ceiling. The desired note price is then computed by linear interpolation. Due to liquidity issues, some strike prices appear and

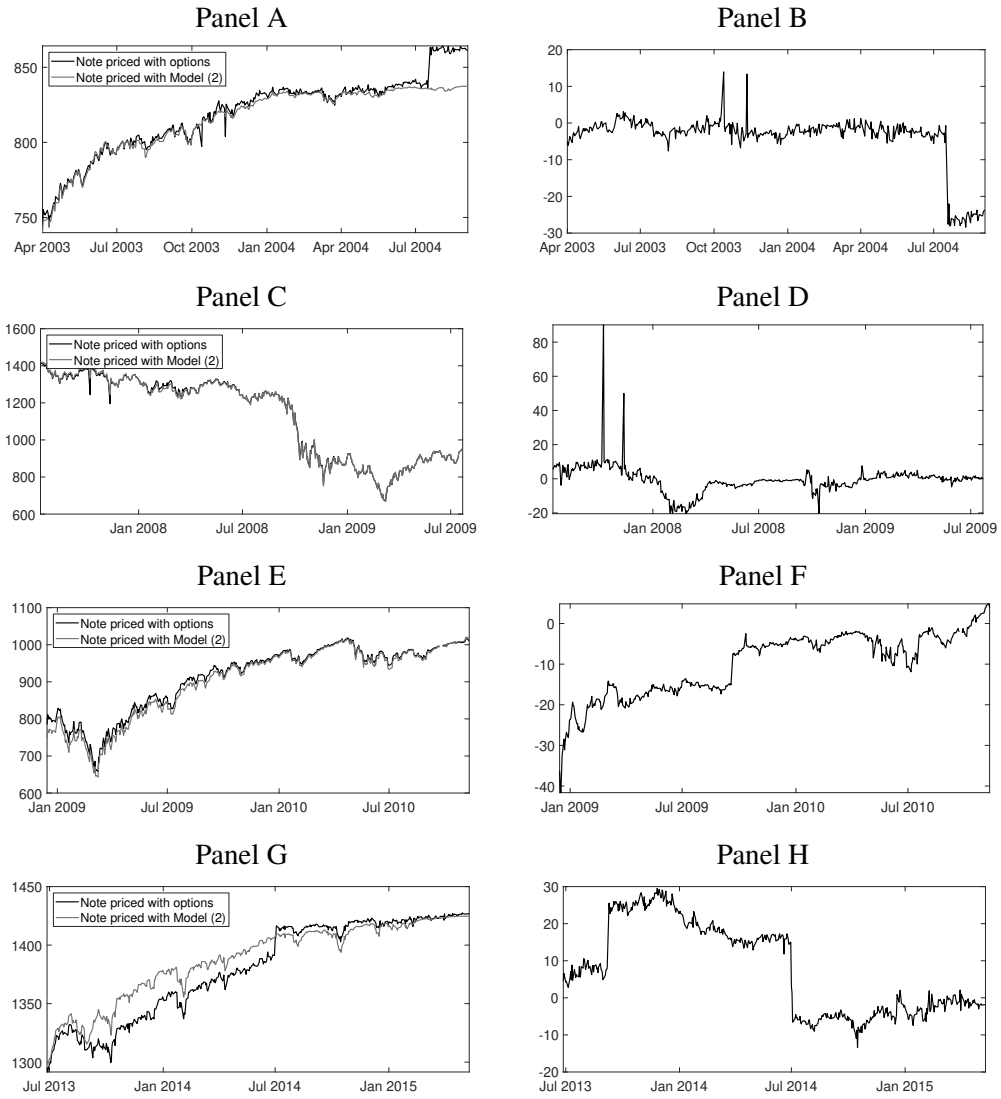
disappear during the note's life, sometimes generating undesired price variation, which can be witnessed in Panels A, C, E, and G of Figure 2.7. This phenomenon does not seem to be related to the economic cycle and can appear at any moment during the note's life.

These undesired price variations can result in unnecessary, large, and erroneous mark-to-market movements. For example, in the two first rows of Figure 2.7, the note price computed with the model-free approach spikes on two occasions at the end of 2003 and 2007. Panel A and Panel G also display long periods where note prices estimated with quoted options are either overvalued or undervalued. However, using Model (2.2)'s implied surfaces generates more stable note prices during the note's life span.

The Carr & Madan formula and its discretization

For the pricing of several other contingent claims which are not call and put options, the interpolation and extrapolation approach serves as a building block within the Carr and Madan, 2001 methodology. Such a method relies on options prices for a continuum of strike prices, which is provided by Model (2.2).

Carr and Madan, 2001 show that any twice differentiable payoff function can be evaluated using infinitely many out-of-the-money put and call option prices with the same time-to-maturity as the payoff horizon. In practice, the valuation of such a payoff is applied in a model-free fashion using a discrete set of traded options. The volatility surface of Model (2.2) improves the numerical implementation in two ways. First, the integrals involving out-of-the-money options can be truncated at levels of moneyness beyond those taken from the data. Second, these same integrals do not need to be discretized with respect to the strike price dimension.



Left panels display the price of a note estimated using (i) the closest observed quoted option prices in conjunction with linear interpolation and (ii) Model (2.2). Right panels display the price difference between the two methods. The first row is a note with a maturity of 550 days and strikes $K_1 = 775$, $K_2 = 850$, and $K_3 = 975$. The second row is a note with a maturity of 800 days and strikes $K_1 = 1425$, $K_2 = 1550$, and $K_3 = 1675$. The third row is a note with a maturity of 730 days and strikes $K_1 = 1000$, $K_2 = 1200$, and $K_3 = 1400$. The last row is a note with a maturity of 730 days and strikes $K_1 = 1350$, $K_2 = 1550$, and $K_3 = 1700$.

Figure 2.7: Mark-to-market of structured notes

The main result of Carr and Madan, 2001 is adapted herein to provide an integration with respect to the moneyness defined in Equation (2.1) rather than the strike price. In what follows, the cash-and-carry relationship $F_{t,\tau} = S_t \exp\left(\left(r - q\right)\tau\right)$ is assumed to hold, where r is the risk-free rate and q is the continuously compounded dividend yield.

For a twice differentiable payoff function f with second derivative f'', the Carr and Madan, 2001 formula becomes

$$\begin{aligned}\mathrm{e}^{-r\tau}\mathrm{E}^{\mathbb{Q}}\left[f\left(S_\tau\right)\right] =& \mathrm{e}^{-r\tau} f\left(F_{0,\tau}\right) + \sqrt{\tau}F_{0,\tau}\int_{-\infty}^{0} f''\left(F_{0,\tau}\mathrm{e}^{-\sqrt{\tau}M}\right) c\left(M,\tau\right)\mathrm{e}^{-\sqrt{\tau}M}dM \\ &+ \sqrt{\tau}F_{0,\tau}\int_{0}^{\infty} f''\left(F_{0,\tau}\mathrm{e}^{-\sqrt{\tau}M}\right) p\left(M,\tau\right)\mathrm{e}^{-\sqrt{\tau}M}dM\end{aligned} \tag{2.7}$$

where $\mathbb{Q}$ is the risk-neutral probability measure. The proof is in Appendix A.5. A numerical implementation of Equation (2.7) requires truncation of the tails of the integral. The lower and upper bounds $\underline{m}$ and $\overline{m}$ are set such that

$$\begin{aligned}\mathrm{e}^{-r\tau}\mathrm{E}^{\mathbb{Q}}\left[f\left(S_\tau\right)\right] \cong& \mathrm{e}^{-r\tau} f\left(F_{0,\tau}\right) + \sqrt{\tau}F_{0,\tau}\int_{\underline{m}}^{0} f''\left(F_{0,\tau}\mathrm{e}^{-\sqrt{\tau}M}\right) c\left(M,\tau\right)\mathrm{e}^{-\sqrt{\tau}M}dM \\ &+ \sqrt{\tau}F_{0,\tau}\int_{0}^{\overline{m}} f''\left(F_{0,\tau}\mathrm{e}^{-\sqrt{\tau}M}\right) p\left(M,\tau\right)\mathrm{e}^{-\sqrt{\tau}M}dM.\end{aligned} \tag{2.8}$$

The bounds selection is determined by the extrapolation capacity of the implied volatility surface model. Because Model (2.2) provides a closed-form solution for option prices for a continuum of moneyness, the two integrals in Equation (2.8) can be computed without being adversely impacted by the discretization bias stemming from the availability of only a finite number of strike prices.

Among all potential applications, a particularly interesting one consists in recovering risk-neutral moments of the underlying asset price. The continuum of out-of-the-money options across moneyness levels is also involved in the calculation of the VIX and the valuation of variance swaps (Neuberger, 1994, Carr and Madan, 2001, Schneider and Trojani, 2015). As a matter of fact, the VIX is the expected log-return of the S&P 500 forward contract. These economic quantities can be assessed with increased accuracy when the implied volatility surface is smoothed and not restricted to available market data points.

	December 1, 2008						December 2, 2019					
	$\tau = 45$			$\tau = 199$			$\tau = 45$			$\tau = 199$		
	A	B	C	A	B	C	A	B	C	A	B	C
VIX	66.72	66.70	66.70	54.95	57.68	58.80	15.48	16.10	16.22	18.27	18.11	18.17
Skewness	-1.36	-1.41	-1.41	-1.35	-1.24	-1.63	-2.40	-2.07	-2.31	-2.43	-2.52	-2.65
Kurtosis	5.98	6.21	6.22	5.08	4.69	7.15	11.84	10.39	14.10	11.70	13.89	15.72
Num options	100			47			218			99		

Estimation of the VIX and risk-neutral moments of the forward contract log-return $\log(F_{\tau,0}/F_{0,\tau})$ using either the quoted options and a discretization of Equation (2.8) (column A) or the Equation (2.8) on fitted surfaces. In the latter case, the moneyness range is the one provided by quoted options of the same maturity (column B) or it is extrapolated (column C). The extrapolated moneyness range is the largest and smallest moneyness (M) observed for all maturities for a given day.

Table 2.4: S&P 500 return risk-neutral moments

Table 2.4 presents the VIX and risk-neutral moments of the forward contract log-return $\log(F_{\tau,0}/F_{0,\tau})$ using either the quoted options and a discretization of Equation (2.8) (column A) or the Equation (2.8) on fitted surfaces. In the latter case, the moneyness range is the one provided by the quoted options with interpolation only[12] (column B), or it is extrapolated (column C). For short maturities, the number of quoted options spans a wider range resulting in similar results for the three approaches. However, for the longest two maturities, there are important disparities across the three methods due to the lack of observations.[13]

[12]The integration is applied on the same moneyness range as that provided by quoted options of the same maturity.

[13]As an additional verification of Model (2.2)'s extrapolation ability, we work with simulated instead of real data. We choose to simulate the Andersen et al., 2015 underlying asset dynamics model with various starting values for the latent variables to create different scenarios. Then we use Model (2.2) to extrapolate the IV surface from simulated option prices. We find a high-quality fit between the extrapolated and the simulated surfaces. Furthermore, we confirm a very accurate estimation of the risk-neutral moments using the extrapolated IV surfaces.

Extraction of the risk-neutral density

A particularly relevant application of option pricing on a continuum of strikes is the extraction of the risk-neutral density of the underlying asset. Breeden and Litzenberger, 1978 initially noted that such distribution is related to the convexity of option prices with respect to the strike, and that it could in theory be retrieved from a continuous implied volatility surface. Although the risk-neutral underlying price distribution is interesting in itself, its extraction provides incremental benefits in terms of derivatives pricing over the two other aforementioned methods. Indeed, some derivatives are not options and do not have a twice differentiable payoff function, rendering the interpolation and the Carr and Madan, 2001 methods inapplicable. Digital options with binary payoffs are a particular example. For such derivatives, obtaining the risk-neutral distribution is a necessary endeavor for pricing.

For a given day, the index spot price risk-neutral density function g_τ in τ years can be calculated through

$$g_\tau(K) = \mathrm{e}^{r\tau}\frac{\partial^2 C(K,\tau)}{\partial K^2}, \quad K > 0.$$

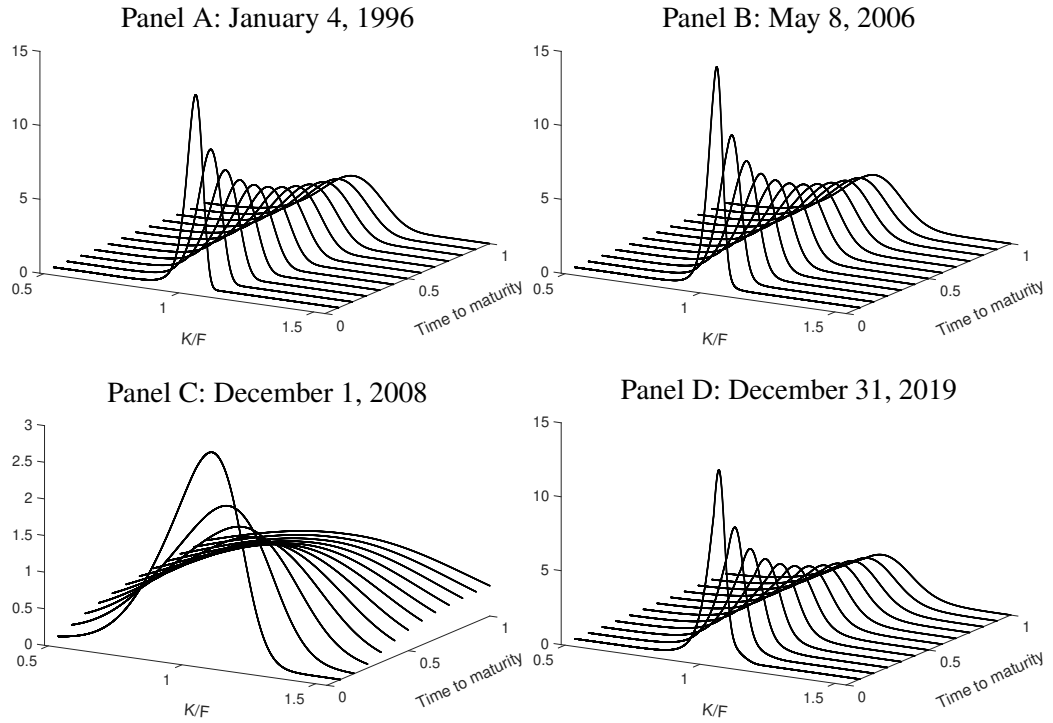
Within the Model (2.2) framework, the risk-neutral density function[14] is

$$\mathrm{e}^{r\tau}\frac{\partial^2 c}{\partial K^2} = \frac{F_{0,\tau}}{K^2}\varphi\left(\delta_1\right)\left(\frac{1}{\sqrt{\tau}}\frac{\partial \delta_1}{\partial M} - \frac{1}{\sqrt{\tau}}\delta_1\frac{\partial \delta_1}{\partial M}\frac{\partial \sigma}{\partial M} + \frac{1}{\sqrt{\tau}}\frac{\partial^2 \sigma}{\partial M^2}\right), \tag{2.9}$$

where $\varphi(z) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}z^2\right)$ is the density function of a standard normal random variable,

$$\begin{aligned}
\frac{\partial \delta_1}{\partial M} &= \frac{1}{\sigma} - \left(\frac{M}{\sigma^2} - \frac{1}{2}\sqrt{\tau}\right)\frac{\partial \sigma}{\partial M}, \\
\frac{\partial \sigma}{\partial M} &= \beta_3 \mathbb{1}_{M\geq 0} + \beta_3\left(1 - \left(\frac{\mathrm{e}^{2M}-1}{\mathrm{e}^{2M}+1}\right)^2\right)\mathbb{1}_{M<0} + \beta_4 2M\mathrm{e}^{-M^2}\log\frac{T}{T_{\max}} \\
&\quad -\beta_5 81 M^2 \mathrm{e}^{27M^3}\log\frac{T}{T_{\max}}\mathbb{1}_{M<0}, \\
\frac{\partial^2 \sigma}{\partial M^2} &= -\beta_3 8\mathrm{e}^{2M}\frac{\mathrm{e}^{2M}-1}{\left(\mathrm{e}^{2M}+1\right)^3}\mathbb{1}_{M<0} + \beta_4 2\left(1-2M^2\right)\mathrm{e}^{-M^2}\log\frac{T}{T_{\max}} \\
&\quad -\beta_5\left(162 + 6561M^3\right)M\mathrm{e}^{27M^3}\log\frac{T}{T_{\max}}\mathbb{1}_{M<0}.
\end{aligned}$$

[14] The arguments of the functions are omitted to simplify the notation.



Risk-neutral density functions $h_\tau(y) = \exp(y)g_\tau(\exp(y))$ of the log-prices derived from Model (2.2) call prices. January 4, 1996, is the first day in the sample. May 8, 2006, is a low volatility day. December 1, 2008, represents the peak of the 2008 financial crisis. December 31, 2019, is the last day of the sample.

Figure 2.8: Log-price risk-neutral densities implied by Model (2.2).

The proof is in Appendix A.6.1. Appendix A.6.1 shows that $\int_0^\infty g_\tau(K)dK = 1$ which is one of the fundamental properties of a density function.

For the same four days that were selected in Figure 2.1, a set of risk-neutral log-price densities $h_\tau(y) = \exp(y)g_\tau(\exp(y))$, each corresponding to a different maturity τ, is displayed in Figure 2.8. In all cases, volatility increases with time-to-maturity. Obviously, the financial crisis shows a greater volatility for all maturities. For all four dates, Model (2.2) implied densities exhibit negative skewness which is particularly apparent in 2008. Appendix A.7 presents the analogue of Figure 2.8 for both the CT and GG models. The polynomial structure of GG model which misbehaves for moneyness levels lying outside

the observed range and the discontinuities in the CT model produce risk-neutral density functions that may take negative values and have irregular patterns, especially in the tails of the distribution.

2.4.2 Risk management for options

The smooth implied volatility surface of Model (2.2) provides the additional benefit, on top of derivatives pricing, to allow for the computation of "Greek" parameter sensitivities of option prices that are essential for replicating option payoffs or, more broadly speaking, managing positions on option portfolios. As shown in Appendix A.6.2, the call option delta and gamma[15] are

$$\Delta = \mathrm{e}^{-q\tau}\left(\Phi\left(\delta_1\right) + \varphi\left(\delta_1\right)\frac{\partial\sigma}{\partial M}\right), \tag{2.10}$$

$$\Gamma = \frac{\mathrm{e}^{-q\tau}}{\sqrt{\tau}S_0}\varphi\left(\delta_1\right)\left(\frac{\partial\delta_1}{\partial M} - \delta_1\frac{\partial\delta_1}{\partial M}\frac{\partial\sigma}{\partial M} + \frac{\partial^2\sigma}{\partial M^2}\right), \tag{2.11}$$

where $\partial d_1/\partial M$, $\partial\sigma/\partial M$, and $\partial^2\sigma/\partial M^2$ are defined in Section 2.4.1. If the implied volatility surface is flat, then $\partial\sigma/\partial M = \partial^2\sigma/\partial M^2 = 0$ and the above formulas simplify into the Black-Scholes Greeks. Therefore, the extra terms are measuring the sensitivity of the implied volatility to the variation of the underlying asset price through the moneyness variation.

Using the factor specification of implied volatility, the computation of $\partial\sigma/\partial M$ and $\partial^2\sigma/\partial M^2$ is immediate and its accuracy is not undermined by the limited availability of traded strikes and maturities. This clearly represents a significant advantage for option risk management purposes.

It should also be noted that the delta and the gamma defined above do not depend on any assumption regarding the dynamics of the underlying asset. Rather, they are consistent with the observed shape of the volatility smile and, as such, they comply with Bates, 2005's

[15]The functions' arguments are omitted to simplify the notation.

definition of smile-implied Greeks:[16]

$$\Delta = \frac{1}{S_0}\left(c - K\frac{\partial c}{\partial K}\right), \qquad \Gamma = \frac{K^2}{S_0^2}\frac{\partial^2 c}{\partial K^2}.$$

This is particularly useful for illiquid options, e.g. OTC transactions, having a moneyness or a maturity that is quite remote from those of publicly quoted ones; a model-free assessment of associated partial derivatives relying purely on finite differences would most likely prove unstable due to the paucity of related observations.

Another Greek parameter of high importance is the vega, i.e., the sensitivity of the option price with respect to the implied volatility. As shown in Appendix A.6.2, the call option vega is

$$\vartheta = \frac{\partial c}{\partial \sigma} = \mathrm{e}^{-r\tau} F_{0,\tau} \varphi\left(\delta_1\right) \sqrt{\tau}.$$

The call price sensitivity to the long-term volatility level is $\frac{\partial \sigma}{\partial \beta_1}\frac{\partial c}{\partial \sigma}$, and its sensitivity to the maturity slope is $\frac{\partial \sigma}{\partial \beta_2}\frac{\partial c}{\partial \sigma}$.

2.5 Conclusion

This study develops a factor model for the representation of volatility surfaces. The design of the model makes it very parsimonious, easy to interpret, seamless to estimate and quick to compute. Factors underlying the representation possess very intuitive meaning, i.e., long-term level, time-to-maturity slope, moneyness slope, smile attenuation and smirk. They are designed in accordance with the most salient empirical features of volatility surfaces as evidenced by the commonalities between loading vectors obtained from a PCA analysis on historical surfaces and the designed model factors.

The construction of the model factors leads to volatility-surface-implied underlying asset densities that are well-behaved and smooth. In particular, the convenient asymptotic behaviour of the surface imposed by the model for large maturities and moneyness levels

[16] Bates, 2005 formulas for the delta and the gamma rely on the scale invariance of option prices, a property verified by Model (2.2). The implied volatility is a function of the moneyness $M = \frac{1}{\sqrt{\tau}} \ln \frac{F_{0,\tau}}{K}$ which is not affected when both the strike price and the underlying price are multiplied by a constant.

allows for extrapolation beyond ranges observed in the data. This key characteristic of the model is paramount in the applications discussed in this paper which crucially rely upon extrapolation ability. Moreover, performing such extrapolation is a necessary endeavour for multiple market participants requiring frequent marking-to-market of illiquid options. Furthermore, our model produces IV surfaces that are twice continuously differentiable with respect to the moneyness level, thereby limiting instances of model prices inconsistent with no-arbitrage principles.

The fitting performance is assessed on historical S&P 500 option prices obtained from the OptionMetrics database. Benchmarking against alternative literature models provides evidence of strong calibration outperformance versus the competing Heston, 1993, Goncalves and Guidolin, 2006, and Chalamandaris and Tsekrekos, 2011 models, especially in recent periods. Indeed, the specification of the two latter factor-based benchmarks exhibits explosive asymptotic behaviour, leading to poor fitting performance for deep-out-of-the-money puts. By contrast, our model is much better suited for extrapolation to extreme moneyness levels.

A screening procedure based on the Davis and Hobson, 2007 methodology highlights that option prices generated by the model developed herein very infrequently violate no-arbitrage restrictions, and that the model tends to smooth out theoretical arbitrage opportunities found in the data.

Several applications of the model are related to derivatives pricing and risk management. Three different derivatives pricing methods, namely (i) pure interpolation or extrapolation, (ii) the Carr-Madan formula and (iii) risk-neutral density extraction through the Breeden and Litzenberger, 1978 methodology, are shown to be applicable in conjunction with the model, using fitted surfaces stemming from the latter as their main input. The presented pricing methods enable the valuation of different classes of illiquid derivatives for which the mark-to-market is not straightforward. Closed-form expressions for European option Greek letters (e.g., the delta and the gamma) in the context of the implied volatility surface model are presented. The straightforward computation of option sensitivities in this setting can facilitate the implementation of hedging procedures for options, especially

for those whose strike or maturity is not quoted on public exchanges.

Further work expanding on the present study could include designing hedging strategies consistent with the volatility surface factor representation and assessing the adequacy of the model for options on individual stocks or alternative underlying assets instead of equity indices.

References

Ackerer, D., Tagasovska, N., & Vatter, T. (2019). Deep smoothing of the implied volatility surface. *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*.

Alexander, C., & Nogueira, L. M. (2007). Model-free price hedge ratios for homogeneous claims on tradable assets. *Quantitative Finance*, *7*(5), 473–479.

Ammann, M., & Feser, A. (2019). Robust estimation of risk-neutral moments. *Journal of Futures Markets*, *39*(9), 1137–1166.

Andersen, T. G., Bondarenko, O., & Gonzalez-Perez, M. T. (2015). Exploring return dynamics via corridor implied volatility. *The Review of Financial Studies*, *28*(10), 2902–2945.

Azzone, M., & Baviera, R. (2022). Additive normal tempered stable processes for equity derivatives and power-law scaling. *Quantitative Finance*, *22*(3), 501–518.

Bakshi, G., Cao, C., & Chen, Z. (1997). Empirical performance of alternative option pricing models. *The Journal of Finance*, *52*(5), 2003–2049.

Bakshi, G., & Kapadia, N. (2003). Delta-hedged gains and the negative market volatility risk premium. *The Review of Financial Studies*, *16*(2), 527–566.

Bakshi, G., & Madan, D. (2000). Spanning and derivative-security valuation. *Journal of Financial Economics*, *55*(2), 205–238.

Bates, D. S. (2005). Hedging the smirk. *Finance Research Letters*, *2*(4), 195–200.

Bayraktar, E., & Yang, B. (2011). A unified framework for pricing credit and equity derivatives. *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics*, *21*(3), 493–517.

Birru, J., & Figlewski, S. (2012). Anatomy of a meltdown: The risk neutral density for the S&P 500 in the fall of 2008. *Journal of Financial Markets*, *15*(2), 151–180.

Black, F. (1976). The pricing of commodity contracts. *Journal of Financial Economics*, *3*(1-2), 167–179.

Black & Scholes. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, *81*(3), 637–654.

Bliss, R. R., & Panigirtzoglou, N. (2002). Testing the stability of implied probability density functions. *Journal of Banking & Finance*, *26*(2-3), 381–422.

Breeden, D. T., & Litzenberger, R. H. (1978). Prices of state-contingent claims implicit in option prices. *Journal of Business*, *51*(4), 621–651.

Carr, P., & Madan, D. (2001). Towards a theory of volatility trading. *Option Pricing, Interest Rates and Risk Management, Handbooks in Mathematical Finance*, *22*(7), 458–476.

Carr, P., & Madan, D. B. (2005). A note on sufficient conditions for no arbitrage. *Finance Research Letters*, *2*(3), 125–130.

Carr, P., & Wu, L. (2014). Static hedging of standard options. *Journal of Financial Econometrics*, *12*(1), 3–46.

Célérier, C., & Vallée, B. (2017). Catering to investors through security design: Headline rate and complexity. *The Quarterly Journal of Economics*, *132*(3), 1469–1508.

Chalamandaris, G., & Tsekrekos, A. E. (2011). How important is the term structure in implied volatility surface modeling? Evidence from foreign exchange options. *Journal of International Money and Finance*, *30*(4), 623–640.

Conrad, J., Dittmar, R. F., & Ghysels, E. (2013). Ex ante skewness and expected stock returns. *The Journal of Finance*, *68*(1), 85–124.

Cont, R., & Da Fonseca, J. (2002). Dynamics of implied volatility surfaces. *Quantitative Finance*, *2*(1), 45–60.

Cox, J. C., & Ross, S. A. (1976). The valuation of options for alternative stochastic processes. *Journal of Financial Economics*, *3*(1-2), 145–166.

Daglish, T., Hull, J., & Suo, W. (2007). Volatility surfaces: Theory, rules of thumb, and empirical evidence. *Quantitative Finance*, *7*(5), 507–524.

Davis, M. H., & Hobson, D. G. (2007). The range of traded option prices. *Mathematical Finance*, *17*(1), 1–14.

Delbaen, F., & Schachermayer, W. (1994). A general version of the fundamental theorem of asset pricing. *Mathematische Annalen*, *300*(1), 463–520.

Diebold, F. X., & Li, C. (2006). Forecasting the term structure of government bond yields. *Journal of Econometrics*, *130*(2), 337–364.

Dumas, B., Fleming, J., & Whaley, R. E. (1998). Implied volatility functions: Empirical tests. *The Journal of Finance*, *53*(6), 2059–2106.

Fengler, M. R. (2009). Arbitrage-free smoothing of the implied volatility surface. *Quantitative Finance*, *9*(4), 417–428.

Figlewski, S. (2018). Risk-neutral densities: A review. *Annual Review of Financial Economics*, *10*, 329–359.

François, P., & Stentoft, L. (2021). Smile-implied hedging with volatility risk. *Journal of Futures Markets, forthcoming*.

Gao, G. P., Gao, P., & Song, Z. (2018). Do hedge funds exploit rare disaster concerns? *The Review of Financial Studies*, *31*(7), 2650–2692.

Gao, G. P., Lu, X., & Song, Z. (2019). Tail risk concerns everywhere. *Management Science*, *65*(7), 3111–3130.

Garman, M. B., & Kohlhagen, S. W. (1983). Foreign currency option values. *Journal of International Money and Finance*, *2*(3), 231–237.

Goncalves, S., & Guidolin, M. (2006). Predictable dynamics in the S&P 500 index options implied volatility surface. *Journal of Business*, *79*(3), 1591–1635.

Henderson, B. J., & Pearson, N. D. (2011). The dark side of financial innovation: A case study of the pricing of a retail financial product. *Journal of Financial Economics*, *100*(2), 227–247.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The review of financial studies*, *6*(2), 327–343.

Israelov, R., & Kelly, B. T. (2017). Forecasting the distribution of option returns. *Available at SSRN 3033242*.

Jackwerth, J. C. (2000). Recovering risk aversion from option prices and realized returns. *The Review of Financial Studies*, *13*(2), 433–451.

Jackwerth, J. C. (2004). *Option-implied risk-neutral distributions and risk aversion*. Charlotteville: Research Foundation of AIMR.

Jackwerth, J. C., & Rubinstein, M. (1996). Recovering probability distributions from option prices. *The Journal of Finance*, *51*(5), 1611–1631.

Martin, I. (2017). What is the expected return on the market? *The Quarterly Journal of Economics*, *132*(1), 367–433.

Martin, I. W., & Wagner, C. (2019). What is the expected return on a stock? *The Journal of Finance*, *74*(4), 1887–1929.

Nelson, C. R., & Siegel, A. F. (1987). Parsimonious modeling of yield curves. *Journal of Business*, *60*(4), 473–489.

Neuberger, A. (1994). The log contract. *Journal of Portfolio Management*, *20*, 74–80.

Neumann, M., & Skiadopoulos, G. (2013). Predictable dynamics in higher-order risk-neutral moments: Evidence from the S&P 500 options. *Journal of Financial and Quantitative Analysis*, 947–977.

Rebonato, R. (2005). *Volatility and Correlation: The Perfect Hedger and The Fox*. John Wiley & Sons.

Ross, S. (2015). The recovery theorem. *The Journal of Finance*, *70*(2), 615–648.

Schneider, P., & Trojani, F. (2015). Fear trading. *Swiss Finance Institute Research Paper*, (15-03).

Skiadopoulos, G. (2001). Volatility smile consistent option models: A survey. *International Journal of Theoretical and Applied Finance*, *4*(03), 403–437.

Chapter 3

Joint dynamics for the underlying asset and its implied volatility surface: A new methodology for option risk management

Abstract*

The factor-based representation of implied volatility surfaces developed in François et al., 2022 is fitted to daily observed implied volatility surfaces from 1996 to 2019. The extracted time series of factor weight associated with the implied volatility surface's characteristics is then jointly modelled alongside the S&P 500 log-returns. The complex joint dynamics are captured with a stochastic model reminiscent of the GARCH family with Normal Inverse Gaussian innovations and a Gaussian copula. The model is the first step towards predicting option prices and is required for the risk assessment of volatility positions such as straddles, strangles, and the VIX on horizons ranging from one to five days.

Keywords: Implied volatility, Dynamic factor model, Risk management, VIX.

*Joint work with Pascal François, Geneviève Gauthier, and Frédéric Godin. Fraçois and Gauthier are affiliated with HEC Montréal and Godin is affiliated with Concordia university.

3.1 Introduction

The Black and Scholes, 1973 model revolutionized the practice of option trading. Not only because it provided the first arbitrage-based, explicit option pricing formula, but also because it highlighted the bijective relationship between the option premium and its implied volatility (IV). Since then, the IV surface has become the standard representation of option market prices. It is also the output to which the performance of option pricing models is benchmarked (Andersen et al., 2015). A rich option pricing literature has expanded over the last four decades to extend the original Black and Scholes framework. Most contributions to this literature build upon the standard option pricing approach (SOPA). Despite many different modelling assumptions, the SOPA typically proceeds in two steps. First, it posits the dynamics of the underlying asset return under the historical probability measure. Various additional state variables can be specified (such as volatility, interest rates, and convenience yields) to augment the model realism.[2] Next, the SOPA establishes the rules for the arbitrage-free valuation of contingent claims, which implies characterizing a change of probability measure.

In this paper, we opt for a radically different route to model the dynamics of option prices. We build on an early, yet underdeveloped literature that suggests using the IV surface not as a model output but rather as an input. We propose a dynamic extension of the parametric IV surface of François et al., 2022 that we couple with an asymmetric GARCH process with non-Gaussian innovations for the underlying asset return. Our approach (labelled the JIVR model, which stands for Joint Implied Volatility and Return) forecasts the future distributions of S&P 500 index straddle positions and of VIX in a very accurate manner. It successfully does so because using the IV surface integrates the entire market information and contributes to properly assessing higher-order moments and capturing tail risk. A noteworthy merit of the JIVR model is its easy implementation. The estimation of

[2] A non-exhaustive list of modelling innovations includes: GARCH processes in discrete time (Glosten et al., 1993, Duan, 1995, Heston and Nandi, 2000), stochastic volatility (Hull and White, 1987, Heston, 1993, Bates, 1996, Duffie et al., 2000), jumps in returns and in volatility (Merton, 1976, Broadie et al., 2007, Bollerslev and Todorov, 2011), two-factor volatility (Bates, 2000, Christoffersen et al., 2008, Andersen et al., 2015), non-normal innovations (Barndorff-Nielsen, 1998, Carr and Wu, 2004, Christoffersen et al., 2010).

parameters is fast and only requires standard techniques such as least-square regressions and maximum likelihood.[3] To highlight the benefits of our approach, one can make a comparison with the modelling of the yield curve. As noted by Carr and Wu, 2016, 2020, using the IV surface as a model input is reminiscent of the Heath-Jarrow-Morton (Heath et al., 1992) framework for the term structure: a tight fit with current market data is guaranteed, and forecasting option prices takes full advantage of the whole market information.[4] First attempts at directly modelling the IV surface include Zhu and Avellaneda, 1998, Schönbucher, 1999, Fengler, 2006, and Daglish et al., 2007. These early works assume diffusion processes for implied volatilities, and they highlight the difficulty of deriving constraints on the risk-neutral drift to prevent arbitrage. Carr and Wu, 2016 also emphasizes the problematic fit with the current shape of the IV surface. For that reason, Carr and Wu, 2016 restrict the modelling of the IV surface to near-term dynamics. They suggest using their framework in conjunction with a parametric specification for the underlying asset return. In the same spirit, Carr and Wu, 2020 limits the diffusion modelling of the IV to the management of the instantaneous P&L of an option position. Aside from diffusions, a related approach consists in extracting the IV surface explanatory factors in a non-parametric fashion (Cont and Da Fonseca, 2002, Israelov and Kelly, 2017). This method, however, only applies to dense regions of the IV surface. It, therefore, rejects peripheral options (deep-out-the-money and long-maturity) that are very informative about higher-order moments. By contrast, the JIVR model works with an asymptotically well-behaved, parametric IV surface representation that allows for clean interpolation and extrapolation.[5]

[3] This is in sharp contrast with the most recent models of the SOPA that work with several latent variables (e.g., volatility components) and must therefore rely on heavy filtering techniques for estimation (see Bates, 2022, for a recent review).

[4] Pursuing the analogy with term structure modelling, Bates, 2022 writes: "Implied volatility surfaces describe the pricing failures of [the Black-Scholes] model, in the same way that yields inferred from a bond pricing model premised on identical discount rates for all maturities are used to describe nonflat term structures of bond yields." That argument also applies to all Black-Scholes extensions that yield a closer, still imperfect fit on the IV surface.

[5] In their study on S&P 500 options, François et al., 2022 show that this IV surface specification leaves little room for arbitrage. They also document that its fitting performance compares favourably with existing benchmarks.

Most importantly, we show that a dynamic IV surface can be consistently incorporated with a model for underlying asset returns. Focusing on S&P 500 index options, we opt for an asymmetric GARCH with non-Gaussian innovations to capture the large variations observed in returns and in the characteristics of the IV surface. The variance has two components, as suggested by Christoffersen et al., 2013. Oh and Park, 2022 show that the adequate estimation of a two-factor variance process requires additional sources of information from the derivatives market. In contrast to the literature, our framework exploits the available forward-looking information by connecting one of the variance factors to the 1-month, at-the-money (ATM) IV level. This approach displays better fitting performance compared to a conventional one-factor NGARCH while preserving the stability of parameters. Furthermore, the first coefficient of the IV surface, representing the long-term ATM implied volatility, is shown to have a volatility that is proportional to the 1-month ATM implied volatility level – a result in support of Carr and Wu, 2016. The other IV factors follow a GARCH-type process. To complete the JIVR model, a Gaussian copula captures the dependence structure between the S&P 500 log-returns and the IV factors.

While our framework for joint underlying asset returns and IV surface dynamics has many relevant applications, we focus on the risk management of volatility strategies. Two validation tests are considered. First, we perform the backtests of the Value-at-Risk (VaR) for S&P 500 index straddles and strangles from January 2, 1996, to December 31, 2020. Our 1-day and 5-day VaR estimates successfully pass the coverage test (Kupiec et al., 1995) on both tails of the distribution. Second, we forecast the distribution of the VIX index. We use as a benchmark a GARCH model with non-normal innovations directly applied to the VIX. Using an expanding window starting in 2014, the yearly comparison of log-likelihoods documents the superior performance of the JIVR model. Overall, the two aforementioned tests show the ability of the JIVR model to adequately manage volatility positions through the accurate forecasting of IV surface.

The rest of the paper is organized as follows. Section 3.2 presents the data. Section 3.3 reviews the parametric specification that serves as a building block for our dynamic IV surface model. Section 3.4 describes and assembles the components of the JIVR model, which is estimated in Section 3.5. Section 3.6 explains the risk management applications.

Section 3.7 concludes.

3.2 Data

The OptionMetrics database provides the dataset, which includes daily quoted bid and ask prices of European call and put options on the S&P 500 index (SPX options) from the CBOE. The dataset extends from January 4, 1996, to December 31, 2020. On any given day t, the data includes the option strike price K, its maturity, and the associated underlying asset forward price $F_{t,\tau}$, with τ denoting the time-to-maturity. The OptionMetrics database also includes the zero-coupon yield curve and dividend yields.[6]

Option exclusion filters are applied to the dataset, which mostly follows the Bakshi et al., 1997 guidelines. More precisely, we exclude all in-the-money options as well as any option with any of the following characteristics: a time-to-maturity shorter than six trading days, a price lower than $\$3/8$, a bid price of \$0, or a bid-ask spread larger than 175% of the option mid-price.[7] The final dataset includes 6,292 days and a total of 3,814,217 option quotes.[8]

For an option with strike price K and time-to-maturity $\tau = T - t$, the moneyness is defined as

$$M_t = \frac{1}{\sqrt{\tau}} \log \frac{F_{t,\tau}}{K}. \tag{3.1}$$

According to that definition, OTM calls (puts) are associated with a negative (positive)

[6]The OptionMetrics forward price is computed as

$$F_{t,\tau} = Se^{(r_{t,\tau} - q_t)\tau},$$

where $r_{t,\tau}$ is the time-t continuously compounded risk-free rate for time-to-maturity τ, and q_t is the S&P 500 dividend yield.

[7]This last criterion is similar to that of Azzone and Baviera, 2022. When the ratio of the bid-ask spread over the mid-price is large, the latter induces implied volatilities largely deviating from the rest of the IV surface. Options excluded due to this criterion represent a tiny proportion (0.3%) of the total number of options in the dataset.

[8]The IV surface on October 9, 2006, is removed from the dataset because it is very erratic and most likely due to unreliable data on that day.

value for M. Implied volatilities are calculated by inverting the Black and Scholes, 1973 formula, using the mid-quote price as the observed price. Table 3.1 provides a brief description of the sample IVs. Detailed statistics are reported for buckets of moneyness and maturity. The average IV increases with M, reflecting the well-known smile phenomenon, except for DOTM calls ($M < -0.2$), where index option smiles typically exhibit a smirk.

On average, the term structure of IVs is slightly (but not monotonically) decreasing. The standard deviation of the IV decreases with the time-to-maturity, indicating a time-varying time-to-maturity slope.

	Calls		Puts			
	$M \leq -0.2$	$-0.2 < M \leq 0$	$0 < M \leq 0.2$	$0.2 < M \leq 0.8$	$M \geq 0.8$	All
Mean (%)	18.89	15.85	20.83	29.53	47.69	25.37
Standard deviation (%)	7.57	5.89	6.27	7.25	12.69	11.74
Number of contracts	334,482	839,841	841,813	1,439,416	358,665	3,814,217

	Days-to-maturity					
	$\tilde{\tau} \leq 30$	$30 < \tilde{\tau} \leq 90$	$90 < \tilde{\tau} \leq 180$	$180 < \tilde{\tau} \leq 365$	$\tilde{\tau} \geq 365$	All
Mean (%)	27.27	25.54	25.99	25.36	23.00	25.37
Standard deviation (%)	15.48	12.75	12.22	10.77	8.72	11.74
Number of contracts	329,083	1,115,684	722,542	738,335	908,573	3,814,217

Some descriptive statistics of the daily SPX options implied volatility (IV) surfaces from January 4, 1996, to December 31, 2020, grouped by buckets of moneyness and time-to-maturity. M characterizes the moneyness defined in Equation (3.1), and $\tilde{\tau}$ represents the time-to-maturity of the option in days.

Table 3.1: Descriptive statistics of the SPX options implied volatilities

3.3 Factor-based representation of volatility surfaces

This section recalls the static parametric volatility surface representation model of François et al., 2022 on which the subsequent dynamic model is based.

On any day t, François et al., 2022 represent the implied volatility surface (i.e., IVs for any combination of moneyness M and time-to-maturity τ) with the following five-factor

model:

$$\sigma\left(M,\tau,\beta_t\right) = \underbrace{\beta_{t,1}}_{\substack{\text{Long-term}\\ \text{ATM IV}}} + \beta_{t,2}\ \underbrace{e^{-\sqrt{\tau/T_{\text{conv}}}}}_{\text{Time-to-maturity slope}} + \beta_{t,3}\underbrace{\left(M\mathbb{1}_{\{M\geq 0\}} + \frac{e^{2M}-1}{e^{2M}+1}\mathbb{1}_{\{M<0\}}\right)}_{\text{Moneyness slope}}$$
$$+\beta_{t,4}\underbrace{\left(1-e^{-M^2}\right)\log(\tau/T_{\text{max}})}_{\text{Smile attenuation}} + \beta_{t,5}\underbrace{\left(1-e^{(3M)^3}\right)\log(\tau/T_{\text{max}})\mathbb{1}_{\{M<0\}}}_{\text{Smirk}}, \quad \tau\in[T_{\text{min}},T_{\text{max}}] \tag{3.2}$$

where $\beta_t = (\beta_{t,1},\beta_{t,2},\beta_{t,3},\beta_{t,4},\beta_{t,5})$ are the stochastic factors, subsequently referred to as the factor coefficients.

These factors represent the long-term at-the-money (ATM) level, the time-to-maturity slope, the moneyness slope, the smile attenuation over long maturities, and the smirk, respectively.

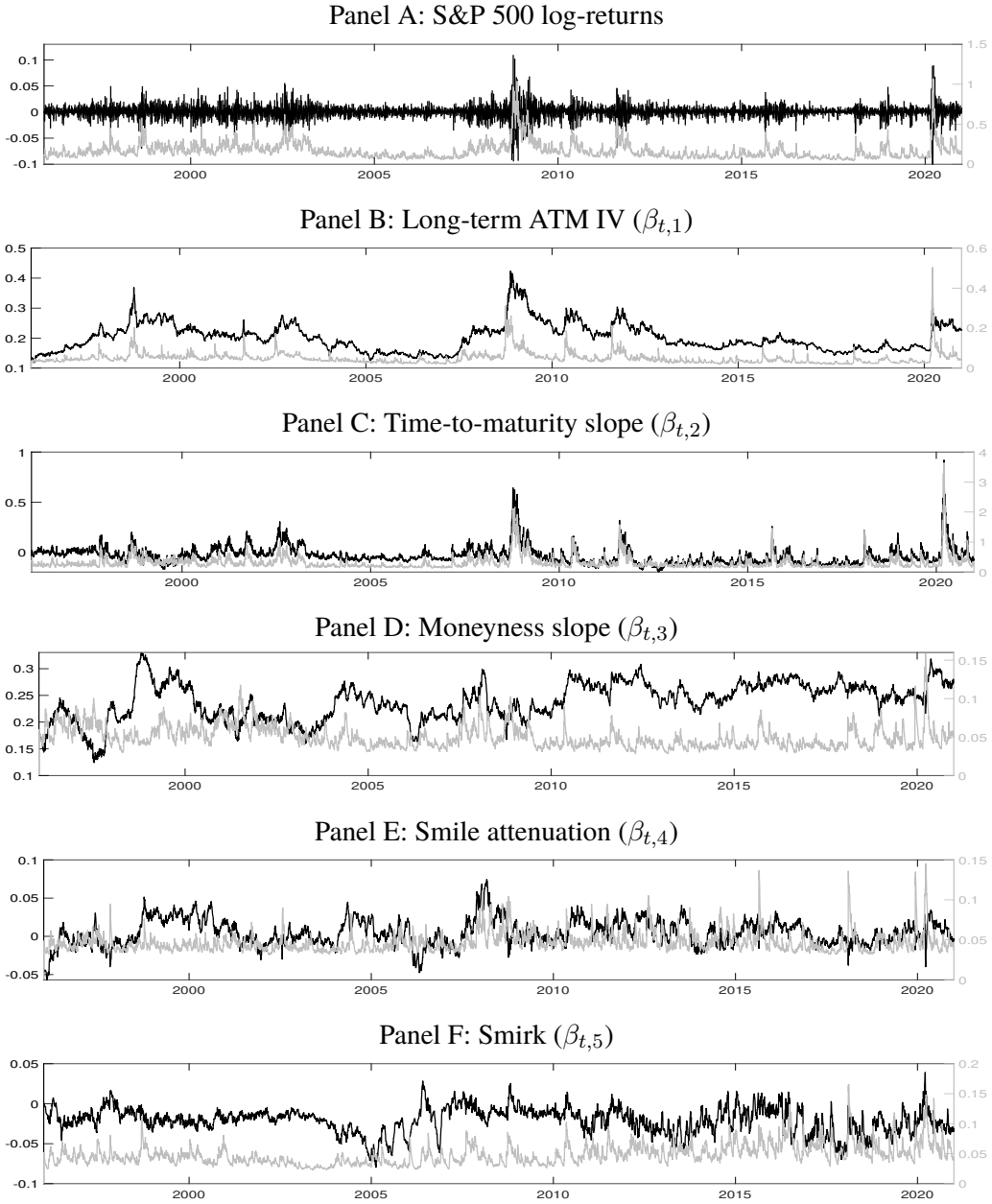
The model is fitted daily (i.e., for each t) to the options prices by minimizing the sum of squared IV differences between the model and the observed prices while incorporating prior information to maintain the financial interpretability of the coefficients.[9]

The black line in Panel A of Figure 3.1 represents the S&P 500 log-returns. The other five panels contain the time series of estimated coefficients $\beta_{t,1},\ldots,\beta_{t,5}$ (black lines). The time-to-maturity slope in Equation (3.2) represents the short-term ATM implied volatility minus the long-term ATM implied volatility.[10] Thus, the coefficient β_2 is negative (resp. positive) when the short-term implied volatility is lower (resp. greater) than the long-term implied volatility. As expected, Panels B and C of Figure 3.1 show that the long-term level and the slope strongly increase during the 2008 subprime crisis and the COVID-19 pandemic.

Table 3.2 presents summary statistics for fitted factor coefficients. The long-term level coefficient $\beta_{t,1}$ varies between 0.12 and 0.42 and displays a mean of 0.2, which is consistent with expectations for a long-term volatility level. The time-to-maturity slope $\beta_{t,2}$ ranges between -0.2 to 0.92, and its skewness is strongly positive at 2.66, indicating that it slightly

[9]Following François et al., 2022, the model horizon T_{max} is set to 5 years and T_{conv} to 0.25 to capture the fast convexity change in the IV term structure. $T_{\text{min}} = \frac{6}{365}$ corresponds to the smallest time-to-maturity in our sample. The parameters are estimated by means of least-square regressions with a Bayesian adjustment.

[10]Due to the fact that $\sigma_t(0,0,\beta_t) - \lim_{\tau\to\infty}\sigma_t(0,\tau,\beta_t) = \beta_{t,2}$.



On a two-scale graph, Panel A presents the S&P 500 log-returns (black line) and its estimated annualized volatility (grey line) obtained from the conditional variance (3.3) of Section's 3.4 dynamic model. The other panels display daily estimates of the factor coefficients (black line) as well as their associated estimated volatility (grey line) computed from Section 3.4's conditional variances (3.4) or (3.5).

Figure 3.1: S&P 500 daily returns, daily IV surface coefficients and their volatilities

	Min	Q1	Median	Q3	Max	Mean	Std	Skew	Kurt
Long-term level ($\beta_{t,1}$)	0.12	0.17	0.19	0.23	0.42	0.20	0.05	1.02	4.54
TmT Slope ($\beta_{t,2}$)	−0.20	−0.09	−0.05	0.00	0.92	−0.03	0.09	2.66	17.08
Moneyness Slope ($\beta_{t,3}$)	0.12	0.21	0.24	0.27	0.33	0.24	0.04	−0.38	2.87
Smile attenuation ($\beta_{t,4}$)	−0.06	−0.01	0.00	0.02	0.07	0.00	0.02	0.19	3.71
Smirk ($\beta_{t,5}$)	−0.08	−0.03	−0.02	−0.01	0.04	−0.02	0.02	−0.44	3.48
Short-term vol. ($\beta_{t,1} + \beta_{t,2}$)	0.02	0.09	0.15	0.21	1.21	0.17	0.11	2.40	13.14

Summary statistics of the factor coefficient estimates. The (ATM) long-term level is $\beta_{t,1}$. TmT slope is the term structure slope $\beta_{t,2}$, that is, the difference between the short-term and the long-term ATM implied volatilities. The moneyness slope corresponds to $\beta_{t,3}$ and the smile attenuation to $\beta_{t,4}$. The moneyness smirk for call options is captured by $\beta_{t,5}$. The last row (short-term volatility) is obtained by summing the long-term level factor with the time-to-maturity slope factor ($\beta_{t,1} + \beta_{t,2}$). It accounts for the interaction between $\beta_{t,1}$ and $\beta_{t,2}$.

Table 3.2: Summary statistics of the factor coefficients

decreases during periods of calm while it strongly increases during periods of turmoil. The three quartiles for the time-to-maturity slope are all negative, showing that the IV term structure is rarely increasing.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1) S&P 500 log-returns	1	**-0.54**	**-0.71**	**0.06**	**-0.21**	**-0.29**	**-0.77**
(2) Long-term level ($\Delta\beta_{t,1}$)	-0.62	1	**0.16**	**-0.06**	**0.24**	**0.15**	**0.29**
(3) TmT Slope ($\Delta\beta_{t,2}$)	-0.76	0.33	1	**-0.03**	**0.13**	**0.31**	**0.98**
(4) Moneyness Slope ($\Delta\beta_{t,3}$)	0.11	-0.06	-0.09	1	**0.27**	**0.13**	**-0.03**
(5) Smile attenuation ($\Delta\beta_{t,4}$)	-0.14	0.18	0.06	0.24	1	**-0.05**	**0.16**
(6) Smirk ($\Delta\beta_{t,5}$)	-0.24	0.15	0.33	0.11	-0.06	1	**0.33**
(7) Short-term vol. ($\Delta(\beta_{t,1} + \beta_{t,2})$)	-0.81	0.46	0.99	-0.09	0.08	0.33	1

Pearson (below the diagonal) and Spearman (bold numbers above the diagonal) correlations between the S&P 500 log-returns and the variations of the coefficients $\Delta\beta_{t,i} = \beta_{t,i} - \beta_{t-1,i}$ from January 4, 1996, to December 31, 2020.

Table 3.3: Correlation matrix of factor coefficient variations

Table 3.3 displays the sample correlation matrix applied to the S&P 500 log-returns and the factor coefficient estimate daily variations $\Delta\beta_t = \beta_{t,i} - \beta_{t-1,i}$, $i = 1, \ldots, 5$. The S&P 500 log-returns are strongly negatively correlated with variations of the long-term volatility coefficient $\Delta\beta_1$ and with those of the time-to-maturity slope $\Delta\beta_2$, but even more so with those of the short-term ATM implied volatility $\Delta(\beta_{t,1} + \beta_{t,2})$. This is a manifestation of the leverage effect generating higher short-term volatility that is associated with negative S&P 500 log-returns. Thus, negative index returns generally impact long-term volatility but not as much as short-term volatility. The almost perfect correlation between $\Delta\beta_2$ and $\Delta(\beta_1 + \beta_2)$ highlights the stability of long-term IVs.

3.4 The IV surface dynamics

Early attempts at directly modelling the IV dynamics (i.e. the “reverse approach" discussed in the introduction) consist in treating the implied volatilities as state variables (e.g. Schönbucher, 1999). But, as Carr and Wu, 2020 emphasized, assuming diffusion dynamics for IVs places strong restrictions on the drift and volatility coefficients to bar arbitrage, hence rendering the approach hardly tractable.

To circumvent this issue, we first rely on a factor representation of the IVs, which has been shown to leave little room for arbitrage (François et al., 2022). That factor representation, given by Equation (2.2), serves as a foundation for a dynamic representation of the IV surface.

Three ingredients are required for the model to be fully characterized: (i) the physical dynamics for the underlying S&P 500 log-returns with time-varying volatility, (ii) the physical dynamics for each of the five-factor coefficients, and (iii) a dependence structure between the underlying asset and the coefficients. The contribution of the three model components to the accurate modelling of IV surface dynamics is tested and validated in a robustness check (see Appendix B.1).

3.4.1 S&P 500 log-returns

In the spirit of Christoffersen et al., 2008's GARCH volatility component model, in which the conditional volatility is mean-reverting around a long-run component, the log-return dynamics follow an adaptation of an NGARCH(1,1)-NIG model where the variance is anchored in the 1-month ATM implied volatility,[11] thereby incorporating the forward-looking information of the option data. The excess log-return[12] $R_{t+1} = \log \frac{S_{t+1}}{S_t} - r_t + q_t$ satisfies

$$\begin{aligned} R_{t+1} &= \lambda h_{t+1,R}\Delta - \psi(\sqrt{h_{t+1,R}\Delta}) + \sqrt{h_{t+1,R}\Delta}\epsilon_{t+1,R}\,, \\ h_{t+1,R} &= V_t + \kappa_R\,(h_{t,R} - V_t) + a_R h_{t,R}(\epsilon_{t,R}^2 - 1 - 2\gamma_R\epsilon_{t,R}), \\ \sqrt{V_t} &= \omega_R\sigma\left(0, \frac{1}{12}, \beta_t\right), \end{aligned} \tag{3.3}$$

where $\Delta = \frac{1}{252}$ represents the daily time step.[13] The parameter λ reflects the equity risk premium. The sequence of innovations $\{\epsilon_{t,R}\}_{t=1}^T$ is constituted of independent standardized NIG random variables with two parameters ζ and ϕ which influence the skewness and the kurtosis of the distribution.[14] The convexity correction $\psi(\sqrt{h_{t+1,R}\Delta})$ is derived from the cumulant generating function ψ of the standardized NIG distribution, which is described

[11] As shown by Ledoit and Santa-Clara, 1998 and Yan, 2011, the very short end of the IV surface at the ATM point converges to the instantaneous volatility under the Equivalent Martingale Measure, i.e. ω_R should be very close to 1 in the absence of volatility risk premium. As shown in Christoffersen et al., 2008, the parameter ω_R should be different than 1. Note that in our framework, $\sigma\left(0, \frac{1}{12}, \beta_t\right) = \left(\beta_{t,1} + \beta_{t,2}\exp\left(-\sqrt{\frac{1}{12}\frac{1}{T_{\text{conv}}}}\right)\right)$.

[12] S denotes the index level, r is the daily risk-free rate and q stands for the daily dividend yield.

[13] The conditional variance dynamics can be rewritten as

$$h_{t+1,R} = (1-\kappa_R)V_t + \left(\kappa_R - a_R(\gamma_R^2+1)\right)h_t + a_R h_t\left(\epsilon_{t,R} - \gamma_R\right)^2.$$

Therefore, the conditional variance stays nonnegative if $V_t > 0$, $0 \le \kappa_R \le 1$ and $|\gamma_R| \le \sqrt{\frac{\kappa_R - a_R}{a_R}}$.

[14] Their expectation, variance, skewness and excess kurtosis are, respectively:

$$\mathrm{E}[\epsilon_{t,R}] = 0, \quad \mathrm{E}[\epsilon_{t,R}^2] = 1, \quad \mathrm{E}[\epsilon_{t,R}^3] = \frac{3\zeta}{\phi^2} \quad \text{and} \quad \mathrm{E}[\epsilon_{t,R}^4] - 3 = 3\Big(\frac{\phi^2 + 5\zeta^2}{\phi^4}\Big).$$

in Appendix B.5.[15]

According to Equation (3.3), the conditional annualized daily variance of the S&P 500 log-returns, $h_{t,R}$, exhibits mean-reverting behaviour around a fraction of the 1-month ATM squared implied volatility. The conditional variance noise term $\left(\epsilon_{t,R}^2 - 1 - 2\epsilon_{t,R}\gamma_R\right)$ is centered around 0. Results in Appendix B.1 substantiate the largely superior fitting performance of the standalone S&P 500 log-returns representation stemming from this novel characterization of the variance process over a NGARCH-NIG(1,1) model.

3.4.2 Factor coefficient dynamics

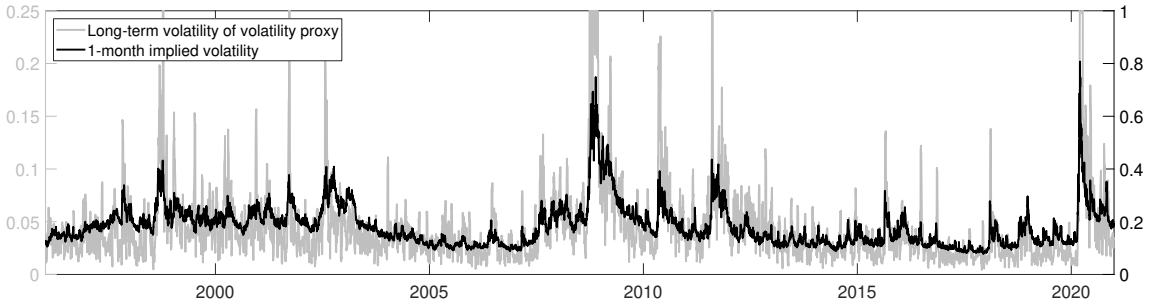
The second model component specifies the long-term ATM surface level dynamics for $\beta_{t,1}$. As in Carr and Wu, 2016, we assume that the volatility of the implied volatility (the volvol) is proportional to the implied volatility level. That assumption is substantiated by Figure 3.2, which reports the time series of the 1-month ATM implied volatility obtained from Equation (2.2) and a proxy of the volvol consisting of the sample standard deviation of $\Delta\beta_1$ computed with a 5-day rolling window. On a two-scale graph, Figure 3.2 highlights the similarity of the two time series.

In line with the evidence in Figure 3.2, we propose a volatility process for the long-term level ($\beta_{t,1}$) that is structurally similar to that of the underlying asset volatility process. The ATM long-term level of the IV surface ($\beta_{t,1}$) evolution is therefore modelled with an AR-NGARCH(1,1)-NIG model:

$$\begin{aligned}
\beta_{t+1,1} &= \alpha_1 + \sum_{j=1}^{5} \theta_{1,j}\beta_{t,j} + \sqrt{h_{t+1,1}\Delta}\epsilon_{t+1,1}, \\
h_{t+1,1} &= U_t + \kappa_1\left(h_{t,1} - U_t\right) + a_1 h_{t,1}(\epsilon_{t,1}^2 - 1 - 2\epsilon_{t,1}\gamma_1), \\
\sqrt{U_t} &= \omega_1\sigma\left(0, \frac{1}{12}, \beta_t\right),
\end{aligned} \tag{3.4}$$

[15] For $-\sqrt{\zeta^2+\phi^2} - \zeta < z < \sqrt{\zeta^2+\phi^2} - \zeta$, the cumulant generating function is given by

$$\psi(z) = \frac{\phi^2}{\phi^2+\zeta^2}\Big(-\zeta z + \phi^2 - \phi\sqrt{\phi^2+\zeta^2-(\zeta+z)^2}\Big).$$



The long-term volatility of volatility proxy corresponds to the sample standard deviation of $\Delta\beta_1$ over the last five days, annualized (grey line, left scale). The 1-month implied volatility is $\sigma\left(0, \frac{1}{12}, \beta_t\right) = \left(\beta_{t,1} + \beta_{t,2} \exp\left(-\sqrt{\frac{1}{12}\frac{1}{T_{\text{conv}}}}\right)\right)$ (black line, right scale).

Figure 3.2: Comparing the 1-month ATM IV to the long-term IV volatility proxy

where the variance process $\{h_{t,1}\}$ exhibits mean-reversion around a fraction ω_1 of the 1-month ATM implied volatility.

The evolution of the four other daily coefficients of Equation (2.2) is represented by an AR-NGARCH(1,1)-NIG process. For $i = 2, 3, 4, 5$,

$$\begin{aligned} \beta_{t+1,i} &= \alpha_i + \sum_{j=1}^{5} \theta_{i,j}\beta_{t,j} + \nu\beta_{t-1,2}\mathbb{1}_{\{i=2\}} + \sqrt{h_{t+1,i}\Delta}\epsilon_{t+1,i} \\ h_{t+1,i} &= \sigma_i^2 + \kappa_i\left(h_{t,i} - \sigma_i^2\right) + a_i h_{t,i}\left(\epsilon_{t,i}^2 - 1 - 2\epsilon_{t,i}\gamma_i\right). \end{aligned} \tag{3.5}$$

A second-order lag for the time-to-maturity slope coefficient is included in the specification to capture the auto-correlation present in its level and its variations. The IV surface coefficients exhibit strong autocorrelation, parameters $\theta_{i,i}$, $i = 1, \ldots, 5$ are expected to be close to 1.

3.4.3 Dependence structure

Specifying a dependence structure completes the modelling framework. A Gaussian copula captures the dependence among the NIG innovations $(\epsilon_{t,R}, \epsilon_{t,1}, ..., \epsilon_{t,5})$. Interactions between the IV surface coefficients are captured both through auto-regressive parameters $\theta_{i,j}$, $i \neq j$, and through the dependence between the innovations $\epsilon_{t,i}$, $i \in \{R, 1, \ldots, 5\}$.

Equations (3.3)-(3.5) coupled with the dependence structure of the Gaussian copula comprehensively describe the dynamics of the joint implied volatility and return (JIVR) model.

3.5 Estimation

The parameters from the dynamic model presented in Section 3.4 are estimated through a two-step approach. In the first step, parameters of the marginal processes $\{R_t\}, \{\beta_{t,1}\}, \ldots, \{\beta_{t,5}\}$ are estimated separately by maximum likelihood.[16] In the second step, parameters of the Gaussian copula are estimated from the model residuals obtained in the first step.[17]

	P-value
S&P500 log-returns	30.0%
Long-term level ($\beta_{t,1}$)	63.9%
TmT Slope ($\beta_{t,2}$)	78.6%
Moneyness Slope ($\beta_{t,3}$)	60.8%
Smile attenuation ($\beta_{t,4}$)	65.5%
Smirk ($\beta_{t,5}$)	30.7%

The table presents p-values of the Cramér-von Mises test applied to residuals of the AR-NGARCH-NIG models displayed in Equations (3.3)-(3.5) over the whole period ranging from January 4, 1996, to December 31, 2020.

Table 3.4: Cramér-von Mises goodness-of-fit tests

To test the statistical adequacy of the model, Cramér-von Mises tests are applied to the residuals of each marginal process. The null hypothesis is that the residuals have a NIG

[16]The backward parameter selection algorithm with the Bayesian Information Criterion (BIC) is implemented. Such an iterative procedure is detailed, for instance, in James et al., 2013.

[17]The Gaussian copula is estimated by converting the residuals whose marginals are approximately NIG into pseudo-residuals with approximately standard Gaussian marginals. This is done through the successive application of the NIG cdf and the Gaussian inverse cumulative distribution function to original residuals. A correlation matrix is computed from the set of residuals with Gaussian marginals, which corresponds to the Gaussian copula parameters.

distribution. Table 3.4 presents the p-values of the tests, with the null hypothesis never being rejected.

Table 3.5 displays the estimated model parameters. Regarding the S&P 500 dynamics (3.3), the constant ω linking the 1-month implied ATM volatility to the physical instant volatility is estimated at around 0.98, implying that the physical volatility factor is, on average, smaller than the 1-month implied ATM volatility. The negative skewness and the positive excess kurtosis of the NIG innovation distributions indicate the presence of extreme return movements. As a result, the speed of reversion of the S&P 500 conditional variance ($\kappa_R = 89\%$) is not as close to 1 as it would have been if noises were Gaussian. Indeed, a smaller persistence makes scenarios with prolonged extreme volatility due to a single large S&P 500 return less likely. As expected, the asymmetry parameter γ_R is positive, implying that the S&P 500 variance reacts more strongly to negative return shocks than to positive shocks.

The grey lines in Figure 3.1 display the time series of the estimated S&P 500 log-returns annualized volatility ($\sqrt{h_R}$) as well as the time series of the factor coefficients annualized volatility ($\sqrt{h_i}$ for $i = 1, 2, 3, 4, 5$). The volatility time series of the log-returns and the long-term level factor coefficient closely follow the IV surface level, which is consistent with the specification of the respective variance processes. As expected, the volatility sharply increases during periods of market turmoil, such as the 2008 financial crisis or during the COVID pandemic, and is relatively low during periods where the market is calm. Interestingly, the time-to-maturity slope volatility closely follows the time-to-maturity slope level itself. The volatility time series of the other three factor coefficients do not exhibit any clear pattern related to the underlying log-returns volatility or the IV surface level.

For all five factor coefficients (β_i) displayed in Table 3.5, large values for the auto-regressive parameter $\theta_{i,i}$ imply strong persistence in their dynamics. Moreover, the speed of reversion parameters κ_i are quite high, indicating that the volatilities of implied volatility coefficients are also persistent. For the long-term implied volatility level β_1, the asymmetry parameter γ_1 is negative, which is expected because a positive shock on the long-term implied volatility is a sign of market uncertainty and increases the variability of the volatility

surface. Because β_2 represents the difference between the short- and the long-term implied volatility levels, β_2 increases during financial turmoil. Again, positive shocks on β_2 have a larger impact on the variability of the implied volatility surface than negative ones, resulting in a negative, yet non-statistically significant asymmetry parameter γ_2.

	β_1	β_2	β_3	β_4	β_5		S&P500
α	0.0009^*	0.0084^*	0.0008^*	-0.0014^*	0.0007^*	λ	2.7324^*
	(0.0002)	(0.0009)	(0.0003)	(0.0003)	(0.0002)		(0.0002)
θ_1	0.9963^*	-0.0139^*		0.0028^*			
	(0.0009)	(0.0030)		(0.0008)			
θ_2	0.0037^*	0.8778^*	0.0013^*				
	(0.0005)	(0.0119)	(0.0006)				
θ_3		-0.0326^*	0.9971^*	0.0037^*	-0.0042^*		
		(0.0039)	(0.0011)	(0.0011)	(0.0009)		
θ_4				0.9803^*			
				(0.0028)			
θ_5		-0.0478^*			0.9860^*		
		(0.0073)			(0.0023)		
ν		0.0894^*					
		(0.0121)					
$\sigma\sqrt{252}$		0.3803^*	0.0522^*	0.0486^*	0.0515^*		
ω	0.2676^*						0.9774^*
	(0.0064)						(0.0009)
κ	0.8382^*	0.9658^*	0.9743^*	0.9454^*	0.9808^*		0.8891^*
	(0.0279)	(0.0032)	(0.0054)	(0.0110)	(0.0041)		(0.0100)
a	0.1342^*	0.0983^*	0.0926^*	0.1022^*	0.1005^*		0.0561^*
	(0.0150)	(0.0072)	(0.0101)	(0.0113)	(0.0100)		(0.0041)
γ	-0.1118^*	-2.9657	0.1935	0.1211	-0.2060		2.5064^*
	(0.0081)	(2.2242)	(0.1068)	(0.1248)	(0.1125)		(0.1125)
β_{NIG}	0.1438^*	0.8529^*	0.0291^*	-0.1591^*	0.0927^*		-0.6412^*
	(0.0375)	(0.0008)	(0.0007)	(0.0004)	(0.0004)		(0.0004)
γ_{NIG}	1.3511^*	1.5389^*	2.2848^*	1.4500^*	1.4285^*		2.0398^*
	(0.0717)	(0.0828)	(0.1880)	(0.0766)	(0.0719)		(0.0719)
Skew	0.24	1.08	0.02	-0.23	0.14		-0.4623
Ex. Kurt	1.74	3.21	0.58	1.51	1.50		1.0772
Log. Lkhd.	$-28,314$	$-16,940$	$-27,532$	$-28,040$	$-27,874$		$-20,673$
Log. Lkhd. All	$-28,322$	$-16,897$	$-27,535$	$-28,042$	$-27,879$		

JIVR model parameters estimated over the whole daily sample ranging from January 4, 1996, to December 31, 2020. The standard errors are displayed under the estimates in parentheses. The model is regularized using a backward selection method with the BIC criterion. The log-likelihoods (Log. Lkhd.) and the log-likelihood of the model where no parameter from the θ matrix is set to 0 (Log. Lkhd. All) are reported. The skewness (Skew) and excess kurtosis (Ex. Kurt) of the NIG distributions for residuals are shown for all coefficients and the S&P 500 log-returns. Parameters with a star superscript ($*$) are significantly different from 0 at the 5% confidence level.

	(1)	(2)	(3)	(4)	(5)	(6)
(1) S&P 500 log-returns (R_t)	1.00					
(2) Long-term level ($\beta_{t,1}$)	-0.55	1.00				
(3) TmT Slope ($\beta_{t,2}$)	-0.69	0.13	1.00			
(4) Moneyness Slope ($\beta_{t,3}$)	0.03	-0.03	-0.01	1.00		
(5) Smile attenuation ($\beta_{t,4}$)	-0.22	0.25	0.12	0.28	1.00	
(6) Smirk ($\beta_{t,5}$)	-0.34	0.17	0.37	0.13	-0.05	1.00

Estimated Gaussian copula parameters for the innovations of the JIVR model $\{\epsilon_{t,R}, \epsilon_{t,1}, \ldots, \epsilon_{t,5}\}$. The Gaussian copula is estimated on the residuals extracted from estimated models illustrated in Equations (3.3)-(3.5) over the whole sample ranging from January 4, 1996, to December 31, 2020.

Table 3.6: Gaussian copula

Table 3.6 presents estimates for the Gaussian copula parameter matrix. Results indicate that the log-return innovations are strongly negatively associated with shocks on the first two coefficients, i.e., the long-term level and time-to-maturity slope.

3.6 Risk management applications

We test the ability of the JIVR model to accurately estimate risk metrics for option portfolios on real data. In a first set of numerical experiments, we examine standard positions in volatility management. Then, we compare VIX index forecasts produced by the JIVR model to those provided by an approach directly modelling the VIX time series. The goal is to assess how well the model does in processing the information from remote areas of the IV surface to capture higher moments and tail behaviour.

In the applications presented below, backtesting and forecasting procedures rely on an expanding window methodology. More precisely, the total sample period is divided by year. For each iteration N, where $N = 2007, \ldots, 2020$, the model is first estimated over the training sample, which covers years 1996 to $N - 1$. Then, for each day t in year N, multiple d-day-ahead IV surface predictions are generated using the estimated model, the

latter being denoted by m_N. Such predictions are then used to calculate daily outcomes. Algorithm 1 summarizes the procedure.

Algorithm 1 The expanding window

for $N = 2007 : 2020$ **do** ▷ 2007 to 2020 is the out-of-sample period.
  Compute m_N, the estimated model over the training period 1996 to $N - 1$.
  for $t = 1 : D_N$ **do** ▷ D_N is the number of trading days in year N
    for $i = 1 : s$ **do** ▷ s is the number of simulations
      Simulate $\tilde{\beta}_{t+d}^{\{i\}}$ and $\tilde{Y}_{t+d}^{\{i\}}$ from model m_N using the information set at time t.
    end for
  end for
end for

3.6.1 Straddle and strangle positions

Straddles and strangles are standard option strategies that can be used to take positions in the underlying asset volatility. To evaluate the risk of such strategies, we consider the Value-at-Risk (VaR), a popular risk metric used by practitioners. The accuracy of VaR estimates is evaluated through a standard backtesting procedure. VaR estimates are produced for the six strategies being considered (1-month, 3-month and 6-month straddles and strangles) over two possible time horizons, namely, one and five trading days.

Daily out-of-sample d-day-ahead VaR estimates for various confidence levels over the years 2007 to 2020 are produced according to the expanding window approach.[18] To forecast the strategy return distribution for a horizon of d days, i.e. a return between times t and $t + d$, one must first compute the current price of the strategy at time t using the fitted

[18]The risk-free rate and the dividend yield are kept constant during the simulations over one and five trading days. Under these assumptions, the forward price is simulated as follows:

$$\tilde{F}_{t+d,\tau-d}^{\{i\}} = F_{t,\tau} e^{\sum_{u=1}^{d} \tilde{R}_{t+u}^{\{i\}}}$$

where $F_{t,\tau}$ is the forward price at time t with maturity τ and $\tilde{R}_{t+u}^{\{i\}}$ is the simulated daily excess log-return for path i.

IV surface. The return $\frac{V_{t+d}^{\{i\}}-V_t}{V_t}$ is then computed for each simulated[19] scenario i, where V_t denotes the time-t strategy value. The return (not price) VaR is estimated to allow for comparisons across straddles and strangles.

VaR coverage tests are conducted to assess the backtesting performance (Kupiec et al., 1995). A VaR breach occurs when an observation falls below (above) the return quantile associated with the 1% or 5% (95% or 99%) confidence level, addressing potential losses for investors taking long (short) positions. The VaR coverage test is a likelihood ratio test that determines if the proportion of realized VaR breaches is significantly different from the VaR confidence level. More details are provided in Appendix B.3.

Table 3.7 exhibits the proportion of observed daily VaR breaches over the out-of-sample period (from 2007 to 2020) for each confidence level. VaR coverage test p-values are disclosed in parentheses. Panel A and B exhibit results for the 1-day-ahead and the 5-day-ahead forecast horizons, respectively. The counts of VaR breaches are close to their expected theoretical value, and therefore the results are economically conclusive. It is also possible to perform tests to verify whether the differences are statistically significant. The p-values are reported in parenthesis in the table. Overall, results of the VaR coverage tests provide comfort about the ability of the developed methodology to produce reliable VaR estimates.

3.6.2 Forecasting the VIX index distribution

The VIX index has high practical importance since it encompasses information related to market perceptions about the future volatility of the S&P 500 index over a 30-day horizon.

The Chicago Board Options Exchange (CBOE) computes the VIX index ($\text{VIX}_t^{\text{index}}$) from a portfolio of available put and call options as explained in detail in CBOE, 2014. For a given time-to-maturity τ,

$$\text{VIX}_{t,\tau} = 100\sqrt{\left(\frac{2}{\tau}\sum_{i=1}^{N_{t,\tau}}\frac{\Delta K_{i,\tau}}{K_{i,\tau}^2}e^{r_\tau \tau}P_{t,\tau}(K_{i,\tau}) - \frac{1}{\tau}\left(\frac{F_{t,\tau}}{K_{j,\tau}}-1\right)^2\right)} \tag{3.6}$$

[19]The Monte Carlo simulation is based on 75,000 paths.

where $K_{1,\tau} < ... < K_{j,\tau} \leq F_{t,\tau} < K_{j+1,\tau} < ... < K_{N_t,\tau}$ are the strike prices of quoted options with maturity τ, $P_{t,\tau}(K_i)$ is the time t out-of-the-money option price.[20] The strike price variations are $\Delta K_{1,\tau} = K_{2,\tau} - K_{1,\tau}$, $\Delta K_{i,\tau} = \frac{1}{2}(K_{i+1,\tau} - K_{i-1,\tau})$ for $1 < i < N_{t,\tau}$ and $\Delta K_{N,\tau} = K_{N_{t,\tau},\tau} - K_{N_{t,\tau}-1,\tau}$. The $(\text{VIX}^{\text{index}}_{t,T})^2$ of maturity $T = 30$ days is a linear interpolation between VIX^2_{t,τ_1} and VIX^2_{t,τ_2}, where $\tau_1 \leq T \leq \tau_2$ are the two nearest available time-to-maturities surrounding T:

$$\text{VIX}^{\text{index}}_{t,T} = \sqrt{\frac{\tau_2 - T}{\tau_2 - \tau_1}\text{VIX}^2_{t,\tau_1} + \frac{T - \tau_1}{\tau_2 - \tau_1}\text{VIX}^2_{t,\tau_2}}. \tag{3.7}$$

We use the JIVR model to generate forecasts for the VIX variation

$$\Delta\text{VIX}^{\text{JIVR}}_{t,t+d} = \text{VIX}^{\text{JIVR}}_{t+d} - \text{VIX}^{\text{JIVR}}_{t}$$

over a prediction horizon of d days. More precisely, $\text{VIX}^{\text{JIVR}}_{t,\tau}$ is obtained from Equation (3.7) by replacing the OTM option quoted prices by the ones obtained from the fitted IV surface that day. This step requires the identification of the moneyness levels $M_1, ..., M_{N_{t,\tau}}$ corresponding to the available strike prices $K_1, ..., K_{N_{t,\tau}}$. From Monte Carlo simulations, the JIVR model generates scenarios for the IV surface and the underlying asset log-return for horizon of d days. Using the same available moneyness levels and maturities as for time t, the VIX forecast $\text{VIX}^{\text{JIVR}}_{t+d}$ is computed from the corresponding option prices on the predicted IV surfaces. Lastly, the $\text{VIX}^{\text{index}}_{t+d,T}$ forecast is

$$\text{VIX}^{\text{index}}_{t+d,T} = \text{VIX}^{\text{index}}_{t,T} + \Delta\text{VIX}^{\text{JIVR}}_{t,t+d}.$$

This approach is compared to a direct modelling of the VIX index time series through:

$$\text{VIX}_t = \alpha_{\text{VIX}} + \beta_{\text{VIX}}\text{VIX}_{t-1} + \sqrt{h_{t,\text{VIX}}\Delta}\,\epsilon_{t,\text{VIX}} \tag{3.8}$$

$$h^{\text{VIX}}_{t+1} = \sigma^2_{\text{VIX}} + \kappa\left(h^{\text{VIX}}_t - \sigma^2_{\text{VIX}}\right) + a h^{\text{VIX}}_t(\epsilon^2_{t,\text{VIX}} - 1 - 2\epsilon_{t,\text{VIX}}\gamma) \tag{3.9}$$

where the variance process $\{h^{\text{VIX}}_{t+1}\}$ exhibits mean-reversion around a fixed parameter σ^2_{VIX}. The sequence of innovations $\{\epsilon_{t,\text{VIX}}\}_{t=1}^T$ is constituted of independent standardized NIG random variables.

[20] These prices are obtained from the bid-ask spread mid-points. They correspond to a put option for $K_{i,\tau} \leq F_{t,\tau}$, and a call option for $K_{i,\tau} > F_{t,\tau}$.

To compare the log-likelihoods of both models, we introduce the average likelihood ratio (ALR). The ALR is defined as the geometric average of the ratio of likelihood scores for observations of one model over those of another,

$$\begin{aligned} \text{ALR} &= \left(\frac{\mathcal{L}^{(1)}(O_{1:T}|\theta^{(1)})}{\mathcal{L}^{(2)}(O_{1:T}|\theta^{(2)})}\right)^{\frac{1}{T}}. \\ &= \exp\left(\frac{1}{T}\left(\sum_{t=1}^{T}\log\mathcal{L}_t^{(1)}\left(O_t|O_{1:t-1},\theta^{(1)}\right)-\log\mathcal{L}_t^{(2)}\left(O_t|O_{1:t-1},\theta^{(2)}\right)\right)\right) \end{aligned} \tag{3.10}$$

where T is the total number of days in the set of out-of-sample folds, $\mathcal{L}^{(1)}$ and $\mathcal{L}^{(2)}$ are the likelihoods of the first and second models respectively, $O_{1:T}$ is the time series of out-of-sample observations including all out-of-sample folds, and $\theta^{(1)}$ and $\theta^{(2)}$ represent the model parameter sets considered when computing $\mathcal{L}^{(1)}$ and $\mathcal{L}^{(2)}$. In our case, model 1 corresponds to the JIVR model, while model 2 corresponds to the direct model. The ALR indicates how much more or less likely an observation is, on average, in one model versus the other in relative terms.

Out-of-sample yearly log-likelihoods are computed using the expanding window methodology described by Algorithm 1.[21] However, the out-of-sample period is reduced to between 2014 and 2020, which corresponds to the period where reported VIX values are computed with the most recent calculation method published in CBOE, 2014.

[21]The log-likelihood cannot be computed directly from the simulated VIX distribution generated by the JIVR model. To circumvent this issue, a kernel density estimate (`ksdensity` function from the `MATLAB` software with the default bandwidth) is applied to the simulated VIX values to obtain a density estimate. The latter is then used to compute the log-likelihood for the JIVR model.

	JIVR	Direct approach	ALR	P-values
2014	**-343.80**	-352.33	1.03	11.0%
2015	-446.86	**-430.55**	0.94	96.5%
2016	-385.73	**-379.62**	0.98	76.9%
2017	**-261.53**	-269.93	1.03	19.0%
2018	**-427.71**	-445.06	1.07	2.6%
2019	**-359.82**	-362.01	1.01	24.0%
2020	**-517.91**	-537.27	1.08	1.6%
Total	**-2,743.36**	-2,776.77	1.02	8.2%

Log-likelihoods for each of the out-of-sample years as well as for the aggregated out-of-sample period (Total). The log-likelihood for each year is computed using both the JIVR model and the direct model. The respective parameters of both models are estimated over previous years' observations. Bold numbers highlight which of the two models outperforms the other, either for a specific year or in aggregate. The last column of the table (p-values) corresponds to the Diebold and Mariano, 1995 test p-values. The Diebold and Mariano, 1995 test, described in Appendix B.4, verifies if the predictive accuracy of two models is equal for a specified performance metric (log-likelihood). If the null hypothesis of the test is rejected, then one model statistically outperforms the other.

Table 3.8: Out-of-sample performance for VIX distribution forecasting

Table 3.8 exhibits the log-likelihood, the ALR, and the p-values from the Diebold and Mariano, 1995 test computed for each out-of-sample year as well as for the whole out-of-sample period. The null hypothesis of Diebold and Mariano, 1995 test, described in Appendix B.4, assumes that the predictive accuracy of the models is statistically equal. For a significance level of 5%, a p-value under 2.5% (resp. over 97.5%) indicates that the JIVR model (resp. direct model) significantly outperforms the direct model (JIVR model). Results show that the JIVR model largely outperforms the direct model, with ARL above 1 for five out of the seven out-of-sample years and for the entire aggregated period (2014–2020). The Diebold and Mariano, 1995 test p-values reveal that the JIVR model statistically outperforms the direct model for two of the seven out-of-sample years.

		Straddles			Strangles		
	TmT in months	1	3	6	1	3	6
	Panel A: 1 day						
1%	% of VaR breaches	1.02%	0.77%	0.79%	1.02%	0.60%	0.57%
	p-values	(89.93%)	(14.52%)	(20.30%)	(89.93%)	(0.91%)	(0.50%)
5%	% of VaR breaches	4.06%	3.32%	3.35%	5.19%	3.74%	3.69%
	p-values	(0.80%)	(0%)	(0%)	(60.41%)	(0.04%)	(0.02%)
95%	% of VaR breaches	5.82%	4.68%	4.82%	6.18%	5.22%	4.99%
	p-values	(3.01%)	(37.97%)	(62.71%)	(0.18%)	(55.19%)	(98.46%)
99%	% of VaR breaches	1.79%	1.22%	1.50%	1.76%	1.48%	1.50%
	p-values	(0%)	(20.47%)	(0.52%)	(0%)	(0.81%)	(0.52%)
	Panel B: 5 days						
1%	% of VaR breaches	0.57%	1.33%	1.50%	0.96%	0.94%	1.28%
	p-values	(0.50%)	(5.84%)	(0.52%)	(83.14%)	(70.02%)	(11.35%)
5%	% of VaR breaches	3.86%	5.19%	5.36%	4.11%	5.59%	5.93%
	p-values	(0.12%)	(60.41%)	(32.99%)	(1.28%)	(11.52%)	(1.38%)
95%	% of VaR breaches	5.16%	4.57%	5.19%	4.99%	4.62%	5.05%
	p-values	(65.84%)	(23.20%)	(60.41%)	(98.46%)	(30%)	(89.26%)
99%	% of VaR breaches	1.28%	1.13%	1.65%	1.36%	1.30%	1.56%
	p-values	(11.35%)	(43.12%)	(0.04%)	(4.08%)	(8.22%)	(0.20%)

For strangles, the moneyness of the call option is $M = -0.1$ and that of the put option $M = 0.1$. For straddles, both options are at-the-money. Time-to-maturity (TmT) is in months. The rows (1%, 5%, 95%, and 99%) represent the VaR confidence levels. The distribution forecast horizon is either 1-day- or 5-day-ahead. The backtest period extends from January 2, 2007, to December 31, 2020. The VaR coverage test is described in Section B.3 in which the number of VaR estimates is $N \approx 14 \times 250 = 3500$. Values in parentheses represent the p-values of the tests.

Table 3.7: VaR coverage test for straddles and strangles

3.7 Conclusion

This study develops the JIVR model, a characterization of the joint dynamics of the S&P 500 index and of its associated implied volatility surface. The approach is reminiscent of the dynamic Nelson-Siegel model of Diebold and Li, 2006 or the Heath-Jarrow-Morton framework of Heath et al., 1992; the current implied volatility surface is used as an input to the model, thereby greatly enhancing the ability of the approach to depict current market conditions accurately. The parametric model of François et al., 2022 is leveraged to decompose the implied volatility surface into contributions from five economically interpretable factors. The parametric model has been shown to capture well the implied volatility shape while supporting extrapolation beyond observable areas of the surface and leaving minimal room for arbitrage when applied to real data.

The JIVR model relies on joint NGARCH-type dynamics with fat-tailed and asymmetric NIG innovations to represent the evolution of the five IV surface factors and of the S&P 500 log-returns. The NGARCH processes for log-returns and the long-term level of the IV surface both include two variance components and are anchored in respective proportions of the one-month ATM IV. This novel characterization effortlessly integrates information from the IV surface into the variance dynamics of the S&P 500 log-returns and of the long-term IV factor. Such model specification proves consistent with the Carr and Wu, 2016 postulate expressing a proportionality relationship between the implied volatility level and the volvol. All other IV factors are represented by regular NGARCH-NIG processes. The dependence structure between innovations of all factors and of the underlying return is captured by a Gaussian copula.

The estimation of the model is performed through least-squares and conventional maximum likelihood procedures and is quite seamless. Cramér-von Mises goodness-of-fit tests applied to the residuals illustrate the appropriateness of the model specification. The JIVR model is shown to provide a vastly superior fit to observations than conventional GARCH processes estimated independently on each of the factors and on the underlying returns. In particular, the marginal specification of the S&P 500 returns process extracted from the JIVR model exhibits significantly higher performance than a standalone GARCH process, thereby highlighting the strong value added of information borrowed from the IV surfaces

when forecasting the S&P 500 return distribution.

Two exercises are conducted to demonstrate the capability of the JIVR model to accurately generate scenarios for the future implied volatility surface and the underlying return. The first exercise consists in assessing the risk of straddle and strangle positions with various times-to-maturity and one-day or five-day forecasting horizons. VaR coverage tests following Kupiec et al., 1995 confirm the overall accuracy of VaR estimates provided by the model.

The second exercise compares the predictive performance of the JIVR model and of a conventional time series counterpart model to forecast the VIX index distribution. Diebold-Mariano tests applied on out-of-sample likelihood scores confirm that the outperformance provided by the JIVR model is statistically significant.

Early attempts at jointly modelling the underlying asset return and the corresponding IV surface dynamics have been facing implementation challenges that were initially deemed unsurmountable. The JIVR model proposed here shows that effective S&P 500 index options risk management is feasible when asset returns with non-Gaussian NGARCH processes and two-factor volatility are combined with a robust, parametric IV surface specification. More importantly, our contribution suggests that this stream of literature deserves further investigation, as it may have much more applicability than originally thought. Further tests involving option pricing and option replication, potentially applied to other types of underlying assets, will help determine the exact potential of this approach.

References

Andersen, T. G., Fusari, N., & Todorov, V. (2015). The risk premia embedded in index options. *Journal of Financial Economics*, *117*(3), 558–584.

Azzone, M., & Baviera, R. (2022). Additive normal tempered stable processes for equity derivatives and power-law scaling. *Quantitative Finance*, *22*(3), 501–518.

Bakshi, G., Cao, C., & Chen, Z. (1997). Empirical performance of alternative option pricing models. *The Journal of Finance*, *52*(5), 2003–2049.

Barndorff-Nielsen, O. E. (1998). Processes of normal inverse Gaussian type. *Finance and Stochastics*, *2*(1), 41–68.

Bates, D. S. (1996). Jumps and stochastic volatility: Exchange rate processes implicit in deutsche mark options. *The Review of Financial Studies*, *9*(1), 69–107.

Bates, D. S. (2000). Post-'87 crash fears in the S&P 500 futures option market. *Journal of Econometrics*, *94*(1-2), 181–238.

Bates, D. S. (2022). Empirical option pricing models. *Annual Review of Financial Economics*, *14*, 369–389.

Black & Scholes. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, *81*(3), 637–654.

Bollerslev, T., & Todorov, V. (2011). Tails, fears, and risk premia. *The Journal of Finance*, *66*(6), 2165–2211.

Broadie, M., Chernov, M., & Johannes, M. (2007). Specification and risk premiums: The information in S&P 500 futures options. *Journal of Finance*, *62*(3), 1453–1490.

Carr, P., & Wu, L. (2004). Time-changed Lévy processes and option pricing. *Journal of Financial Economics*, *71*(1), 113–141.

Carr, P., & Wu, L. (2016). Analyzing volatility risk and risk premium in option contracts: A new theory. *Journal of Financial Economics*, *120*(1), 1–20.

Carr, P., & Wu, L. (2020). Option profit and loss attribution and pricing: A new framework. *The Journal of Finance*, *75*(4), 2271–2316.

CBOE. (2014). The CBOE volatility index-VIX. *White Paper*.

Christoffersen, P., Elkamhi, R., Feunou, B., & Jacobs, K. (2010). Option valuation with conditional heteroskedasticity and nonnormality. *The Review of Financial Studies*, *23*(5), 2139–2183.

Christoffersen, P., Jacobs, K., & Ornthanalai, C. (2013). GARCH option valuation: Theory and evidence. *The Journal of Derivatives*, *21*(2), 8–41.

Christoffersen, P., Jacobs, K., Ornthanalai, C., & Wang, Y. (2008). Option valuation with long-run and short-run volatility components. *Journal of Financial Economics*, *90*(3), 272–297.

Cont, R., & Da Fonseca, J. (2002). Dynamics of implied volatility surfaces. *Quantitative Finance*, *2*(1), 45–60.

Daglish, T., Hull, J., & Suo, W. (2007). Volatility surfaces: Theory, rules of thumb, and empirical evidence. *Quantitative Finance*, *7*(5), 507–524.

Diebold, F., & Mariano, R. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–63.

Diebold, F. X., & Li, C. (2006). Forecasting the term structure of government bond yields. *Journal of Econometrics*, *130*(2), 337–364.

Duan, J.-C. (1995). The GARCH option pricing model. *Mathematical Finance*, *5*(1), 13–32.

Duffie, D., Pan, J., & Singleton, K. (2000). Transform analysis and asset pricing for affine jump-diffusions. *Econometrica*, *68*(6), 1343–1376.

Fengler, M. R. (2006). *Semiparametric modeling of implied volatility*. Springer Science & Business Media.

François, P., Galarneau-Vincent, R., Gauthier, G., & Godin, F. (2022). Venturing into uncharted territory: An extensible implied volatility surface model. *Journal of Futures Markets*, *42*(10), 1912–1940.

Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, *48*(5), 1779–1801.

Heath, D., Jarrow, R., & Morton, A. (1992). Bond pricing and the term structure of interest rates: A new methodology for contingent claims valuation. *Econometrica: Journal of the Econometric Society*, 77–105.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The review of financial studies*, *6*(2), 327–343.

Heston, S. L., & Nandi, S. (2000). A closed-form GARCH option valuation model. *The Review of Financial Studies*, *13*(3), 585–625.

Hull, J., & White, A. (1987). The pricing of options on assets with stochastic volatilities. *The Journal of Finance*, *42*(2), 281–300.

Israelov, R., & Kelly, B. T. (2017). Forecasting the distribution of option returns. *Available at SSRN 3033242*.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning : With applications in R* (Vol. 112). Springer. https : / / doi . org / https : //doi.org/10.1007/978-1-4614-7138-7

Kupiec, P. H., et al. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, *3*(2), 73–84.

Ledoit, O., & Santa-Clara, P. (1998). Relative pricing of options with stochastic volatility. *University of California-Los Angeles finance working paper*, 9–98.

Merton, R. C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, *3*(1-2), 125–144.

Oh, D. H., & Park, Y.-H. (2022). GARCH option pricing with volatility derivatives. *Journal of Banking & Finance*, 106718.

Schönbucher, P. J. (1999). A market model for stochastic implied volatility. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, *357*(1758), 2071–2092.

Yan, S. (2011). Jump risk, stock returns, and slope of implied volatility smile. *Journal of Financial Economics*, *99*(1), 216–233.

Zhu, Y., & Avellaneda, M. (1998). A risk-neutral stochastic volatility model. *International Journal of Theoretical and Applied Finance*, *1*(2), 289–310.

Chapter 4

Foreseeing the worst: Forecasting electricity DART spikes

Abstract*

Statistical learning models are proposed for the prediction of the probability of a spike in the electricity DART (day-ahead minus real-time price) spread. Assessing the likelihood of DART spikes is of paramount importance for virtual bidders, among others. The model's performance is evaluated on historical data for the Long Island zone of the New York Independent System Operator (NYISO). A tailored feature set encompassing novel engineered features is designed. Such a set of features makes it possible to achieve excellent predictive performance and discriminatory power. Results are shown to be robust to the choice of the predictive algorithm. Lastly, the benefits of forecasting the spikes are illustrated through a trading exercise, confirming that trading strategies employing the model-predicted probabilities as a signal generate consistent profits.

Keywords: Power markets, Spikes prediction, DART spreads, NYISO, Predictive analytics, Statistical learning.

*Joint work with Geneviève Gauthier and Frédéric Godin. Gauthier is affiliated with HEC Montréal and Godin is affiliated with Concordia university.

4.1 Introduction

Electricity generation and consumption must happen simultaneously, which makes the electricity market particularly volatile. The slow reaction time of large producers of inexpensive electricity, combined with bottlenecks and failures in the transmission grid, or sudden increases (decreases) in demand, give rise to a phenomenon known as price spikes, where electricity trades at extremely high (or low negative) prices.

The New York Independent System Operator (NYISO) administers electricity flow operations for a large area in New York State. Electricity transactions are performed through two main markets: the day-ahead (DA) market and the real-time (RT) market. The DA market allows for the scheduling of power production and consumption one day in advance and contains the bulk of traded electricity volumes. Conversely, the RT market acts as a balancing market, correcting for the real-time departure of electricity volumes previously booked in the DA market. The NYISO is responsible for calculating DA and RT prices, which are decided through an auction system matching supply and demand while preserving the integrity of the power system. The DA market closes at 5:00 the day before the generation and distribution of electricity take place, and NYISO publishes DA prices at 11:00 on the same day. Market participants, thus, learn about DA prices only after the DA market closes. On the DA market, the scheduling of electricity is performed on an hourly basis, which allows participants to submit different bids for each hour. Conversely, the RT prices are updated every five minutes, and the hourly RT prices are obtained by aggregating all 5-minute RT prices published during the hour of interest.

On each transmission grid node, hourly DA and RT prices are determined through a locational-based marginal pricing (LBMP) approach, reflecting the marginal cost of consumption of an additional MWh of electricity on the node for that hour. When reported by the NYISO, the LBMP is further decomposed into three components: (1) energy cost, (2) congestion cost, and (3) losses. Congestion costs occur when the grid's electrical transmission capacity is exceeded under the most economical dispatching scenario. The NYISO is then constrained to dispatch more-expensive power generation units from local power plants, leading to substantial price increases for the associated nodes.

The present study is concerned with a quantity referred to as the *DART spread*, which is the difference between the DA and RT prices of power for a given grid node and hour. Since DA prices encompass market participants' expectations about the next-day RT prices, the DART spread could loosely be thought of as the market's price forecast error, up to a risk premium typically embedded in DA prices (Longstaff & Wang, 2004).

A thorough understanding of DART spread dynamics is essential for several market participants. For instance, virtual bidders who do not possess production or supply capacity and who must therefore reverse DA commitments in the RT markets are exposed to DART spreads instead of standalone prices from the DA or RT markets. A long position on the DA market puts the virtual bidders at risk when the DART is negative. DART spread dynamics also have implications for production facility and retailer risk managers who must decide on the volumes to be locked in ahead of time on the DA market to optimize risk-reward trade-offs faced by their institution.

Spike events are strong sources of risk for electricity market participants. Most of the literature is concerned with spikes in the electricity prices because generators, retailers, and large electricity consumers are exposed to extreme price levels. This has led several authors to explore price spike forecasting, e.g., Christensen et al., 2009, Christensen et al., 2012, Eichler et al., 2014, and Sandhu et al., 2016.

The present study instead considers the perspective of virtual bidders who are concerned with spikes in DART spreads rather than in prices. Therefore, this study considers the problem of forecasting these extreme DART events, or more precisely, the probability that a DART spread spike occurs in a given hour based on available information. Such a problem is expressed as a supervised learning problem which is tackled with four machine learning algorithms: (1) logistic regression, (2) random forests, (3) gradient boosting trees, and (4) deep neural networks (DNN).

To illustrate the developed approach, this study focuses on the Long Island zone as it is well known for being susceptible to DART spikes. This phenomenon results from Long Island's geographic location–it is a peninsula–which entails a smaller capacity to carry electricity from inexpensive power plants situated outside of the zone. This reduced grid capacity creates frequent bottlenecks, thereby raising the congestion price component for

Long Island.

From an economic standpoint, the added value of the model-predicted spike probabilities is assessed through trading backtests involving trading strategies that rely on these as signals. Such strategies are compared to a base-case strategy that systematically holds long positions on the DART spread. Such an approach seeks to collect the DART premium (the average positive DART spread), which rewards investors for exposing themselves to negative DART spread spikes caused by high RT prices. However, the flip side of this strategy is constant exposure to sudden significant losses stemming from such spikes. The signal generated by the predictive models makes it possible to modify the base-case strategy to develop novel strategies that avoid long positions when the likelihood of a price spike is too high. Such strategies are shown to lead to significantly better profitability and lower risk than the base-case strategy, outlining the contribution of the spike probability signals generated by the models. Results show that trading strategy performance is robust to the choice of the predictive model.

In summary, this paper offers two main contributions. The first is the comparison of multiple statistical and machine learning models producing predictions of DART spread spike occurrence. Since the bulk of the literature is concerned with price spikes, considering DART spreads instead of prices is a key differentiating feature of our study. The second contribution consists in showcasing the usefulness of the informational content embedded in spike probability forecasts by integrating such signals into trading strategies, which improves trading performance.

The paper is subdivided as follows. Section 4.2 provides a review of the raw data, describes the spike labeling methodology, and discusses the engineered feature set that is considered. In Section 4.3, four predictive algorithms are trained on the data, and their performance is assessed. Furthermore, individual features' contribution to predictive performance is assessed. Section 4.4 proposes simple trading strategies integrating the model-generated signals to determine investment positions, with their profitability and risk assessed during the conduction of an out-of-sample backtest. Section 4.5 concludes.

4.2 Data description

This section discusses the raw data and their transformation for subsequent predictive analysis with supervised learning algorithms. In particular, the construction of labels and features for each observation is outlined.

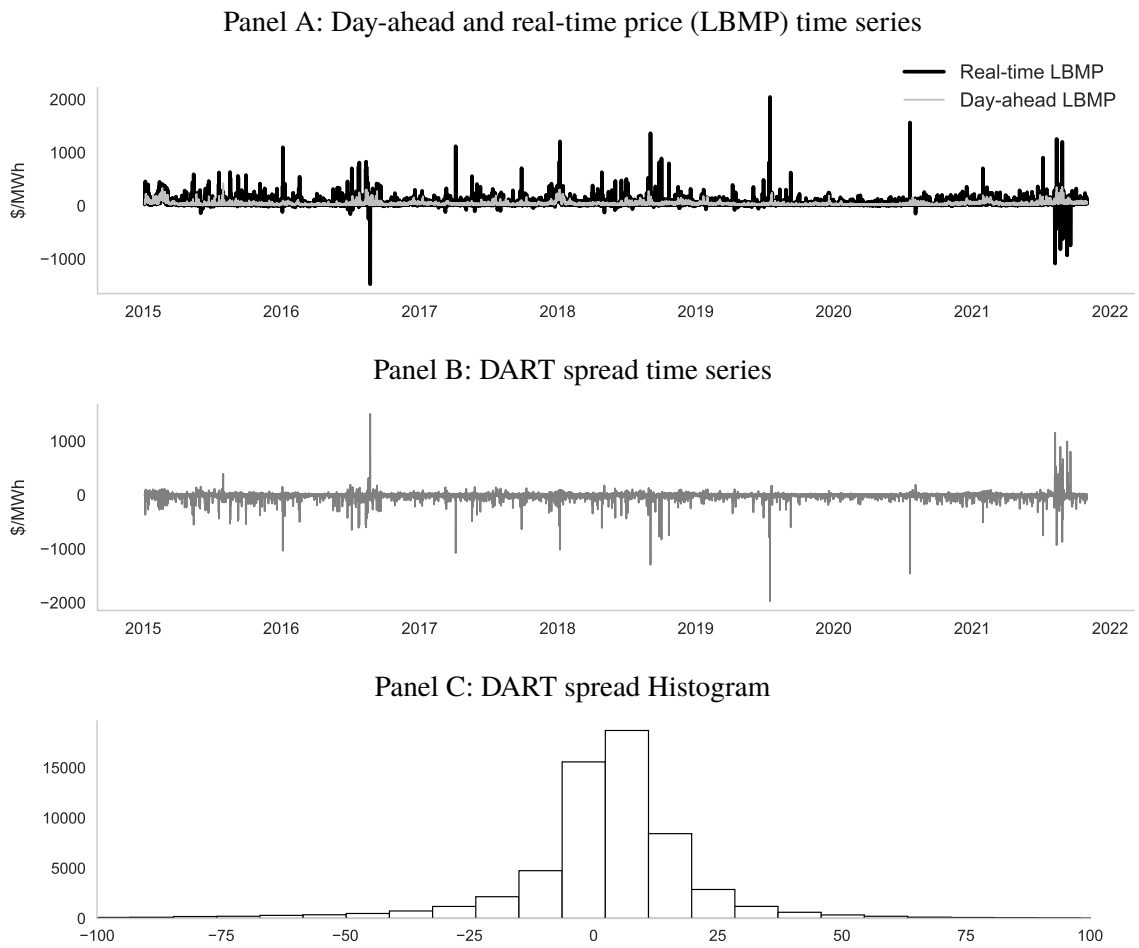
4.2.1 Raw Data

This research project focuses on the Long Island zone overseen by the NYISO. The NYISO provides historical data on day-ahead and real-time electricity prices and loads for its various zones, including Long Island, with hourly granularity, as well as the grid transfer capacity of the multiple interfaces supplying Long Island.[2] The electricity price (LBMP) and load data considered extend from January 1, 2015, to October 31, 2021. The DART spread for hour t is calculated by subtracting the hour-t real-time price RT_t from the corresponding day-ahead price DA_t, that is

$$\text{DART}_t = \text{DA}_t - \text{RT}_t.$$

Time series of the day-ahead and real-time prices are reported in Panel A of Figure 4.1, whereas Panel B provides the associated DART spread time series. Panel C exhibits the historical distribution of DART spreads through a histogram. As expected, real-time prices are much more volatile than day-ahead prices. The sample average of the DART spread is \$0.45/MWh, implying a positive DART premium compensating for aversion to spike risk.

[2]The Long Island zone's grid transfer capacity consists of the total transfer capacity from the Con EDLIPA, NPX-1385, NPX-CSC, SprainBrooke-Dunwoodie South lines Y50 and Y49, and PJM-NEPTUNE interfaces as described in NYISO, 2016. In this work, the grid transfer capacity considered is the sum of reported capacities for all such interfaces.



Panel A displays the hourly real-time and day-ahead price (LBMP) time series for the Long Island zone of the NYISO from January 1, 2015, to February 15, 2021. Panel B displays the corresponding DART spreads (difference between the day-ahead and real-time prices) for the same dates. Panel C displays a histogram characterizing the DART spread distribution on the same period.

Figure 4.1: Historical data for the real-time price, day-ahead price and DART spread

Various weather-related variables are obtained from the data provider *Openweathermap*. First, hourly realized temperature (in degrees Celcius) is collected for Long Island between January 1, 2015, and October 31, 2021.

Second, temperature forecast data are included, which consist of daily weather forecasts as of 18:00 with horizons ranging from 30 to 54 hours in three-hour increments. More

details about temperature forecasts are provided in Appendix C.1.

Historical temperature forecast data are only available as of October 7, 2017. Thus, to complete the sample and make up for missing temperature forecasts between January 1, 2015, and October 6, 2017, synthetic forecasts are generated by adding statistical noise on realized temperature data. Appendix C.1.2 explains this procedure in detail. Throughout the study, synthetic temperature forecasts are never included in test samples used for performance assessment, but only in training sets. This prevents spurious performance assessment due to information leakage from training to test sets where information related to realized temperature, rather than genuine forecasts, would unduly be provided to the predictive model.

4.2.2 Identifying electricity price spikes

This study aims to perform a daily computation of probabilities of a DART spread spike occurring in each hour of a given day. The task is approached as a supervised learning exercise. The response variable has a binary format: either "1" for hours in which a spike occurs or "0" otherwise. Such labels are not readily observable, and the first step is to determine which observations are considered to be spikes.

Spike identification criteria

While there is consensus in the literature on the notion that price spikes are extreme price events, no single, objective definition of "price spike" has emerged. Weron, 2007 and Janczura et al., 2013 highlight this lack of consensus in the literature and explain that such a definition is a subjective matter. Notwithstanding disparities among the various possible definitions, many authors such as Sandhu et al., 2016 or Janczura et al., 2013 characterize price spikes as extreme high prices that exceed a certain threshold and are short-lived.

Several approaches to identify electricity price spikes have been considered in the literature, among which the following three have proven quite popular:

- *Fixed price threshold*: Occurrences exceeding some selected fixed price threshold are classified as spikes. See for instance Klüppelberg et al., 2010, Amjady and Key-

nia, 2010, Christensen et al., 2012, Herrera and González, 2014, Eichler et al., 2014, Clements et al., 2015 and Manner et al., 2016, He and Chen, 2016.

- *Variable price threshold*: Occurrences exceeding a given sample quantile of observed values are flagged as spikes. For example, the highest (lowest) 5% of sample prices are considered spikes. A non-exhaustive list of works applying this criterion includes Trueck et al., 2007 and Sandhu et al., 2016.

- *Statistical filtering of spikes*: A stochastic process capturing price dynamics is selected and fitted to the data, and statistical filtering methods such as Sequential Monte-Carlo algorithms are applied to disentangle the portion of prices caused by spikes from that caused by normal price movements. See for instance Benth et al., 2007 and Gudkov and Ignatieva, 2021.

The aforementioned studies apply the threshold to electricity prices. However, other studies, such as Cartea and Figueroa, 2005 and Weron and Misiorek, 2008, identify spikes through price variations rather than through the level itself. For a more in-depth review of spike identification methodologies, see Janczura et al., 2013.

The first two approaches (fixed and variable price thresholds) are conceptually similar as both set pre-determined thresholds and directly assign a spike label to any prices exceeding such thresholds. The beauty of such methods lies in their simplicity. The third approach based on statistical filtering differs vastly from the first two. An a priori stochastic generative model for prices embedding the spike-generating mechanism must be specified, and statistical inference, i.e., filtering, methods are applied to estimate the spike component of prices based on its posterior distribution given observed prices. The use of sophisticated statistical filtering methods and the requirement to design a stochastic process matching the complex stylized facts of electricity prices are associated with higher inherent complexity.

This study uses a fixed threshold approach to identify spikes; it is a common choice made in the literature and by practitioners, which makes it possible to avoid the technical complexities that stem from the statistical filtering approach.

Results for DART spread spike identification

In the literature, the target variable used for spikes labeling is often the real-time price. However, for certain market participants, such as virtual bidders who take positions on the day-ahead market and revert them on the real-time market, the payoff is the DART spread. In this study, fixed thresholds are thus applied to DART spreads. The negative spikes are obtained through

$$S_t^- = \begin{cases} \text{DART}_t & \text{if DART}_t < \gamma^-, \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

for some fixed threshold $\gamma^- < 0$. Thus, observations are labeled as spikes, i.e., $S_t^- \neq 0$, when the DART spread is smaller than the specified threshold for negative spikes. For the remainder of this study, only negative spikes are considered. The size of a spike, when one occurs, is considered to be the DART spread itself; the DART spread is not subdivided into regular and spike components during an occurrence of a spike.

DART spikes thus embed an element of surprise associated with a sudden change of circumstances within a one-day horizon. This element entails that DART spikes are most likely to last less than one day, although definition (4.1) does not explicitly enforce short-livedness.

The first three columns of Table 4.1 exhibit summary statistics for the DA prices, RT prices and DART spreads. The standard deviation of RT prices (46.98) is larger than that of DA prices (27.81), highlighting the more volatile nature of RT prices. The large negative DART spread skewness (-7.62) stresses the significant risks to which the virtual bidders are exposed when taking long positions on the DART. As seen in the Long Island electricity price and DART spread time series exhibited in Figure 4.1, numerous extreme price events of various sizes are displayed. The following threshold values are considered to capture several spike magnitudes: $\gamma^- = -30, -45, -60$.[3] The last three columns of Table 4.1

[3]The selection of the -$30/MWh, -$45/MWh, and -$60/MWh values is driven by (i) discussions with industrial partners, and (ii) statistical considerations. The first threshold is set to -$30/MWh since it is the largest negative DART spread considered as an extreme economic event. The smallest threshold is set to -$60/MWh since it is among the lowest threshold values providing enough spike observations to train the

				Spikes		
	DA	RT	DART	$\gamma^- = -30$	$\gamma^- = -45$	$\gamma^- = -60$
Count				3534	2294	1605
Proportion				0.06	0.04	0.03
Mean	39.41	38.96	0.45	-84.44	-110.34	-135.44
Standard Deviation	27.81	46.98	37.58	98.91	114.68	129.19
Median	32.49	27.91	3.54	-55.62	-76.16	-135.44
Min	2.57	-1476.07	-1971.57	-1971.57	-1971.57	-1971.57
Max	424.00	2045.79	1506.74	-30.01	-45.01	-60.02
10%-level quantile	18.40	14.31	-16.88	-157.75	-191.38	-229.92
25%-level quantile	24.04	19.86	-3.28	-90.08	-119.79	-146.09
75%-level quantile	44.00	41.97	10.56	-39.34	-57.06	-73.95
90%-level quantile	65.28	72.15	19.29	-33.34	-49.04	-64.57
Skewness	3.42	7.01	-7.62	-6.82	-6.09	-5.54

Various summary statistics for non-null values of (S^-) labelled spikes. All numbers are expressed in $\$/\text{MWh}$.

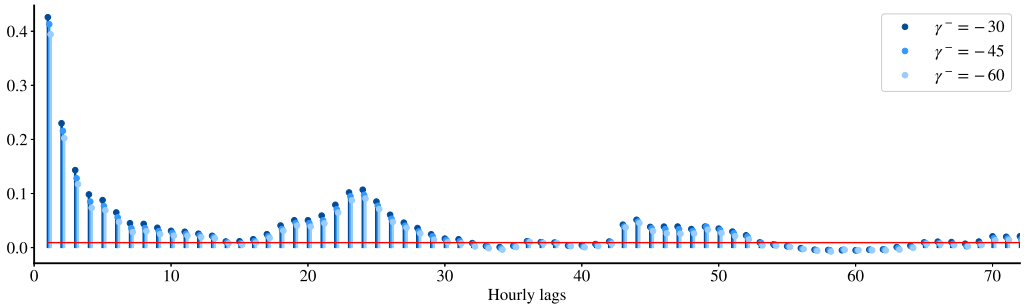
Table 4.1: Summary statistics for spikes

display summary statistics of the spikes, i.e., non-null values of S_t^-, for each threshold. The proportion of labeled spikes varies between 6% and 3%, indicating that only a minority of observations are labeled as spikes.

Figure 4.2 displays the autocorrelations of the DART spikes time series $\{S_t^-\}$ for lags extending from 1 to 72 hours. Results show that for the three considered thresholds, the autocorrelations are high and statistically significant at lags 24, 48, and 72, indicating the presence of spike clusters lasting multiple days.

Strong seasonal effects are detected in spike occurrences. Indeed, Figure 4.3 depicts

statistical learning models adequately. Indeed, Table 4.1 highlights that a threshold of -$60/MWh allows capturing 3% of the observations (1605 data points), indicating events that are sufficiently rare to be considered spikes, but frequent enough to retain sufficient training data.



Autocorrelations of the DART spikes time series $\{S_t^-\}$ for the three considered thresholds. The considered lags extend from 1 hour to 72 hours. The autocorrelations are computed over the whole sample period (2015-2021). The red line exhibits the upper bound of the 95% confidence intervals.

Figure 4.2: Autocorrelation of the DART spikes time series

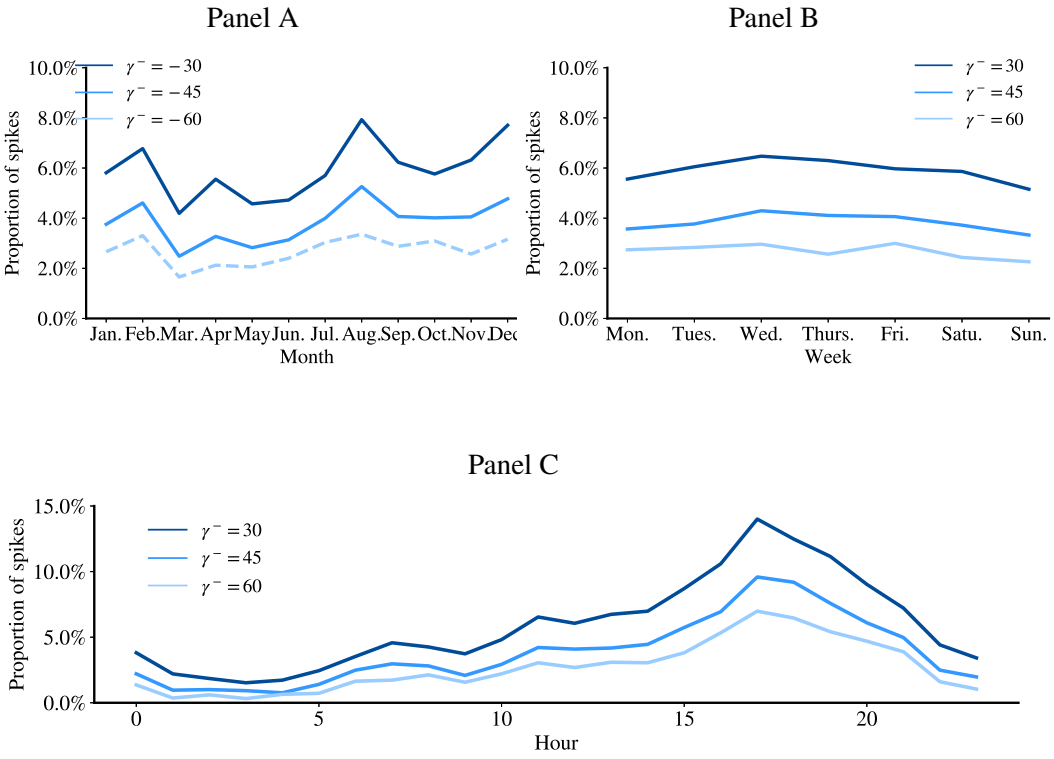
the proportion of observed spikes across the various months, days of the week, or times of day. Panel A indicates more frequent spikes in either summer or winter months but fewer in fall and spring. Panel C shows more frequent spikes during the late afternoon and fewer at night and in the early morning hours. Surprisingly, the week-versus-weekend effect is not striking, as seen in Panel B.

4.2.3 Features used for prediction

This section discusses and defines the various features, i.e. explanatory variables, considered in the spike prediction analyses. While some features are directly extracted from the raw data, others are obtained through data transformation. These engineered features aim to complement the information set provided by the conventional sources of information available. The steps involved in constructing such features are outlined.

Prediction generation timeline

Features should be included as predictive variables only if they are available at the time when predictions are being generated. The perspective of a market participant placing bids on the day-ahead market is considered herein. The timeline that determines availability of the predictors is now explained.



Proportion of spikes (number of spikes divided by the number of observations in the corresponding hourly/daily/monthly bucket). Fixed thresholds considered are $\gamma^- = -\$30/\text{MWh}$, $\gamma^- = -\$45/\text{MWh}$ and $\gamma^- = -\$60/\text{MWh}$. The data sample extends from January 1, 2015, to October 31, 2021.

Figure 4.3: Proportion of spikes per month, day of the week and hour

Each daily round of predictions being performed is associated with a window of three consecutive days. The third and last day, referred to as the *target day*, is the day on which all hourly predictions apply. Predictions are performed on the first day, coined as the *prediction day*. The second day (*trading day*) is where day-ahead bids are placed for all hours of the target day. The reason to include a delay between the time at which predictions are performed and the moment at which day-ahead bids are placed is to reflect that participants would typically require some time to process their predictive analysis outputs and determine their bids. Key elements of the timeline are now presented.

- Prediction day (day 1): At 11:00, the NYISO publishes hourly load forecasts for each hour of the target day. At 18:00, the temperature forecasts for the target day are published. All such information is combined with other available features to produce hourly spike probability predictions at 18:00.

- Trading day (day 2): Bids for the day-ahead participants are placed by 5:00. At 11:00, the NYISO publishes day-ahead prices for the target day.

- Target day (day 3): Real-time prices are revealed throughout the day, allowing for the computation of realized DART spreads.

In summary, all features entering the predictions applying to the target day must be available by 18:00 of the prediction day, i.e., two days in advance. Figure 4.4 provides an illustration of the timeline.

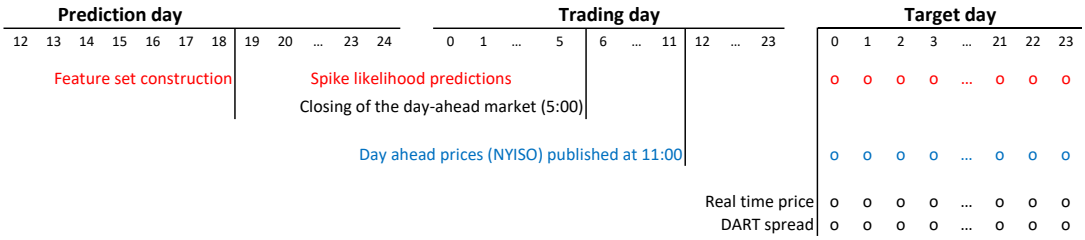


Figure 4.4: Timeline for spike predictions and subsequent DART spread realization

The list of features and their construction

The electricity literature identifies multiple features which are known to embed informational content that is useful to forecast prices and price spikes, see for instance Lago et al., 2021. Even though DART spread spike forecasting is a different exercise than price and price spike forecasting, we nevertheless consider features similar to these proposed in such literature.

The set of all selected features, which are listed in Table 4.2, can be divided into three categories: (1) forward-looking features, (2) seasonal features, and (3) backward-looking features.[4]

Forward-looking	**Seasonal**	**Backward-looking**
HDD forecast	Hour	Past spikes
CDD forecast	Month	Past day-ahead price error
Load/Grid	Week-end/Holidays	Past day-ahead load error

Table 4.2: Feature variables used for spike prediction

The forward-looking features encompass information related to market participants' expectations about the future realization of various variables. For any observation, i.e., target hour, such features include the 48-hours-ahead load forecast to grid capacity ratio (see below for details), and the time-18:00 prediction day's temperature forecast associated with the corresponding target hour, i.e., the 30- to 54-hours-ahead forecast depending on the target hour. As explained below, two non-linear transformations of the latter temperature forecast are considered.

The NYISO (see Itron, 2008) as well as the literature (see Fan et al., 2019, Zahedi et al., 2013 or Yi-Ling et al., 2014) consider non-linear transformations of temperature metrics. This makes it possible to reflect the non-linear relationship between electricity consumption and temperature. Indeed, more electricity is consumed when temperatures are either very low (as heaters are turned on) or very warm (as air conditioners are turned on). A popular methodology described in the literature and adopted by the NYISO (see Itron, 2008) consists in transforming the temperature feature into *heating degree day* (HDD) and *cooling degree day* (CDD) features. The HDD and CDD thus reflect the electricity demand for

[4] An experiment reported in Section C.4 of the Appendix integrates an additional feature called the *cumulative temperature and humidity index* (CTHI), which is used as predictor by the NYISO to forecast the load. Such inclusion does not improve the general performance of the models. Furthermore, a graphical exploration exercise reveals that the CTHI feature exhibits strong dependence with other features (HDD, CDD and Load/Grid). Therefore, its inclusion in the feature set increases the likelihood of multicollinearity-related issues. For such reasons, the CTHI is not included in the set of features.

heating and cooling and are expressed as

$$\text{HDD}_t = \max\left(\text{BP} - T_t, 0\right), \quad \text{CDD}_t = \max\left(T_t - \text{BP}, 0\right),$$

where T_t is the hour-t temperature measurement and BP $= 18.3°$ C ($65°$ F) is the breakpoint considered by the NYISO, which is also used herein. When used in conjunction with predictive algorithms that do handle automatically non-linear relationships, the HDD and CDD transformed variables are most likely more appropriate than the original temperature forecast as a predictive feature.

The last forward-looking feature, *load/grid*, corresponds to the ratio of the 48-hours-ahead load forecast over the grid transfer capacity supplying Long Island. Indeed, the load-to-capacity ratio has been suggested by Anderson and Davison, 2008 as a driver of spike likelihood, although the latter paper considers generation capacity instead of transmission capacity. The interface transfer capacity is used in the denominator to reflect that, for the same amount of load, a curtailed transmission capacity is associated with higher spike risk due to an increase in the likelihood of bottlenecks.

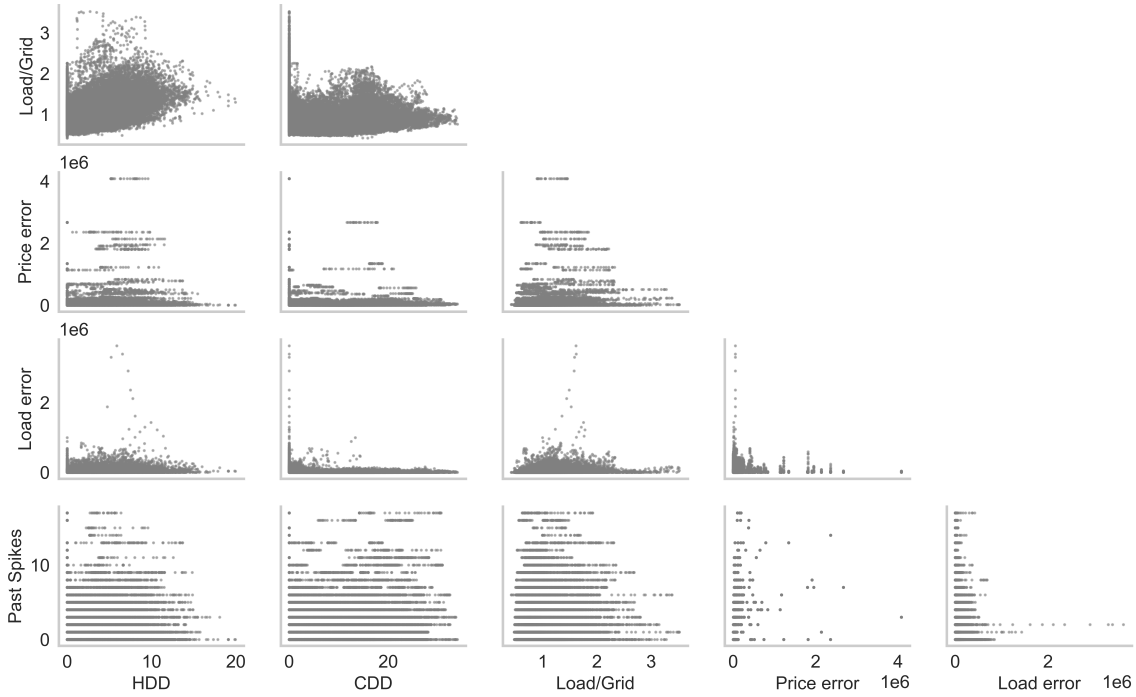
The second class of features, namely the seasonal features, aim to capture well-known seasonal patterns in electricity markets. They include dummy variables for each hour of the day and month of the year, and another dummy variable indicating (additionally) if the day of the target hour is either a weekend day or a holiday.

The last category, the backward-looking features, are engineered features that consist of metrics computed from historical observations. Such features are meant to capture market conditions of the recent past. The three backward-looking features are calculated once at the prediction time, and each of them has identical values for all 24 hours of the target date. The first backward-looking feature, *past spikes*, is the number of observed spikes in the 24 hours leading up to the prediction. This reflects the tendency of spikes to occur in clusters, as highlighted for instance in Klüppelberg et al., 2010, Christensen et al., 2012, Herrera and González, 2014, He and Chen, 2016, and Manner et al., 2016. Thus, the presence of many recent spikes indicates a higher likelihood of observing spikes in the near future.

The second backward-looking feature, *past day-ahead load error*, aims to indicate periods where the estimation of near-term future load consumption by the market proves more difficult. Such a feature is helpful because sudden and unexpected surges or drops in load

increase the likelihood of observing a spike. The *past day-ahead load error* is computed by summing the hourly squared load forecast errors (real-time load minus the day-ahead load) over the 24-hour period leading up to the prediction.

The construction of the third backward-looking feature, *past day-ahead price error*, is analogous to that of the past day-ahead load error; it is also calculated by summing the last 24 hourly observed squared price forecast error (real-time price minus the day-ahead price) in the period prior to the prediction. It is meant to capture periods with higher price volatility.

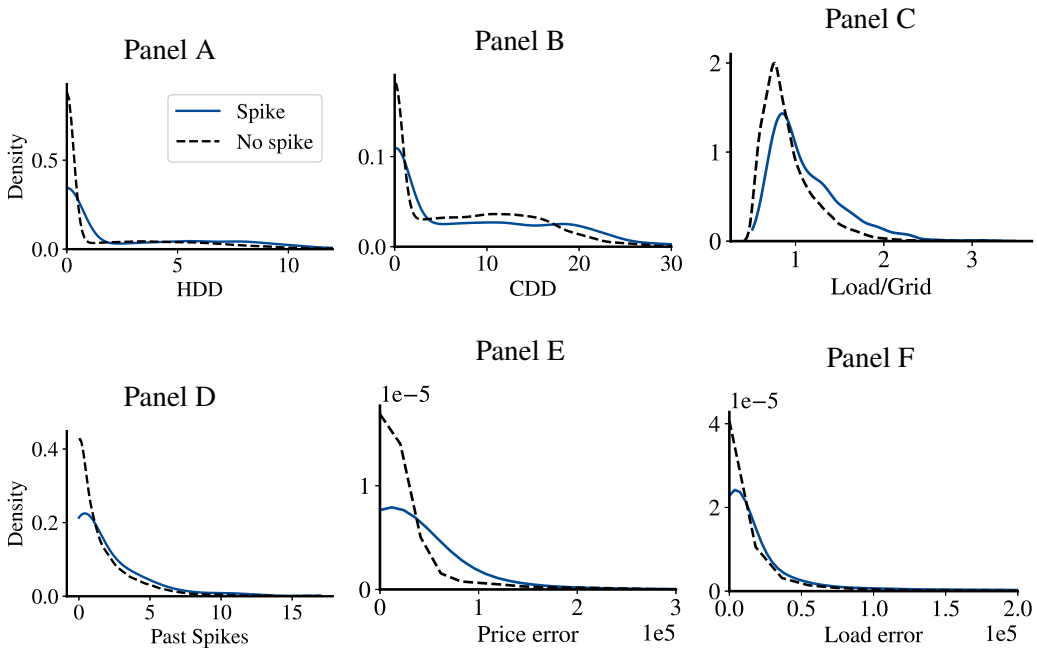


Scatterplots of realized values for model features. The presented features are *heating degree days* (HDD), *cooling degree days* (CDD), *Load/Grid* representing the load forecast to grid transfer capacity ratio, *past spikes*, *past day-ahead price error* (Price error), and *past day-ahead load error* (Load error).

Figure 4.5: Scatterplots of model features

Figure 4.5 illustrates the scatterplots of realized values for all considered features in each of the hourly observations, thereby illustrating the relationship between the features. As expected, the HDD and CDD features are positively correlated with *load/grid*. The

relationship between the other features does not display any clear dependence structure.



Each panel illustrates the kernel density estimate of a feature distribution conditional on either the presence (blue continuous line) or the absence (black dashed line) of a DART spike. The presented features are *heating degree days* (HDD), *cooling degree days* (CDD), *Load/Grid* representing the load forecast to grid transfer capacity ratio, *past spikes*, *past day-ahead price error* (Price error), and *past day-ahead load error* (Load error).

Figure 4.6: Kernel density estimates of feature distributions

Figure 4.6 illustrates kernel density estimates of feature distributions conditional on either the presence or the absence of a DART spike. All panels indicate that larger feature values are more likely to occur when spikes are observed.

4.3 Spike prediction model

This section illustrates the prediction of DART spread negative spike probabilities based on available information. Four predictive algorithms are applied to the data, and their performance is assessed through conventional statistical metrics. A feature importance assess-

ment evaluating the contribution of each feature to predictive performance is also provided. We hereby focus solely on the spike occurrence probability and leave the challenging task of predicting the spike magnitude to a future study.

4.3.1 The predictive models

The four predictive algorithms considered are (1) logistic regressions, (2) random forests, (3) gradient boosting trees, and (4) feed-forward deep neural networks (DNN).[5] Details about their implementation are presented in Appendix C.2.

The logistic regression is a conventional base case for any binary classification problem. It expresses the logit of the probability of a spike as a linear function of predictors, which makes the model easy to interpret and straightforward to estimate with conventional regression tools. However, logistic regression is not necessarily well suited to handling non-linear relationships with the target variable and interactions between features. Since electricity market price data might be fraught with such complex relationships, it is desirable to contemplate alternative predictive models. Therefore, random forests, boosting trees, and neural networks are also considered as they can automatically represent complex and non-linear interactions.

Except for logistic regression, the models considered here cannot be trained out-of-the-box and require the user to select a set of hyperparameters. The hyperparameter tuning methodology is described in greater detail in Appendix C.2.2.

4.3.2 Model performance

Model performance is assessed through two statistical metrics: the area under the receiver operating curve (AUC) and the average log-likelihood. The former is meant to measure the discriminatory power of the models, whereas the latter characterizes the precision of

[5] A stacked classifier combining the predictions of the four models (logistic regression, random forests, gradient boosting trees, and DNN) is considered in Section C.5 of the Online appendix. Stacking the models does not improve the out-of-sample predictive performance metrics (i.e. the log-likelihood and AUC) in comparison to the standalone models.

the predictive model. The receiver operating curve (ROC) provides the set of all possible trade-offs between false positive and false negative error rates obtained across the possible choices of probability thresholds for classification (see, for instance, James et al., 2013). The value of the AUC must lie between 0 and 1, with a higher value indicating a higher ability to distinguish between the two classes. An AUC under or equal to 0.5 indicates that the model has no predictive power. The second performance metric, the average log-likelihood, is computed by comparing spike labels and the predicted probabilities:

$$\ell = \frac{1}{\tau}\sum_{t=1}^{\tau} \log\left(p_t\right)\mathbb{1}_{\{S_t^- < 0\}} + \log\left(1-p_t\right)(1-\mathbb{1}_{\{S_t^- < 0\}}) \tag{4.2}$$

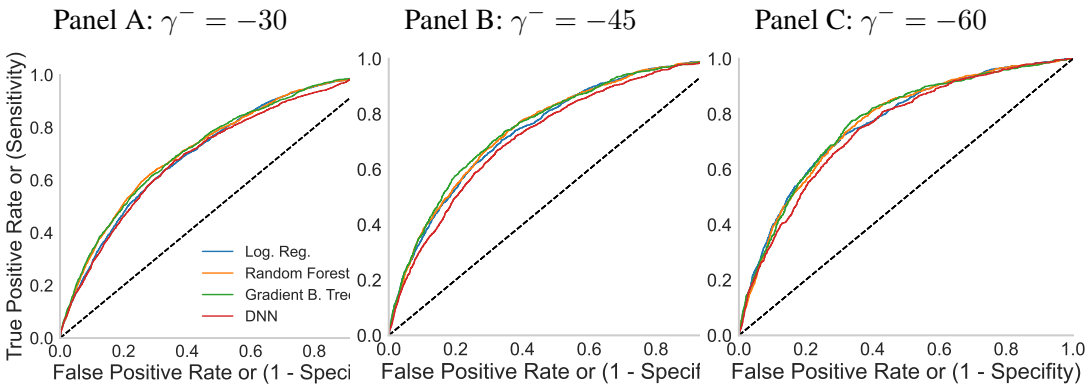
where p_t corresponds to the model generated probability of observing a spike in hour t, τ is the sample size and S_t^- is zero if and only if no negative spike occurs in hour t.

The performance assessment relies on an expanding window approach consisting in iteratively training the model over an expanding training set for each testing iteration. During the first iteration, the model is trained over the first three years of the dataset. The out-of-sample performance metrics are computed over the following year, i.e., the fourth year. One year is added to the training dataset for the subsequent iteration while generating predictions for the following year. Performance metrics (AUC and average log-likelihood) are computed for training and test set observations.

Panel A of Table 4.3 displays the AUC for negative spikes. The results presented are for the training sets (in-sample) and test sets (out-of-sample). The last row of each panel displays the aggregated out-of-sample results, i.e., the computed AUC over the merged out-of-sample sets. A larger AUC indicates that the model is more powerful at discriminating between the binary classes. All Panel A entries show an AUC considerably above 0.5 (more precisely, always above 0.65), indicating that the four models exhibit material discriminatory power for every threshold considered. The in-sample AUC is only slightly higher than the out-of-sample AUC, implying that models are not plagued with over-fitting issues. The gradient boosting trees displays the highest aggregated out-of-sample AUC for each threshold ($\gamma^- = -\$30/\text{MWh}$ (0.722), $\gamma^- = -\$45/\text{MWh}$ (0.755), and $\gamma^- = \$60/\text{MWh}$ (0.769)).

However, all models display quite similar out-of-sample AUC across all thresholds.

This result is confirmed in Figure 4.7, which illustrates aggregated out-of-sample ROC curves for all models and threshold γ^-. The displayed ROC curves are similar in shape and height for all panels, except for the DNN, which is slightly lower than the others. Thus, there is no apparent domination of one model over the others.



Each panel displays the ROC curves for the four predictive models at a specific threshold (γ^-). The ROC curve illustrates the attained true positive rate on the y-axis against the corresponding false-positive rate on the x-axis. The formula for the true positive rate and false positive rate is True positives$/$(True positives $+$ False negatives), and $1 -$ True negatives$/$(True negatives $+$ False positives) respectively. The ROC curves are computed over the aggregated out-of-sample set (2018 to 2021).

Figure 4.7: ROC curves

Panel B of Table 4.3 displays the average log-likelihood for each model and every threshold considered (γ^-) and confirms previous findings. Indeed, the gap between the in-sample and out-of-sample model performance is quite small. Furthermore, the models demonstrate similar performance for each threshold, although the gradient boosting trees display a slightly higher aggregated log-likelihood. This showcases that prediction performance is robust to the choice of predictive models. To assess which of the models are the best-performing ones from a statistical standpoint, the Hansen et al., 2011 model confidence set approach is considered.[6] Stars in Panel B's last row identify the best model(s) associated with each threshold. The model confidence set approach indicates that for each threshold, the gradient boosting tree model significantly outperforms the other models,

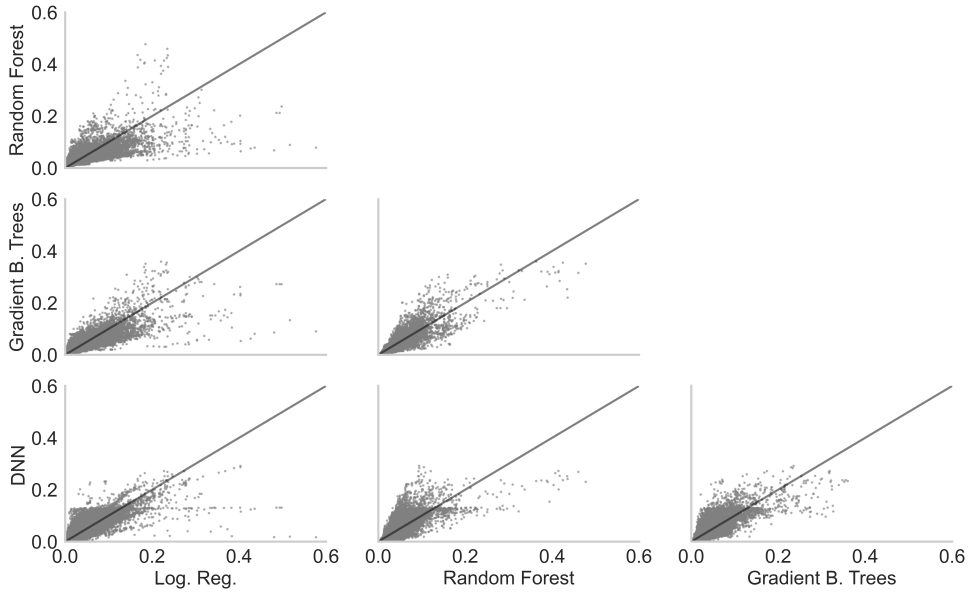
[6]The implementation of the Hansen et al., 2011 model confidence approach is described in detail in Appendix C.3.

with an exception for the $\gamma^- = -\$60/\text{MWh}$ threshold where the gradient boosting tree statistically outperforms the random forest and the DNN models but not the logistic regression.

Figure 4.8 illustrates the scatterplot matrix of model-generated out-of-sample probabilities for $\gamma^- = -\$60/\text{MWh}$. Most of the time, spike likelihoods are moderate. Therefore, one should not expect to predict spikes with very a high degree of certainty. This result raises the question of what level of spike likelihood could be considered substantial enough to become actionable. This issue is investigated in Section 4.4. Despite the very close performance of the models, material dissimilarities between the individual predicted probabilities are observed, especially for the DNN model. For instance, unlike the other models, the DNN rarely outputs probabilities of observing a spike of over 20%. This result implies that although they all exhibit similar statistical performance, the models might not be considered fully interchangeable. This is further investigated in the next section, where trading strategies based on each model are examined.

Figure 4.9 reports the relationship between the proportion of observed spikes and the spike probabilities generated by each model over the out-of-sample period (2018 to 2021). To obtain such figure, the observations are regrouped into buckets based on their model generated spike probability.[7] Each bar indicates the proportion of observed spikes within each bucket. The precision of the model is deemed adequate if the proportions are close to the associated probabilities for each bucket, i.e. if the bars closely follow the identity function (the 45-degree diagonal). This relationship seems to hold reasonably well for buckets associated with low probabilities that contain large numbers of observations. For high-probability buckets, unreported statistical tests highlight that the variability can be attributed to the small number of observations. Indeed, the green dots indicate that only a few dozen observations are associated with high spike probabilities.

[7]The non-overlapping bucket intervals are of size 2%. The first and last bucket intervals are $[0\% - 2\%[$ and $[48\% - 50\%]$ respectively.



The panels display the scatterplots for all pairs of predicted spike probabilities belonging to the aggregated out-of-sample set (2018 to 2021). The threshold is $\gamma^- = -60$. The identity curve is also displayed in each panel.

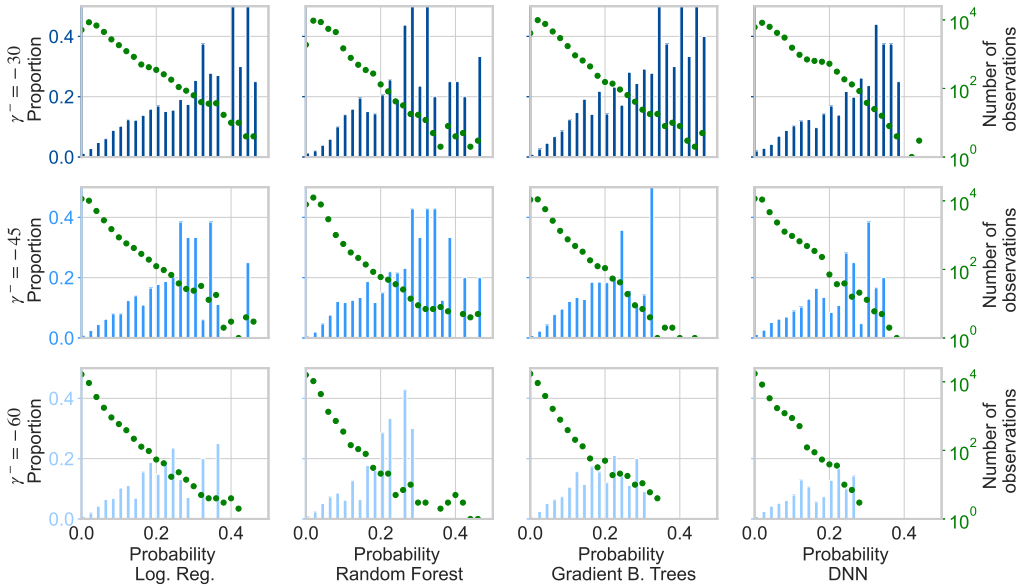
Figure 4.8: Scatterplots of predicted spike probabilities across models for $\gamma^- = -60$

4.3.3 Feature importance assessment

Spike prediction probabilities are constructed by combining the informational content of several features. This section aims to provide information about how each feature contributes to the overall model predictive performance, thereby making it possible to rank features in terms of their absolute and relative importance.

Each feature's importance is assessed through two approaches: (1) Shapley decompositions and (2) marginal performance loss through feature removal.

The Shapley, 2016 decomposition has recently been integrated into the machine learning literature through algorithms referred to as SHAP (Lundberg & Lee, 2017) or SAGE (Covert et al., 2020). They make it possible to decompose individual predictions (in SHAP) or their total predictive performance (in SAGE) into a sum of contributions from the various features, thereby making it possible to evaluate their respective importance. This study focuses on the SHAP algorithm, in which the feature i contribution to the spike probability



Each panel's x-axis reports spike probability intervals, while the y-axis displays the proportion of spikes observed relative to the number of observations inside the interval. The figures are obtained using the out-of-sample data period (2018 to 2021). The green dots display the number of observations in each interval with a log-scale y-axis.

Figure 4.9: Proportion of spikes vs predicted probability

predictions made for hour t is defined as

$$\phi_{i,t} = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{t,S \cup \{i\}}) - f_S(x_{t,S}) \right]$$

where F is the set of all predictors, $|\cdot|$ denotes the cardinality of a set, $x_{t,S}$ is the hour-t features values for the subset of features S and $f_S(x_{t,S})$ is the spike probability generated by the model trained exclusively with predictors S. It quantifies adjustments to predictions when the subsets of features are incremented with predictor i. The Shapley decomposition has the favourable property of explaining each prediction as the sum of its contributions:

$$f_F(x_{t,F}) = \phi_{\emptyset,t} + \sum_{i \in F} \phi_{i,t}.$$

To measure the importance of each respective feature, the average absolute feature contri-

butions are presented:

$$\psi_i = \frac{1}{\tau}\sum_{t=1}^{\tau}|\phi_{i,t}|, \tag{4.3}$$

with larger values of ψ_i relative to other features meaning that feature i is more impactful when making predictions. SHAP values are computed using the `Python` package `shap`.

Figure 4.10 reports the mean absolute Shapley values (4.3) computed over the out-of-sample period for every feature, model and threshold considered. Results indicate that the load forecast over transfer capacity ratio (*load/grid*), the hourly and monthly indicators (*hour* and *month*), the CDD/HDD and *past spikes* features contribute the most to the predictions. Interestingly, for all predictive models, the *past spikes* feature offers a much higher contribution when the threshold γ^- is large than when it is small. Other features such as *weekend/holidays* and *past day-ahead load error* exhibit generally low contributions to the predictions.

To complement the information provided by SHAP, this study quantifies the marginal performance loss through the decrease in out-of-sample average log-likelihood observed when any of the features are omitted from the set during training. A drop in performance implies that the feature does bring useful information, while minuscule improvements up to degradation in performance suggest that the feature conveys little to no information. Figure 4.11 exhibits such percentage increase/drops in the average log-likelihood. Features with the highest contribution are the load forecast to transfer capacity ratio, hourly indicators, and the number of spikes in the previous day. Conversely, features like *past day-ahead price error*, *past day-ahead load error* and *weekend/holidays* once again have little to no predictive power relative to the other features for every model. Such findings are mostly consistent with those provided by the SHAP algorithm.[8]

The Shapley, 2016 decomposition and the marginal performance loss provide different

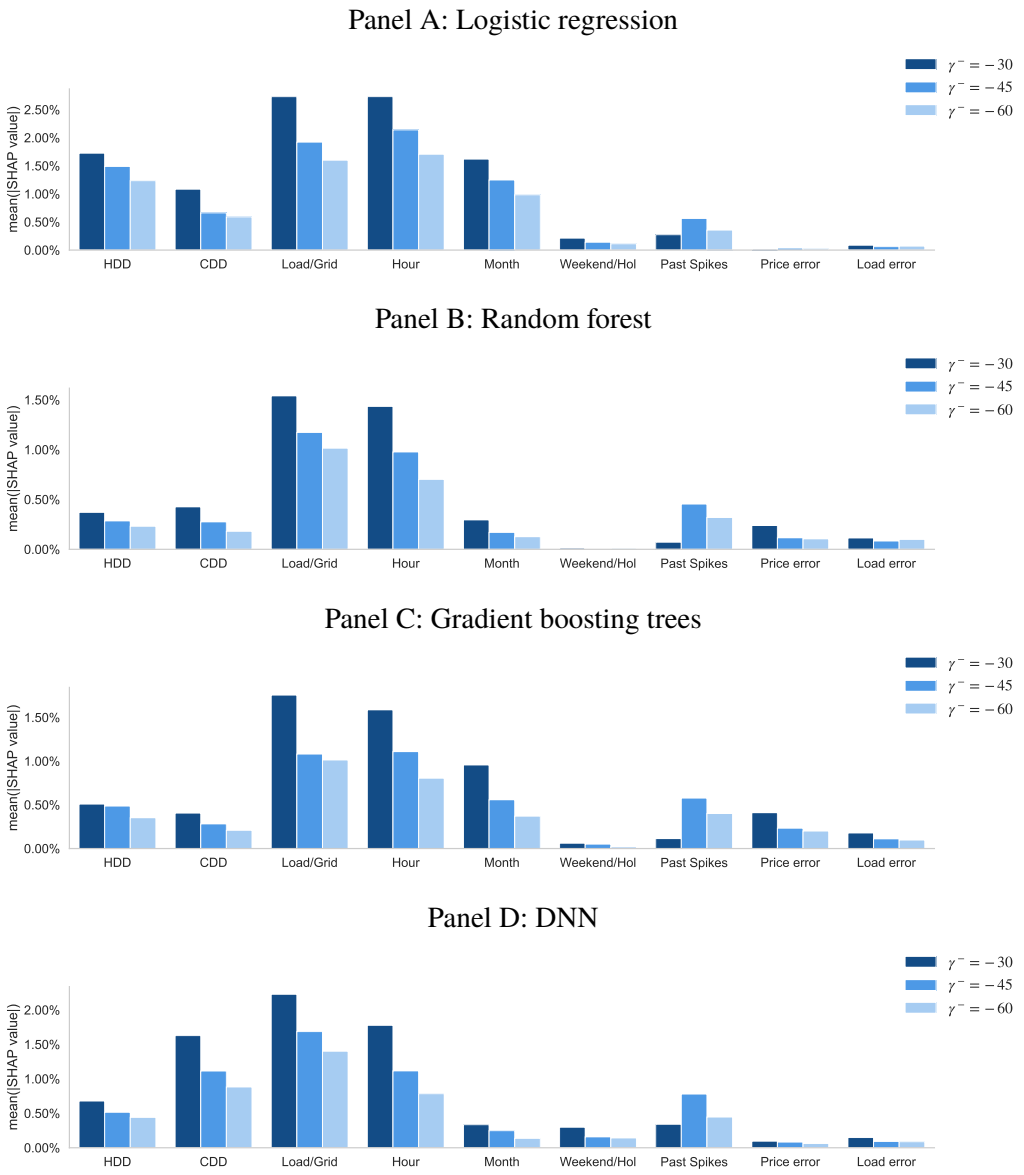
[8] A model performance assessment with a revised feature set selected based on the results of the present section is reported in Section C.6 of the Online appendix. Results are not reported in the body of the article because the model performance with the revised feature set is improved artificially due to data leakage; the decision to remove some features from the feature set described in Section 3.2 is based on performance results computed over the out-of-sample period, thus leaking information from the out-of-sample period into the feature selection procedure.

information about the features contribution. While the Shapley decomposition quantifies the extent to which the model relies on each respective feature, the marginal loss assesses the incremental performance gain/loss when the feature is included into the model. The Shapley, 2016 decomposition indicates that the models strongly rely on some of the variables that lead to low marginal performance gains, or even to a performance loss. This phenomenon is the result of strong dependence between certain features. For example, the *month* feature is extensively used by the models, but its associated marginal loss is mainly negative (i.e. dropping such variable improves out-of-sample performance). This result can be attributable to the fact that other features, such as *load/grid*, *heating degree days*, and *cooling degree days*, exhibit strong dependence with the *month* feature and already intrinsically capture the information provided by the latter quantity related to spike likelihood prediction.

		Logistic Regression			Random Forest			Gradient Boosting trees			DNN		
	γ^-	-30	-45	-60	-30	-45	-60	-30	-45	-60	-30	-45	-60
A	AUC												
In-sample	2015-2017	0.712	0.730	0.742	0.758	0.784	0.818	0.768	0.783	0.796	0.722	0.699	0.715
	2015-2018	0.704	0.727	0.743	0.743	0.769	0.793	0.757	0.774	0.791	0.725	0.751	0.767
	2015-2019	0.710	0.739	0.755	0.747	0.775	0.798	0.752	0.783	0.806	0.727	0.715	0.734
	2015-2020	0.713	0.740	0.756	0.749	0.776	0.800	0.752	0.791	0.794	0.725	0.746	0.766
Out-of-sample	2018	0.669	0.708	0.729	0.672	0.710	0.729	0.680	0.719	0.726	0.657	0.686	0.718
	2019	0.717	0.771	0.793	0.755	0.789	0.820	0.740	0.785	0.807	0.738	0.745	0.765
	2020	0.740	0.762	0.786	0.741	0.746	0.761	0.740	0.756	0.780	0.705	0.736	0.757
	2021	0.701	0.730	0.756	0.707	0.737	0.740	0.706	0.747	0.756	0.684	0.717	0.741
	Aggregated	0.710	0.745	0.765	0.722	0.751	0.766	**0.722**	**0.755**	**0.769**	0.700	0.723	0.748
B	Average log-likelihood												
In-sample	2015-2017	-0.210	-0.154	-0.121	-0.202	-0.147	-0.113	-0.201	-0.147	-0.115	-0.209	-0.158	-0.124
	2015-2018	-0.213	-0.154	-0.120	-0.207	-0.149	-0.115	-0.203	-0.148	-0.116	-0.209	-0.151	-0.118
	2015-2019	-0.203	-0.145	-0.112	-0.197	-0.140	-0.107	-0.196	-0.138	-0.106	-0.200	-0.148	-0.114
	2015-2020	-0.197	-0.141	-0.107	-0.191	-0.136	-0.103	-0.190	-0.133	-0.103	-0.195	-0.140	-0.106
Out-of-sample	2018	-0.224	-0.155	-0.120	-0.223	-0.156	-0.121	-0.222	-0.155	-0.121	-0.226	-0.157	-0.120
	2019	-0.163	-0.109	-0.079	-0.162	-0.111	-0.079	-0.159	-0.109	-0.079	-0.162	-0.111	-0.080
	2020	-0.172	-0.122	-0.086	-0.169	-0.121	-0.085	-0.167	-0.119	-0.084	-0.172	-0.124	-0.088
	2021	-0.279	-0.196	-0.140	-0.278	-0.195	-0.142	-0.278	-0.194	-0.140	-0.284	-0.201	-0.143
	Aggregated	-0.206	-0.143	-0.105*	-0.205	-0.143	-0.105	**-0.203***	**-0.142***	**-0.104***	-0.208	-0.146	-0.106

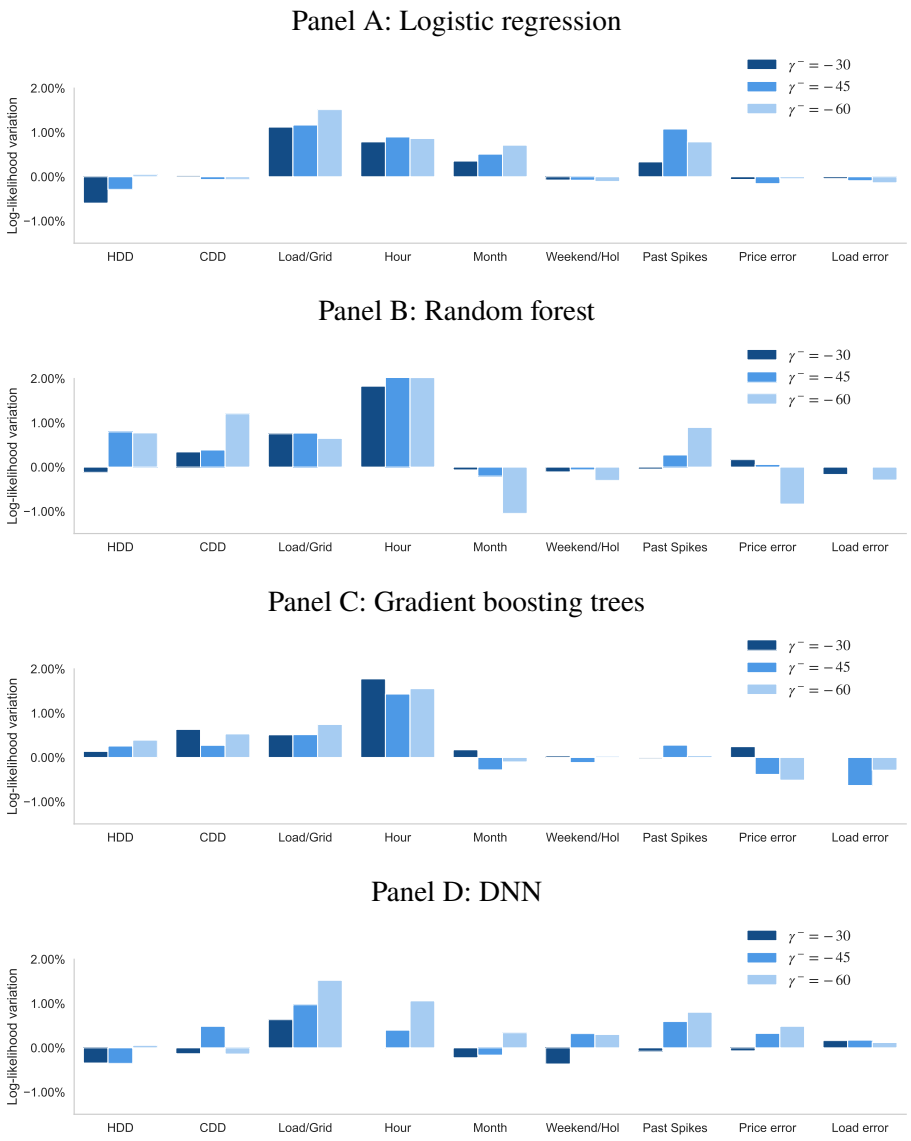
The four models are the logistic regression, the random forest, gradient boosting trees, and the deep neural network (DNN). Panel A's performance metric is the area under the curve (AUC), while Panel B is the average log-likelihood. The models generate out-of-sample predictions for 2018 to 2021. The models are trained on the previous years' observations for each out-of-sample forecast. For example, to generate out-of-sample forecasts for 2019, the models are trained on the observations from 2015 to 2018. For each threshold, the Hansen et al., 2011 confidence set approach is applied to the aggregated out-of-sample log-likelihood to identify the set of models whose performance cannot be distinguished from that with the highest performance. The best models remaining in the model confidence set at a level of significance 5% are identified with a star in the table. The testing procedure is described in Appendix C.3.

Table 4.3: In-sample and out-of-sample performance metrics



Each panel reports, for the three thresholds considered, the features' mean absolute SHAP contributions over the out-of-sample period (2018 to 2021). The features are *heating degree days* (HDD), *cooling degree days* (CDD), *hour* indicators, *month* indicators, *weekend/holidays* indicators (Weekend/Hol.), *past spikes*, *past day-ahead price error* (Price error), and *past day-ahead load error* (Load error). The SHAP values for the categorical features *month* and *hour*, which are divided into buckets for the logistic regression, are computed by summing the SHAP values of each category.

Figure 4.10: Shapley additive explanation values



Each panel reports the percentage decrease in the out-of-sample log-likelihood when the feature is excluded from the feature set. More precisely, the model is re-trained with a reduced feature set where only the targeted feature is removed. The table reports the ratio of the difference between the out-of-sample log-likelihood from both models. The features are *heating degree days* (HDD), *cooling degree days* (CDD), *hour* indicators, *month* indicators, *weekend/holidays* indicators (Weekend/Hol.), *past spikes*, *past day-ahead price error* (Price error), and *past day-ahead load error* (Load error).

Figure 4.11: Decrease in average log-likelihood when removing a single predictor

4.4 Trading strategies performance

This section aims at evaluating the performance of the four models from an economical (rather than statistical) perspective. A large strand of the electricity literature integrates price forecasting methods to devise the trading strategies (see Conejo et al., 2005, Zhang et al., 2012, Ziel et al., 2015, Lago et al., 2018, and many others). This study concentrates only on forecasting DART spikes as a complement to such methods.

In a first exercise, two trading strategies are implemented over the in-sample set, i.e. from January 2015 to December 2017. Both strategies take a long position on the DART when the model-predicted probability is lower than a predetermined cut-off point. The first strategy takes no position otherwise, i.e., when the probability exceeds the cut-off point, while the second strategy takes a short position in the DART. The first strategy reflects the situation of a participant trying to collect the DART premium when the risk of a spike is not too considerable, while in the second strategy, the participant tries to also benefit from the occurrence of a spike. The volume of any position taken in the two strategies is always 1 MWh. This first exercise is completely in-sample and aims to (1) better understand the effect of the cut-off point on the strategies' performance and (2) select a cut-off point for the subsequent out-of-sample exercise. The left (respectively right) panels of Figure 4.12 display the cumulative hourly profits and losses (P&L) of the first (respectively second) strategy as a function of the cut-off point over the in-sample horizon.

Every panel of Figure 4.12 unanimously indicates that profits are maximized for relatively small cut-off points ranging between 5% and 12%, depending on the threshold γ^-. This result is explained by the asymmetry of the loss function, where gains associated with the successful prediction of a spike far exceed the losses incurred when falsely predicting a spike. Such asymmetry encourages the participant to short the DART spread as soon as a small potential spike signal is detected. Comparing the two strategies, Strategy 2 outperforms the first one, pointing toward the added value of short DART positions when the spike probability is high.

The two same trading strategies are implemented in a second exercise, this time over the out-of-sample set (from January 2018 to October 2021). To avoid information leakage, the considered cut-off points are selected based on the aforementioned first in-sample

exercise exhibited in Figure 4.12 and are chosen as respectively 12%, 7% and 5% for $\gamma^- = -\$30/\text{MWh}$, $\gamma^- = -\$45/\text{MWh}$ and $\gamma^- = -\$60/\text{MWh}$.[9] To further assess the added value of the spike probability forecast as a trading signal, the two strategies are compared against a third one, a base case strategy consisting in always taking a long position in the DART.

The left (respectively right) panels of Figure 4.13 illustrate the time evolution of a portfolio value starting at \$0 and invested in the first (respectively second) strategy. Looking at Strategy 1, for all thresholds and models, the portfolio value time series appears to closely follow the upward trends of the base case portfolio from January 2018 to June 2019 and January to June 2021 periods while being much less impacted by the downward trends over June 2019 to January 2021 and May to October 2021 periods. Furthermore, Strategy 1 generates substantially higher profits over the out-of-sample period while holding fewer positions than the base case strategy. The second strategy exhibits a different behaviour where the portfolio value tends to increase steadily over time, except for a few downward stretches.

Table 4.4 illustrates the average P&L per position and the total P&L organized by model, threshold, and year. The results for Strategy 1 indicate that every model generates a positive aggregated P&L, unlike the base case strategy, which generates a negative aggregated P&L. For all predictive models, the cumulative profit is generally more significant for lower thresholds ($-$\$45/MWh and $-$\$60/MWh) than for the higher ones ($-$\$30/MWh). In 2018 and 2021, Strategy 1 generates a strong total P&L, while in 2019 and 2020, it is more modest or even negative. Nonetheless, the four algorithms produce a higher total P&L in most years compared with the base case strategy. As expected, the P&L produced by Strategy 2 is more prominent than that of Strategy 1 since Strategy 2 additionally benefits from the short positions on the DART spread when the spike probability is sufficiently high.

As in Section 4.3, the Hansen et al., 2011 model confidence set approach is harnessed to identify the best-performing model(s) for each threshold using the P&L as the loss function. Stars in the last row of each of the Table 4.4 panels identify models with superior predictive

[9]The cut-off points are selected as the values maximizing the cumulative in-sample P&L of the gradient boosting trees model.

ability. Results show that no single model significantly outperforms all others. The gradient boosting tree model and the logistic regression are always included in the best-performing set, while the random forest is included for the $-\$45$/MWh and $-\$60$/MWh thresholds. However, the DNN model is only included in the best-performing set at the -$45/MWh threshold, indicating that the DNN is, overall, the worst-performing model in terms of generated P&L. Sets of best-performing models using the P&L as the loss function do not coincide with these using the log-likelihood that are presented in Section 4.3. Such disparities are mainly explained by the more volatile nature of the P&L in comparison to log-likelihood scores, which makes the discrimination between models more difficult when using the P&L as the performance metric.

The Hansen et al., 2011 model confidence set approach is also implemented to test if each standalone model, for each threshold, statistically outperforms the base case (12 tests in total). In each of the tests, only two models (the considered model and the base case) are initially included in the model confidence set. We then examine whether or not the base case is removed from the set. Results indicate that for the vast majority of tests (10 out of 12 tests), the models statistically outperform the base case for a confidence level of 5% ($\alpha = 5\%$). Only the random forest and the DNN models for the threshold $\gamma^- = -\$30$/MWh do not statistically outperform the base case.

Table 4.5 reports a risk-adjusted performance measure (the Sortino ratio) and two risk metrics (the semi-deviation and the Value-at-Risk) for both strategies as well as the base case.[10] The Sortino ratio is (1) always positive for the two developed strategies, as opposed to that of the base case, which is negative, and (2) large, i.e., always greater than one and most of the time greater than two. Strategy 2 generally produces Sortino ratios larger than those of Strategy 1, indicating that shorting the DART in periods of high spike probability improves the strategy's risk/reward profile. Both the semi-deviation (Std$^-$) and the Value-

[10]The Sortino ratio is computed as follows:

$$\frac{\frac{8760}{\tau}\sum_{t=1}^{\tau} P_t}{\sqrt{\frac{8760}{\tau}\sum_{t=1}^{\tau}(P_t)^2 \mathbb{1}_{\{P_t<0\}}}}$$

where P_t is the hour-t P&L, and τ is the number of hours in the sample. The constant ($365 \times 24 = 8760$) corresponds to the number of hours in a year and is applied for annualization purposes.

at-Risk at confidence level 1% (VaR 1%), are substantially smaller for both Strategy 1 and Strategy 2 than for the base case strategy. Therefore, the results outline that for every model and threshold, Strategy 1 and Strategy 2 lead to significantly larger P&Ls than the base case strategy and embed lesser risk (especially tail risk), thus highlighting the twofold contribution of integrating the model signals to a trading strategy.

Table 4.6 exhibits the precision and recall metrics as well as the dissected aggregated out-of-sample average P&L of Strategy 2. The precision illustrates the ratio of realized predicted spikes over the total number of predicted spikes. The recall considers the proportion of labeled spikes predicted by the models relative to the sample's total number of labeled spikes.[11] The precision is relatively low for every threshold and all models, i.e., between 9% and 16%, which is not surprising considering the low cut-off points used for the predictions. For all models, the recall is approximately 20%, 33%, and 37% for thresholds $\gamma^- = -\$30$/MWh, $\gamma^- = -\$45$/MWh and $\gamma^- = -\$60$/MWh, respectively.

The average P&L associated with the true positives also corroborates this conclusion, where the average P&L increases with the thresholds. However, it is interesting to note that the average P&L for false positives is surprisingly low (between $-\$10$ and $-\$20$). Therefore, high probabilities of observing spikes predicted by the models appear to be associated with periods of high DART spread volatility and uncertainty. Unsurprisingly, the average P&L of false negatives is strongly negative. However, it is worth noting that the average P&L of false negatives is, in absolute terms, much lower than the average P&L of true positives. This result indicates that, on average, the models capture the more significant spikes. The last two rows of Table 4.6 display the average P&L when the models predict a spike (Avg. pos.) and inversely when the models predict no spikes. Both scenarios generate positive P&L for every model and each threshold. The average P&L is much larger when the model predicts a spike. Nonetheless, the results clearly show the potential of integrating

[11] The precision (P) and recall (R) are calculated as follows:

$$P = \frac{Tp}{Tp + Fp}, \qquad R = \frac{Tp}{Tp + Fn},$$

where Tp denotes the True positives, Fp the false positives, and Fn the false negatives. A false positive is a predicted spike that did not materialize.

the model-generated signals into any trading strategy.

			Logistic Regression			Random Forest			Gradient Boosting Trees			DNN		
		γ^-	-30	-45	-60	-30	-45	-60	-30	-45	-60	-30	-45	-60
Strategy 1	2018	Avg.	0.77	1.02	1.07	0.75	1.35	1.55	0.93	1.32	1.15	0.96	0.98	1.04
		Total	6060	7612	7921	6201	10701	12063	7430	10197	8490	7545	7644	7973
	2019	Avg.	-0.29	-0.13	0.01	-0.34	-0.27	0.08	-0.08	-0.18	-0.25	-0.31	-0.03	-0.10
		Total	-2403	-1077	78	-2881	-2266	690	-685	-1449	-2014	-2581	-251	-799
	2020	Avg.	0.40	0.45	0.52	0.04	0.19	0.24	0.28	0.28	0.44	-0.10	0.23	0.26
		Total	2931	3185	3624	364	1522	1887	2246	2138	3280	-760	1628	1810
	2021	Avg.	0.79	0.75	0.99	0.71	0.75	0.33	0.89	0.86	0.85	0.60	0.28	0.08
		Total	4617	4083	5370	4416	4301	1846	5652	4918	4999	3217	1503	415
	Agg.	Avg.	0.38	0.49	0.61	0.26	0.48	0.56	0.48	0.54	0.51	0.25	0.37	0.34
		Total	11205*	13803*	16992*	8099	14258*	16486*	14644*	15804*	14756*	7422	10524*	9399
Strategy 2	2018	Avg.	0.59	0.94	1.01	0.62	1.65	1.96	0.90	1.53	1.14	0.93	0.95	1.02
		Total	5144	8248	8866	5425	14425	17151	7885	13418	10005	8115	8312	8970
	2019	Avg.	-0.10	0.20	0.47	-0.21	-0.07	0.61	0.29	0.12	-0.01	-0.14	0.39	0.27
		Total	-867	1785	4094	-1823	-592	5320	2570	1040	-88	-1223	3437	2342
	2020	Avg.	0.72	0.78	0.88	0.13	0.40	0.48	0.56	0.54	0.80	-0.12	0.42	0.46
		Total	6315	6824	7700	1180	3497	4226	4945	4728	7013	-1067	3709	4072
	2021	Avg.	1.76	1.61	1.97	1.70	1.67	1	2.04	1.84	1.86	1.38	0.91	0.61
		Total	12838	11770	14343	12435	12206	7295	14908	13440	13602	10038	6610	4433
	Agg.	Avg.	0.70	0.85	1.04	0.51	0.88	1.01	0.90	0.97	0.91	0.47	0.66	0.59
		Total	23430*	28626*	35003*	17217	29536*	33992*	30308*	32628*	30532*	15863	22068*	19817
			2018	2019	2020	2021	Agg.							
Base case		Avg.	0.79	-0.45	-0.05	-0.49	-0.03							
		Total	6976	-3939	-453	-3603	-1020							

The table presents P&L statistics for the two trading strategies considered and the base case. The statistics are divided by year (2018, 2019, 2020, and 2021) and combined over the out-of-sample period (2018 to 2021) under the rows named "Aggregated". Both strategies take a long position on the DART spread when the model predicted probability remains under the pre-determined cut-off point, i.e., 12%, 7% and 5% for $\gamma^- = -\$30$/MWh, $\gamma^- = -\$45$/MWh and $\gamma^- = -\$60$/MWh respectively. The first strategy takes no position otherwise, i.e., when the probability does exceed the cut-off point, while the second strategy takes a short position in the DART otherwise. The summary statistics are the total P&L and the average profit per position. The Hansen et al., 2011 confidence set approach, identifying which of the models have a performance that is statistically indistinguishable from that of the best model, is applied to the total P&L for each threshold. The best models remaining in the model confidence set at a level of significance 5% are identified with a star in the table. The testing procedure is described in Appendix C.3.

Table 4.4: Average and total out-of-sample P&L

		Logistic Regression			Random Forest			Gradient Boosting Trees			DNN		
	γ^-	-30	-45	-60	-30	-45	-60	-30	-45	-60	-30	-45	-60
Strategy 1	Sortino ratio	1.64	2.32	3.03	1.03	2.20	2.73	2.05	2.55	2.32	1.12	1.68	1.58
	Std$^-$	21.80	19.80	18.83	23.42	20.25	19.26	21.80	19.88	20.67	21.11	20.55	20.14
	VaR 1%	-80.77	-75.01	-72.07	-92.08	-82.44	-79.31	-85.28	-78.21	-78.34	-83.41	-80.75	-77.14
Strategy 2	Sortino ratio	2.45	3.08	4.01	1.75	3.16	3.65	3.23	3.59	3.26	1.79	2.38	2.06
	Std$^-$	26.69	25.90	24.30	27.47	26.02	25.94	26.11	25.34	26.09	24.75	25.88	26.78
	VaR 1%	-83.48	-77.85	-76.44	-94.72	-85.94	-84.53	-89.09	-82.46	-82.46	-85.24	-82.73	-80.86
Base case	Sortino ratio	-0.09											
	Std-	32.61											
	VaR 1%	-105.63											

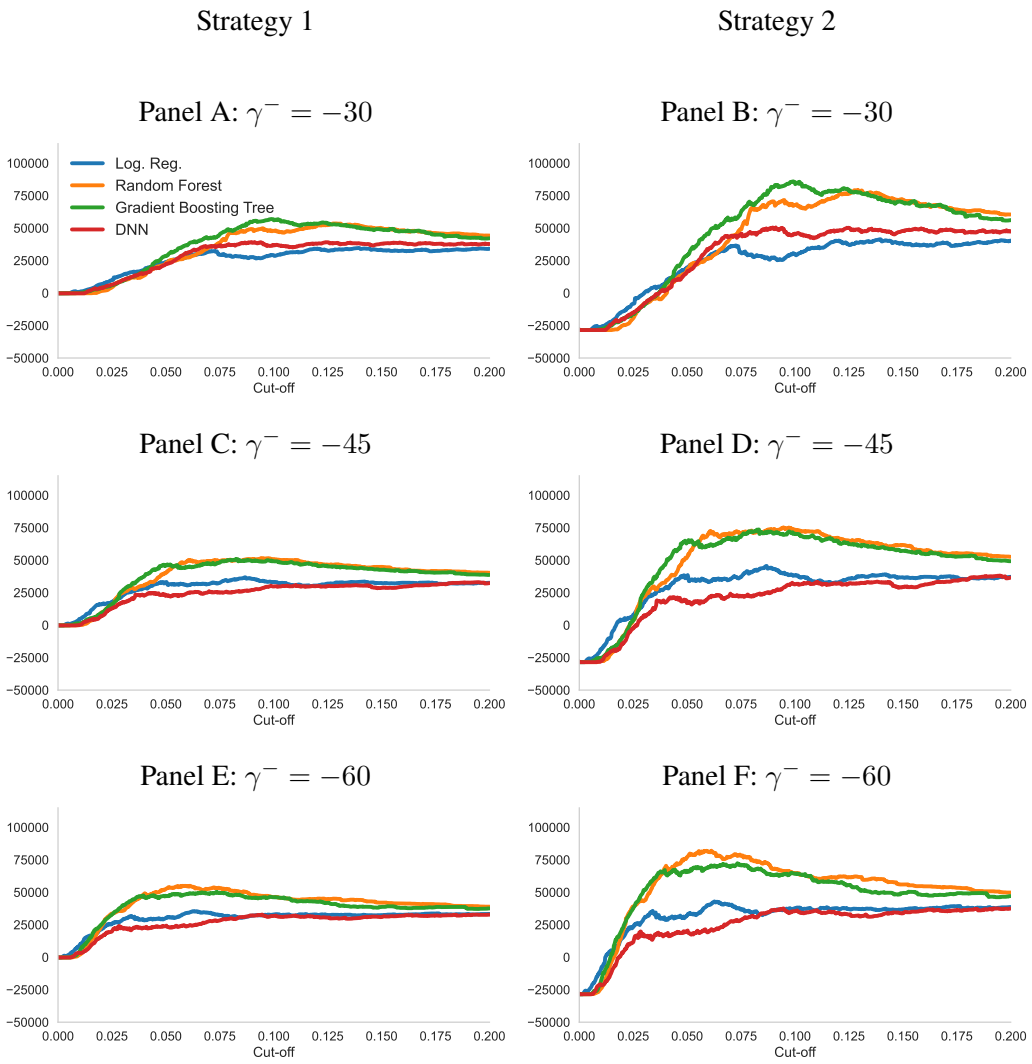
The table reports (1) the Sortino ratio, (2) the semi-deviation (Std.$^-$), and (3) the Value-at-Risk at confidence level 1% (VaR 1%) for hourly P&L of the considered strategies over the out-of-sample period (2018 to 2021). Both Strategies 1 and 2 take a long position on the DART when the model-predicted probability is lower than the pre-determined cut-off point, i.e. 12%, 7% and 5% for $\gamma^- = -\$30$/MWh, $\gamma^- = -\$45$/MWh and $\gamma^- = -\$60$/MWh respectively. The first strategy takes no position otherwise, i.e., when the probability exceeds the cut-off point, while the second strategy takes a short position in the DART otherwise.

Table 4.5: Risk-adjusted metrics for the out-of-sample hourly P&L

		Logistic Regression			Random Forest			Gradient Boosting Trees			DNN		
	γ^-	-30	-45	-60	-30	-45	-60	-30	-45	-60	-30	-45	-60
	Precision	0.16	0.12	0.09	0.16	0.13	0.09	0.18	0.13	0.09	0.16	0.12	0.08
	Recall	0.21	0.34	0.39	0.14	0.24	0.29	0.22	0.34	0.37	0.20	0.33	0.38
Average P&L	True pos.	120.29	136.06	169.37	136.38	152.94	184.01	116.53	138.73	170.27	121.21	141.54	171.93
	False pos.	-17.62	-12.82	-10.03	-18.96	-15.45	-11.95	-17.61	-13.42	-11.59	-15.05	-12.04	-9.03
	False neg.	-70.80	-91.96	-112.44	-72.05	-92.67	-114.39	-71.36	-91.03	-113.01	-71.37	-90.05	-111.78
	True neg.	4.06	3.15	2.61	4.29	3.31	2.69	4.11	3.14	2.52	4.28	3.21	2.68
	Avg Pos.	5.03	5.57	6.00	6.19	6.64	6.17	6.10	5.97	5.39	7.15	6.01	5.89
	Avg neg.	0.37	0.60	0.68	0.30	0.44	0.48	0.43	0.61	0.55	0.51	0.66	0.71

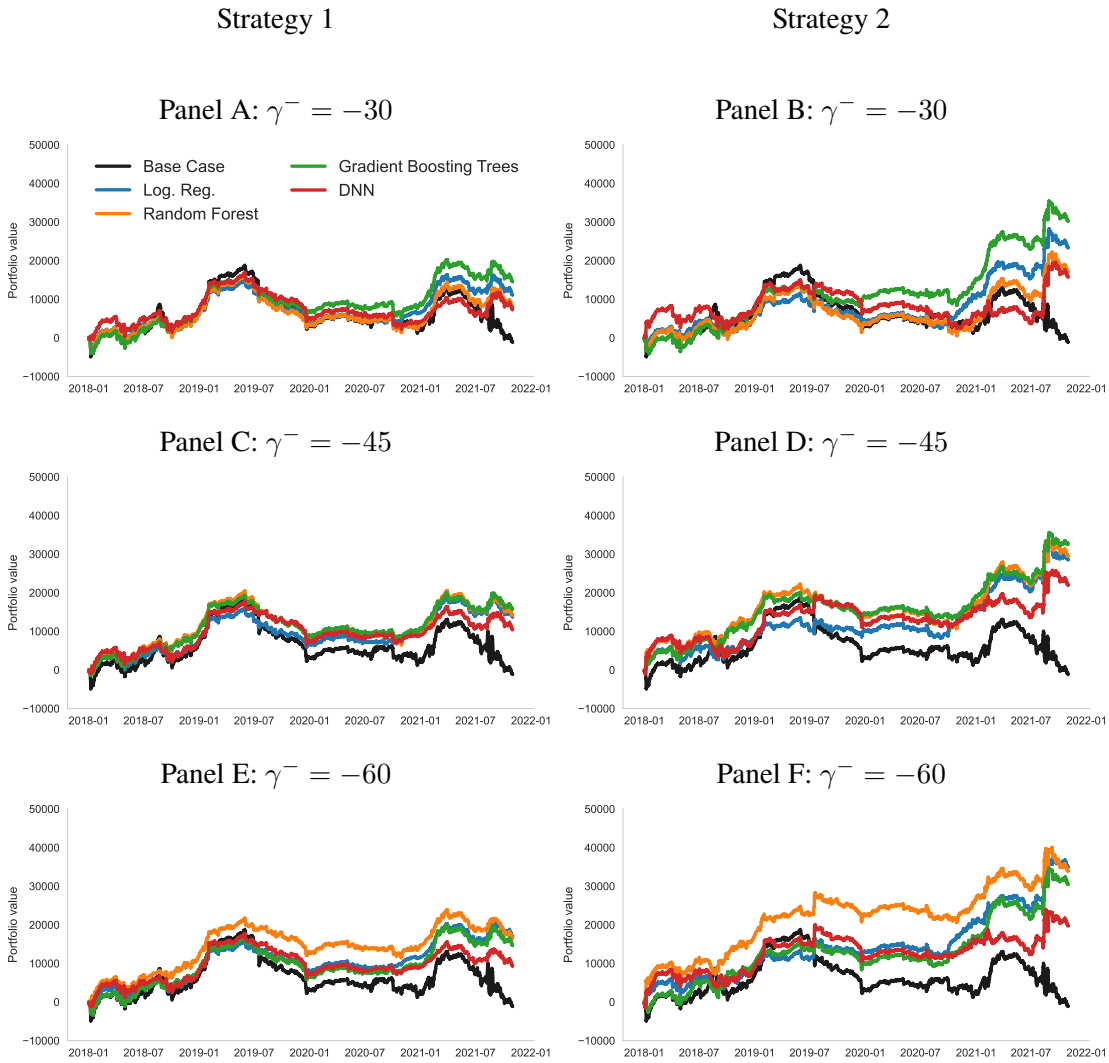
The first two rows of the table exhibit the precision and recall for each model. Rows 3 to 6 illustrate the average P&L associated with either the true positives (True pos.)–predicted spikes that are effectively spikes; false positives (False pos.)–predicted spikes that did not materialize; false negatives (False neg.)–spikes that were not predicted; and true negatives (True neg.). The last two rows detail the average P&L when the model either predicts a spike (Avg pos.) or no spike (Avg neg.). Results are computed over the out-of-sample period (2018 to 2021). The cut-off point for thresholds $\gamma^- = -\$30$/MWh, $\gamma^- = -\$45$/MWh and $\gamma^- = -\$60$/MWh are respectively 12%, 7% and 5%.

Table 4.6: Precision/Recall and Strategy 2 hourly P&L dissected



Each panel displays each model's total P&L as a function of the cut-off value over the training set spanning from 2015 to 2017. Strategy 1 (left panels) and 2 (right panels) take a long position in the DART spread if the predicted spike probability is below the cut-off point. While Strategy 1 takes no position when the spike probability is above the cut-off point, Strategy 2 takes a short position in such case. The volume of the positions taken for both strategies is always 1 MWh.

Figure 4.12: Total P&L for a continuum of cut-off values



Each panel illustrates the time series of a portfolio value starting at \$0 following strategies 1 and 2 over the out-of-sample period (2018 to 2021). Strategy 1 (left panels) and 2 (right panels) take a long position in the DART spread if the predicted spike probability is below the cut-off point. While Strategy 1 takes no position when the spike probability is above the cut-off point, Strategy 2 takes a short position in such a case. The cut-off point for threshold $\gamma^- = -\$30$/MWh, $\gamma^- = -\$45$/MWh and $\gamma^- = -\$60$/MWh are respectively 12%, 7% and 5%. The volume of the positions taken for both strategies is always 1 MWh. The base case strategy consists in taking a long position each period.

Figure 4.13: Portfolio value over time

4.5 Conclusion

This paper studies the forecasting of DART spread spike probabilities. DART spreads of the Long Island zone of the NYISO market are considered in developing the model. A fixed threshold methodology commonly used in the literature is applied to identify spikes in the data. A tailor-made feature set is proposed to perform predictions.

Four statistical learning approaches are considered in predicting spike occurrences. Results indicate that for every threshold considered, all models produce similar predictive performance, although the gradient boosting trees slightly outperform their counterparts. A feature importance assessment highlights the critical importance of forward-looking features such as the load forecast over transfer capacity ratio, predicted heating degree days and predicted cooling degree days, of seasonal features such as hourly and monthly indicators, and of the backward-looking feature counting the number of spikes in the last 24 hours before the prediction. Conversely, the weekly cycle indicators and some of the backward-looking features measuring near-past load or price prediction errors exhibit lesser importance.

Finally, a backtest of two trading strategies integrating model-generated spike probabilities as a market signal is implemented. Such strategies are shown to produce significantly higher profits, lesser risk, and thus larger risk-adjusted performance in comparison to a base case strategy systematically holding long DART spread positions. Therefore, results highlight the added value of the developed signal from an economic perspective.

Future questions worth examining include: (1) applying the prediction scheme to positive DART spikes, (2) determining if the framework developed works well in other nodes of the grid and other power markets, and (3) designing trading strategies where trade volumes are modulated based on the intensity of the spike probability signal to improve profitability, or where the cut-off point driving the trade direction varies depending on the season.

References

Amjady, N., & Keynia, F. (2010). Electricity market price spike analysis by a hybrid data model and feature selection technique. *Electric Power Systems Research*, *80*(3), 318–327.

Anderson, C., & Davison, M. (2008). A hybrid system-econometric model for electricity spot prices: Considering spike sensitivity to forced outage distributions. *IEEE Transactions on Power Systems*, *23*(3), 927–937.

Benth, F. E., Kallsen, J., & Meyer-Brandis, T. (2007). A non-Gaussian Ornstein–Uhlenbeck process for electricity spot price modeling and derivatives pricing. *Applied Mathematical Finance*, *14*(2), 153–169.

Cartea, A., & Figueroa, M. G. (2005). Pricing in electricity markets: A mean reverting jump diffusion model with seasonality. *Applied Mathematical Finance*, *12*(4), 313–335.

Christensen, T. M., Hurn, A. S., & Lindsay, K. A. (2012). Forecasting spikes in electricity prices. *International Journal of Forecasting*, *28*(2), 400–411.

Christensen, T., Hurn, S., & Lindsay, K. (2009). It never rains but it pours: Modeling the persistence of spikes in electricity prices. *The Energy Journal*, *30*(1).

Clements, A., Herrera, R., & Hurn, A. (2015). Modelling interregional links in electricity price spikes. *Energy Economics*, *51*, 383–393.

Conejo, A. J., Contreras, J., Espínola, R., & Plazas, M. A. (2005). Forecasting electricity prices for a day-ahead pool-based electric energy market. *International Journal of Forecasting*, *21*(3), 435–462.

Covert, I., Lundberg, S. M., & Lee, S.-I. (2020). Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, *33*, 17212–17223.

Eichler, M., Grothe, O., Manner, H., & Tuerk, D. (2014). Models for short-term forecasting of spike occurrences in Australian electricity markets: A comparative study. *Journal of Energy Markets*, *7*(1).

Fan, J.-L., Hu, J.-W., & Zhang, X. (2019). Impacts of climate change on electricity demand in China: An empirical estimation based on panel data. *Energy*, *170*, 880–888.

Gudkov, N., & Ignatieva, K. (2021). Electricity price modelling with stochastic volatility and jumps: An empirical investigation. *Energy Economics*, *98*, 105260.

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453–497.

He, D., & Chen, W.-P. (2016). A real-time electricity price forecasting based on the spike clustering analysis. *2016 IEEE/PES Transmission and Distribution Conference and Exposition (T&D)*, 1–5.

Herrera, R., & González, N. (2014). The modeling and forecasting of extreme events in electricity spot markets. *International Journal of Forecasting*, *30*(3), 477–490.

Itron. (2008). *New York ISO climate change impact study* (tech. rep.). New York Independent System Operator.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning : With applications in R* (Vol. 112). Springer. https : / / doi . org / https : //doi.org/10.1007/978-1-4614-7138-7

Janczura, J., Trück, S., Weron, R., & Wolff, R. C. (2013). Identifying spikes and seasonal components in electricity spot price data: A guide to robust modeling. *Energy Economics*, *38*, 96–110.

Klüppelberg, C., Meyer-Brandis, T., & Schmidt, A. (2010). Electricity spot price modelling with a view towards extreme spike risk. *Quantitative Finance*, *10*(9), 963–974.

Lago, J., De Ridder, F., & De Schutter, B. (2018). Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms. *Applied Energy*, *221*, 386–405.

Lago, J., Marcjasz, G., De Schutter, B., & Weron, R. (2021). Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark. *Applied Energy*, *293*, 116983.

Longstaff, F. A., & Wang, A. W. (2004). Electricity forward prices: A high-frequency empirical analysis. *The Journal of Finance*, *59*(4), 1877–1900.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Manner, H., Türk, D., & Eichler, M. (2016). Modeling and forecasting multivariate electricity price spikes. *Energy Economics*, *60*, 255–265.

NYISO. (2016). *NYISO standard template presentation to market participants* (tech. rep.). New York Independent System Operator.

Sandhu, H. S., Fang, L., & Guan, L. (2016). Forecasting day-ahead price spikes for the Ontario electricity market. *Electric Power Systems Research*, *141*, 450–459.

Shapley, L. S. (2016). 17. A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the theory of games (am-28), volume ii* (pp. 307–318). Princeton University Press. https://doi.org/doi:10.1515/9781400881970-018

Trueck, S., Weron, R., & Wolff, R. (2007). *Outlier treatment and robust approaches for modeling electricity spot prices* (MPRA Paper). University Library of Munich, Germany. https://EconPapers.repec.org/RePEc:pra:mprapa:4711

Weron, R. (2007). *Modeling and forecasting electricity loads and prices: A statistical approach* (Vol. 403). John Wiley & Sons.

Weron, R., & Misiorek, A. (2008). Forecasting spot electricity prices: A comparison of parametric and semiparametric time series models. *International Journal of Forecasting*, *24*(4), 744–763.

Yi-Ling, H., Hai-Zhen, M., Guang-Tao, D., & Jun, S. (2014). Influences of urban temperature on the electricity consumption of Shanghai. *Advances in Climate Change Research*, *5*(2), 74–80.

Zahedi, G., Azizi, S., Bahadori, A., Elkamel, A., & Alwi, S. R. W. (2013). Electricity demand estimation using an adaptive neuro-fuzzy network: A case study from the Ontario province–Canada. *Energy*, *49*, 323–328.

Zhang, J., Tan, Z., & Yang, S. (2012). Day-ahead electricity price forecasting by a new hybrid method. *Computers & Industrial Engineering*, *63*(3), 695–701.

Ziel, F., Steinert, R., & Husmann, S. (2015). Forecasting day ahead electricity spot prices: The impact of the EXAA to other European electricity markets. *Energy Economics*, *51*, 430–444.

Chapter 5

Concluding Remarks

The implied volatility surface offers a rich source of information about the participants' view of the market's future conditions.

Volatility and extreme events are present in the distribution of all asset returns. Estimating, quantifying and managing the inherent risk of these assets represent a major challenge for market participants. To better quantify the risk of the assets, the extraction of information from the market is paramount. The implied volatility represents one such source of information. This thesis's first two essays concentrate on leveraging the information content of the implied volatility surface while the last essay departs from the derivatives market and concentrates on the electricity market.

In the first essay, the static implied volatility surface is calibrated using a factor model. The factors adequately capture the moneyness and maturity slopes, the smile attenuation, and the smirk of the implied volatility surface while being economically interpretable. The factor representation of the surface is twice differentiable, is asymptotically well-behaved and allows for interpolating and extrapolating the surface. The factor model can be used to mark-to-market illiquid derivatives since the model can generate coherent and clean interpolations of the implied volatility surface. Other applications related to derivatives pricing and asset pricing are explored. The benefits of a smoothed implied volatility surface are illustrated through the extraction of the risk-neutral density and risk-neutral moments and the calculation of option price sensitivities.

The second essay leverages the implied volatility factor decomposition previously developed to construct the joint implied volatility and return (JIVR) model. The JIVR model is a characterization of the joint dynamics of the S&P 500 index and of its associated implied volatility surface. It integrates the whole implied volatility surface as input to the model. This approach allows for novel characterization of the underlying log-returns and the implied volatility surface level. The JIVR model can efficiently forecast the distribution of any portfolio of options. The capabilities of the JIVR model are exhibited with two exercises. Firstly, the risk metrics of straddle and strangle positions are computed and backtested. Secondly, the predictive performance of the JIVR model is compared to a conventional time series counterpart model to forecast the VIX index distribution. The contribution suggests that further investigation should be invested in this strand of the literature.

The last essay concentrates on the electricity market, which is well-known for its volatility. In this essay, the perspective of virtual bidders, who are exposed to DART spreads, is considered. The occurrence of DART spikes has important risk implications for these participants. To improve the risk-reward profile of their trading strategy, a tailored feature set combined with statistical learning methods are leveraged to forecast these extreme price events. The models' performance is first assessed with standard statistical performance metrics. Results show the capabilities of the four considered statistical learning models to predict the DART spikes. To evaluate the model performance from an economical point of view, a base case trading strategy is compared to two alternative strategies that integrate the models' generated signals. The two alternative trading strategies generate statistically larger profits and embed substantially less downside risk than the base case strategy. The methodology developed could easily be extended to other zones of the NYISO and other electricity markets. Furthermore, forecasting positive DART spikes is an exciting avenue to reduce the risk further and increase the profitability of such strategies.

Chapter A

Appendices of *Venturing into Uncharted Territory: An Extensible Parametric Implied Volatility Surface Model*

A.1 Principal component analysis

Directly applying PCA to the option sample is not possible since option numbers and characteristics (moneyness and time-to-maturity) vary from day to day; PCA requires a stable sample every day. To circumvent this issue, a grid with respect to moneyness and time-to-maturity is constructed. For each point of the grid, the implied volatility can be interpolated using quoted options IV. The interpolation scheme can be achieved through a variety of methods. As in Israelov and Kelly, 2017, a spline interpolation scheme is implemented using the MATLAB fit function with “thinplateinterp" fit type.

Because quoted moneyness levels vary greatly from day to day depending on the market conditions, the grid covers a smaller surface than the one spanned by the quoted moneyness and maturities. Two grids are used in this paper. In Figure 2.2 the moneyness M varies between -0.2 and 0.6 by increments of 0.1 and the days-to-maturity are 30, 60, 91, 122, 152, 182, 270 and 365 days. The other figures and tables of Appendix A.1 are based on the moneyness definition and the grid of Israelov and Kelly, 2017, that is, the moneyness M^* is between -2 and 1 with increments of 0.25 and the days-to-maturity are 30, 60, 91,

122, 152, 182, 270, and 365 days-to-maturity. Because M^* depends on the VIX value, the two grids do not include the same option subsample, especially during financial turmoil.

The PCA is constructed from the sample covariance matrix of the interpolated IV time series. The right panels of Figure 2.2 display the five factors with the greatest explanatory power. Factor 1 can be interpreted as the level, factor 2 is the time-to-maturity slope, factor 3 corresponds to the moneyness slope, factor 4 is a curvature factor and factor five is the smirk. The first factor explains 94.84% of the surface variations.

A.2 Bayesian regression

Model (2.2) is a linear function of the factors, which suggests that the ordinary least square (OLS) estimation approach is straightforward. However, as documented in Gauthier and Simonato, 2012 in the alternative context of zero-coupon yields, there can be several sets of parameters which produce very similar surfaces. To preserve the economic interpretation of each factor, the least squares method is coupled with a Bayesian approach for regularization purposes, thereby avoiding erratic behaviour in parameter time series.

A typical linear model can be expressed as $Y = X\beta + \epsilon$. A linear system incorporates prior information as follows:

$$\begin{bmatrix} Y \\ \beta_{\text{prior}} \end{bmatrix} = \begin{bmatrix} X \\ R \end{bmatrix} \beta + \begin{bmatrix} \epsilon \\ \delta \end{bmatrix} \tag{A.1}$$

where Y represents the observations, β_{prior} the prior's expectation, X the factors, R the matrix linking the parameters to the priors, and (ϵ, δ) the vector of errors which follows a multivariate normal distribution with a diagonal covariance matrix

$$\Omega = \begin{bmatrix} \Sigma_\epsilon & 0 \\ 0 & \Sigma_\delta \end{bmatrix}.$$

The generalized least squares estimator of β with prior information is

$$\hat{\beta} = \left(\begin{bmatrix} X \\ R \end{bmatrix}^\top \Omega^{-1} \begin{bmatrix} X \\ R \end{bmatrix} \right)^{-1} \begin{bmatrix} X \\ R \end{bmatrix}^\top \Omega^{-1} \begin{bmatrix} Y \\ \beta_{\text{prior}} \end{bmatrix}.$$

The value of Σ_ϵ can be estimated using the ordinary least squares estimator (OLS) without any priors and the matrix Σ_δ is a hyperparameter that controls the prior distribution. The smaller the values on the diagonal of Σ_δ are, the closer to the prior expected value the estimated β is.

The observed 1-year ATM IV ($\text{ATM}_{1y,t}$) serves as prior mean for $\beta_{1,t}$. The slope prior is constructed from the one-month ATM IV ($\text{ATM}_{1m,t}$):

$$\text{Slope}_t = \frac{\text{ATM}_{1y,t} - \text{ATM}_{1m,t}}{\exp\left(-\sqrt{4/12}\right)}.$$

The priors for $\beta_{3,t}$ and $\beta_{5,t}$ are the previous day estimates $\beta_{3,t-1}$ and $\beta_{5,t-1}$, respectively. Due to its interconnectedness with the other parameters, no prior is assigned to $\beta_{4,t}$. This entails setting

$$\beta_{\text{prior}} = \begin{bmatrix} \text{ATM}_{1y,t} \\ \text{Slope}_t \\ \beta_{3,t-1} \\ \beta_{5,t-1} \end{bmatrix}, R = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \Sigma_\delta = \begin{bmatrix} 0.38 & 0 & 0 & 0 & 0 \\ 0 & 5.60 & 0 & 0 & 0 \\ 0 & 0 & 0.73 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \times 10^{-4}.$$

The prior variances need to be set. For the first two priors associated with the long-term level and the slope, the sample variance of the observable ($\text{ATM}_{1y,t}$ and Slope_t) is considered. The β_3 prior's variance is the sample variance of the proxy variable for the moneyness slope, namely the one-month ATM IV minus the one-month IV with moneyness $M = 0.4$. Finally, for the last prior of β_5, based on judgmental consideration, the standard deviation is set to half the parameter's level, resulting in a prior variance of 1×10^{-4}. Because the prior variances are large, the estimation procedure has extensive leeway to match the option data, while keeping the economic interpretation of the coefficients due to regularization removing large erratic movements in time series of parameter estimates.

A.3 Benchmarking with a non-parametric PCA approach

In Section 2.3.3, the calibration performance of Model (2.2) is compared to that of two parametric benchmarks. In this section, the non-parametric PCA approach of Israelov and

Kelly, 2017, hereafter IK, is instead considered for benchmarking. As mentioned by the authors, their methodology is best suited to represent densely populated regions of the surface. When applied to our dataset, their model is fitted to a sub-surface inside which only 41% of observed options lie, leaving aside useful information about extreme movement expectations contained in the deep OTM options. Furthermore, PCA methods cannot extrapolate beyond the restricted grid to generate prices for out-of-sample deep OTM options. The main conclusion of this section is that on this restricted sample for which the PCA approach is optimal, Model (2.2) does almost as well, while allowing for IV interpolation and extrapolation on the wider IV surface.

Israelov and Kelly, 2017 rely on an alternative definition of moneyness that is proportional to the 30-day VIX level at day *t*,

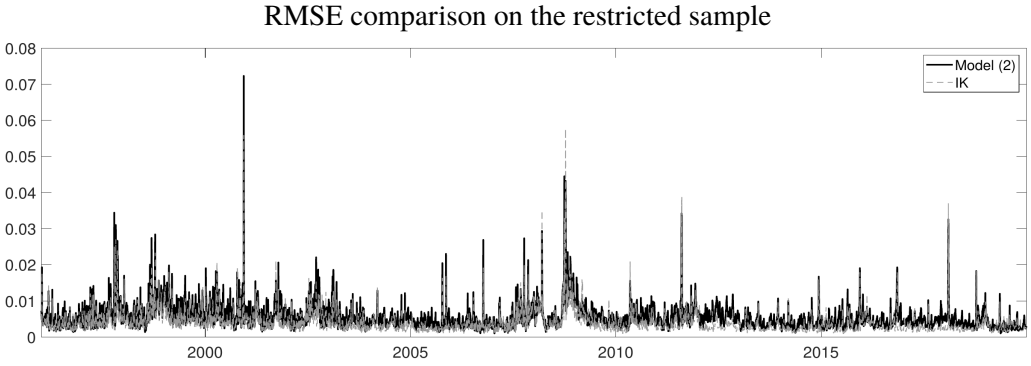
$$M^* = \frac{\log(K/S)}{\text{VIX}_t\sqrt{(\tau)}}.$$

Such specification allows the model to adjust the grid relied upon by the PCA to the prevalent market state, i.e., to include more OTM options in periods of market turmoil. Indeed, to perform a PCA over daily IV surfaces, Israelov and Kelly, 2017 construct a synthetic IV surface on a pre-defined grid covering time-to-maturity values of 30, 60, 91, 122, 152, 182, 273, 365 days, and moneyness M^* values varying between -2 and 1 with increments of 0.25. They interpolate the grid points from available options IV with a thin plate spline. The factors extracted from the PCA represent the most efficient linear decomposition to minimize the squared fitting errors.

The fitting performance is assessed by computing the daily root mean square error (RMSE) between the observed IV and the corresponding fitted values.

Figure A.1 shows that the daily RMSE is very similar between the PCA approach and the specification (2.2) over the restricted sample. Such an outcome was expected due to the similarity of Model (2.2) and PCA factors highlighted by Figure 2.2.

Table A.1 confirms that the ARMSEs of the PCA approach and Model (2.2) are very close within each bucket of moneyness and time-to-maturity. As expected, the PCA approach slightly outperforms Model (2.2) for all buckets because, by construction, the PCA designs factors so as to minimize mean squared discrepancies between data and fitted val-



Model (2.2) is estimated on a restricted sample corresponding to options with a time-to-maturity of between 30 and 365 days, and a M^* moneyness of between -2 to 1.

Figure A.1: RMSE by bucket over the restricted sample

ues. It is reassuring to see that Model (2.2) has a very similar performance to the PCA, while being able to fit over a much larger IV surface and even extrapolate beyond the observable IV.

Model	$M \leq -0.1$	$-0.1 < M \leq 0.1$	$M > 0.1$	All
Model (2.2)	0.0080	0.0059	0.0057	0.0063
IK	0.0060	0.0045	0.0054	0.0052
Number of options	231,356	606,327	587,042	1,424,725
Model	$\tau \leq 60$	$60 < \tau \leq 180$	$\tau > 180$	All
Model (2.2)	0.0089	0.0036	0.0058	0.0063
IK	0.0082	0.0030	0.0031	0.0052
Number of options	435,642	585,176	403,907	1,424,725

The average RMSE over time is reported for each bucket of moneyness (M) and days to maturity (τ). The sample period is January 4, 1996, to December 31, 2019. The restricted sample corresponds to options with a time-to-maturity of between 30 and 365 days and a M^* moneyness of between -2 to 1. Model (2.2) is re-estimated on the restricted sample in Panel B.

Table A.1: Average RMSE over time from IV surface estimation over the restricted sample

A.4 Butterfly and calendar spreads

On a given trading day t, let $D_\tau = \exp(-r_\tau \tau)$ be the risk-free discount factor where r_τ is the OptionMetrics zero-coupon interest rate associated with the maturity τ. For any traded maturity τ, denote by $K_0^{(\tau)} < K_1^{(\tau)} < \ldots < K_{n(\tau)}^{(\tau)}$ the set of all strikes for which quotes are provided in the dataset. Thus, $n(\tau)$ represents the number of available options with time-to-maturity τ. $C(K, \tau)$ denotes the call price for time-to-maturity τ and strike price K.

As outlined in Davis and Hobson, 2007, the butterfly spread value $BS_i^{(\tau)}$, $i = 1, \ldots, n(\tau)-1$, defined as

$$\begin{aligned} BS(K_i^\tau, \tau) = &\frac{C(K_{i+1}^{(\tau)}, \tau)}{\left(K_{i+1}^{(\tau)} - K_i^{(\tau)}\right) D_\tau} + \frac{C(K_{i-1}^{(\tau)}, \tau)}{\left(K_i^{(\tau)} - K_{i-1}^{(\tau)}\right) D_\tau} \\ &- \frac{C(K_i^{(\tau)}, \tau)}{D_\tau} \left(\frac{1}{K_i^{(\tau)} - K_{i-1}^{(\tau)}} + \frac{1}{K_{i+1}^{(\tau)} - K_i^{(\tau)}} \right) \end{aligned}$$

needs to be positive, as otherwise, a butterfly spread arbitrage opportunity would arise.[1]

The calendar spread value $CS_i^{(\tau)}$ is defined as

$$\begin{aligned} CS_i^{(\tau)}(\tau_1, \tau_2, i_1, i_2) = &C(K_i^{(\tau)}, \tau) - \left(\frac{K_{i_2}^{(\tau_2)}/F_{0,\tau_2} - K_i^{(\tau)}/F_{0,\tau}}{K_{i_2}^{(\tau_2)}/F_{0,\tau_2} - K_{i_1}^{(\tau_1)}/F_{0,\tau_1}} \right) \frac{D_\tau F_{0,\tau}}{D_{\tau_1} F_{0,\tau_1}} C(K_{i_1}^{(\tau_1)}, \tau_1) \\ &- \left(1 - \left(\frac{K_{i_2}^{(\tau_2)}/F_{0,\tau_2} - K_i^{(\tau)}/F_{0,\tau}}{K_{i_2}^{(\tau_2)}/F_{0,\tau_2} - K_{i_1}^{(\tau_1)}/F_{0,\tau_1}} \right) \right) \frac{D_\tau F_{0,\tau}}{D_{\tau_2} F_{0,\tau_2}} C(K_{i_1}^{(\tau_2)}, \tau_2). \end{aligned}$$

where $\tau_1,\ \tau_2 > \tau$ and i_1, i_2 are such that the strike prices satisfy $\frac{K_{i_1}^{(\tau_1)}}{F_{0,\tau_1}} \leq \frac{K_i^{(\tau)}}{F_{0,\tau}} \leq \frac{K_{i_2}^{(\tau_2)}}{F_{0,\tau_2}}$. If $CS_i^{(\tau)}(\tau_1, \tau_2, i_1, i_2) \leq 0$, then there is a calendar spread arbitrage opportunity in the data.

There are many combinations of maturities and strike prices that satisfy the calendar spread restrictions. In Section 2.3.5, only one calendar spread per option in the sample is

[1] The butterfly spreads are computed differently at extremities of the strike price set:

$$BS(K_0^\tau, \tau) = 1 - \frac{C(K_0^{(\tau)}, \tau) - C(K_1^{(\tau)}, \tau)}{\left(K_1^{(\tau)} - K_0^\tau\right) D_\tau} \text{ and } BS(K_{n(\tau)}^{(\tau)}, \tau) = \frac{C(K_{n(\tau)-1}^{(\tau)}, \tau) - C(K_{n(\tau)}^{(\tau)}, \tau)}{\left(K_{n(\tau)}^{(\tau)} - K_{n(\tau)-1}^{(\tau)}\right) D_\tau}.$$

tested: for $C(K_i^{(\tau)}, \tau)$, τ_1 and τ_2 are the smallest maturities greater than τ and the ratios $\frac{K_{i_1}^{(\tau_1)}}{F_{0,\tau_1}}$ and $\frac{K_{i_2}^{(\tau_2)}}{F_{0,\tau_2}}$ are the closest to $\frac{K_i^{(\tau)}}{F_{0,\tau}}$. When no such strike and maturity combinations are available, no calendar spread test is performed for that particular option.

A.5 Carr-Madan formula

This appendix contains the proof of Equation (2.7). The Carr and Madan, 2001 formula states that for a twice differentiable payoff function f,

$$\begin{aligned}\mathrm{E}^{\mathbb{Q}}\left[\mathrm{e}^{-r\tau} f\left(S_\tau\right)\right] =& f\left(k\right)\mathrm{e}^{-r\tau} + f'\left(k\right)\left[C\left(k\right) - P\left(k\right)\right] \\ &+ \int_0^k f''\left(K\right) P\left(K\right) dK + \int_k^\infty f''\left(K\right) C\left(K\right) dK\end{aligned} \tag{A.2}$$

where C and P denote the put and call prices written on the underlying index S_τ with maturity τ and strike price K. Due to the put-call parity, the second term vanishes if $k = F_{0,\tau}$. In our framework, the moneyness $M = \left(\log F_{0,\tau} - \log K\right)/\sqrt{\tau}$ can be inverted to retrieve the strike price $K = F_{0,\tau}\mathrm{e}^{-\sqrt{\tau}M}$. Since $dK = -\sqrt{\tau}F_{0,\tau}\mathrm{e}^{-\sqrt{\tau}M}dM$, applying a change of variable in Equation (A.2), assuming that $k = F_{0,\tau}$, leads to Equation (2.7). □

A.6 Greeks and other partial derivatives

This appendix establishes Equations (2.9), (2.10) and (2.11). The functions' arguments are omitted to simplify the notation. Recall that the call price is

$$c = \mathrm{e}^{-r\tau}F_{0,\tau}\left(\Phi\left(\delta_1\right) - \mathrm{e}^{-M\sqrt{\tau}}\Phi\left(\delta_2\right)\right)$$

with $F_{0,\tau} = S_0\exp\left(\left(r-q\right)\tau\right)$, $M = \frac{1}{\sqrt{\tau}}\ln\frac{F_{0,\tau}}{K}$, $\delta_1 = \frac{M}{\sigma} + \frac{1}{2}\sigma\sqrt{\tau}$, $\delta_2 = \delta_1 - \sigma\sqrt{\tau}$, and σ is the implied volatility from Model (2.2).

A.6.1 Risk-neutral density function

Equation (2.9) is derived in this appendix. Note that $\frac{\partial F}{\partial S} = \mathrm{e}^{(r-q)\tau}$, $\frac{\partial M}{\partial F} = \frac{1}{\sqrt{\tau}}\frac{1}{F_{0,\tau}}$,

$$\frac{\partial\delta_1}{\partial M} = \frac{1}{\sigma} - \left(\frac{M}{\sigma^2} - \frac{1}{2}\sqrt{\tau}\right)\frac{\partial\sigma}{\partial M} \text{ and } \frac{\partial\delta_2}{\partial M} = \frac{1}{\sigma} - \left(\frac{M}{\sigma^2} + \frac{1}{2}\sqrt{\tau}\right)\frac{\partial\sigma}{\partial M} = \frac{\partial\delta_1}{\partial M} - \sqrt{\tau}\frac{\partial\sigma}{\partial M}.$$

Moreover,

$$\begin{aligned}\varphi(\delta_2) =& \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}\left(\delta_1-\sigma\sqrt{\tau}\right)^2} = \varphi(\delta_1)\,\mathrm{e}^{\sigma\sqrt{\tau}\delta_1-\frac{1}{2}\sigma^2\tau}\\ =& \varphi(\delta_1)\,\mathrm{e}^{\sigma\sqrt{\tau}\left(\frac{M}{\sigma}+\frac{1}{2}\sigma\sqrt{\tau}\right)-\frac{1}{2}\sigma^2\tau} = \varphi(\delta_1)\,\mathrm{e}^{\sqrt{\tau}M}.\end{aligned}$$

Because $\frac{\partial M}{\partial K} = -\frac{1}{\sqrt{\tau}K}$,

$$\begin{aligned}\frac{\partial c}{\partial K} &= \mathrm{e}^{-r\tau}F_{0,\tau}\left(\Phi(\delta_1) - \mathrm{e}^{-M\sqrt{\tau}}\Phi(\delta_2)\right)\\ &= \mathrm{e}^{-r\tau}F_{0,\tau}\frac{\partial M}{\partial K}\left(\varphi(\delta_1)\frac{\partial \delta_1}{\partial M} + \sqrt{\tau}\mathrm{e}^{-M\sqrt{\tau}}\Phi(\delta_2) - \mathrm{e}^{-M\sqrt{\tau}}\varphi(\delta_2)\frac{\partial \delta_2}{\partial M}\right)\\ &= \mathrm{e}^{-r\tau}F_{0,\tau}\frac{\partial M}{\partial K}\left(\varphi(\delta_1)\frac{\partial \delta_1}{\partial M} + \sqrt{\tau}\mathrm{e}^{-M\sqrt{\tau}}\Phi(\delta_2) - \varphi(\delta_1)\left(\frac{\partial \delta_1}{\partial M} - \sqrt{\tau}\frac{\partial \sigma}{\partial M}\right)\right)\\ &= -\mathrm{e}^{-r\tau}\frac{F_{0,\tau}}{K}\left(\mathrm{e}^{-M\sqrt{\tau}}\Phi(\delta_2) + \varphi(\delta_1)\frac{\partial \sigma}{\partial M}\right)\\ &= -\mathrm{e}^{-r\tau}\left(\Phi(\delta_2) + \mathrm{e}^{M\sqrt{\tau}}\varphi(\delta_1)\frac{\partial \sigma}{\partial M}\right).\end{aligned}$$

The Gaussian density function satisfies $\frac{\partial \varphi}{\partial z}(z) = -z\varphi(z)$. Therefore, the risk-neutral density function is

$$\begin{aligned}\mathrm{e}^{r\tau}\frac{\partial^2 c}{\partial K^2} =& \frac{\partial M}{\partial K}\frac{\partial}{\partial M}\left(\Phi(\delta_2) + \mathrm{e}^{M\sqrt{\tau}}\varphi(\delta_1)\frac{\partial \sigma}{\partial M}\right)\\ =& \frac{\partial M}{\partial K}\left(\varphi(\delta_2)\frac{\partial \delta_2}{\partial M} + \sqrt{\tau}\mathrm{e}^{M\sqrt{\tau}}\varphi(\delta_1)\frac{\partial \sigma}{\partial M} - \mathrm{e}^{M\sqrt{\tau}}\delta_1\varphi(\delta_1)\frac{\partial \delta_1}{\partial M}\frac{\partial \sigma}{\partial M} + \mathrm{e}^{M\sqrt{\tau}}\varphi(\delta_1)\frac{\partial^2 \sigma}{\partial M^2}\right)\\ =& \frac{\partial M}{\partial K}\left(\varphi(\delta_1)\mathrm{e}^{\sqrt{\tau}M}\left(\frac{\partial \delta_1}{\partial M} - \sqrt{\tau}\frac{\partial \sigma}{\partial M}\right)\right.\\ &\left. + \sqrt{\tau}\mathrm{e}^{M\sqrt{\tau}}\varphi(\delta_1)\frac{\partial \sigma}{\partial M} - \mathrm{e}^{M\sqrt{\tau}}\delta_1\varphi(\delta_1)\frac{\partial \delta_1}{\partial M}\frac{\partial \sigma}{\partial M} + \mathrm{e}^{M\sqrt{\tau}}\varphi(\delta_1)\frac{\partial^2 \sigma}{\partial M^2}\right).\\ =& \frac{\partial M}{\partial K}\mathrm{e}^{\sqrt{\tau}M}\varphi(\delta_1)\left(\frac{\partial \delta_1}{\partial M} - \delta_1\frac{\partial \delta_1}{\partial M}\frac{\partial \sigma}{\partial M} + \frac{\partial^2 \sigma}{\partial M^2}\right)\\ =& \frac{F_{0,\tau}}{K^2}\frac{\varphi(\delta_1)}{\sqrt{\tau}}\left(\frac{\partial \delta_1}{\partial M} - \delta_1\frac{\partial \delta_1}{\partial M}\frac{\partial \sigma}{\partial M} + \frac{\partial^2 \sigma}{\partial M^2}\right)\end{aligned}$$

The density integrates to one since

$$\int_0^\infty \mathrm{e}^{r\tau}\frac{\partial^2 c}{\partial K^2}dK = \mathrm{e}^{r\tau}\left(\lim_{K\to\infty}\frac{\partial c}{\partial K} - \lim_{K\to 0}\frac{\partial c}{\partial K}\right)$$

$$
\begin{aligned}
&= -\lim_{M\to-\infty}\left(\Phi\left(\delta_2\right)+\mathrm{e}^{M\sqrt{\tau}}\varphi\left(\delta_1\right)\frac{\partial\sigma}{\partial M}\right)+\lim_{M\to\infty}\left(\Phi\left(\delta_2\right)+\mathrm{e}^{M\sqrt{\tau}}\varphi\left(\delta_1\right)\frac{\partial\sigma}{\partial M}\right)\\
&= \lim_{M\to\infty}\Phi\left(\delta_2\right)+\lim_{M\to\infty}\mathrm{e}^{M\sqrt{\tau}}\varphi\left(\delta_1\right)\frac{\partial\sigma}{\partial M}=1.
\end{aligned}
$$

A.6.2 Greeks computation

Equations (2.10) and (2.11) are derived in this appendix. Because $\frac{\partial\delta_2}{\partial M}=\frac{\partial\delta_1}{\partial M}-\sqrt{\tau}\frac{\partial\sigma}{\partial M}$ and $\varphi\left(\delta_2\right)=\varphi\left(\delta_1\right)\mathrm{e}^{\sqrt{\tau}M}$,

$$
\begin{aligned}
\Delta =&\frac{\partial c}{\partial S}=\frac{\partial F}{\partial S}\frac{\partial c}{\partial F}\\
=&\mathrm{e}^{-r\tau}\frac{\partial F}{\partial S}\left(\left(\Phi\left(\delta_1\right)-\mathrm{e}^{-M\sqrt{\tau}}\Phi\left(\delta_2\right)\right)\right.\\
&\left.+F_{0,\tau}\left(\varphi\left(\delta_1\right)\frac{\partial M}{\partial F}\frac{\partial\delta_1}{\partial M}+\mathrm{e}^{-M\sqrt{\tau}}\sqrt{\tau}\frac{\partial M}{\partial F}\Phi\left(\delta_2\right)-\mathrm{e}^{-M\sqrt{\tau}}\varphi\left(\delta_2\right)\frac{\partial M}{\partial F}\frac{\partial\delta_2}{\partial M}\right)\right)\\
=&\mathrm{e}^{-r\tau}\frac{\partial F}{\partial S}\left(\left(\Phi\left(\delta_1\right)-\mathrm{e}^{-M\sqrt{\tau}}\Phi\left(\delta_2\right)\right)\right.\\
&\left.+F_{0,\tau}\frac{\partial M}{\partial F}\left(\varphi\left(\delta_1\right)\frac{\partial\delta_1}{\partial M}+\mathrm{e}^{-M\sqrt{\tau}}\Phi\left(\delta_2\right)\sqrt{\tau}-\varphi\left(\delta_1\right)\left(\frac{\partial\delta_1}{\partial M}-\sqrt{\tau}\frac{\partial\sigma}{\partial M}\right)\right)\right)\\
=&\mathrm{e}^{-q\tau}\left(\Phi\left(\delta_1\right)+\varphi\left(\delta_1\right)\frac{\partial\sigma}{\partial M}\right).
\end{aligned}
$$

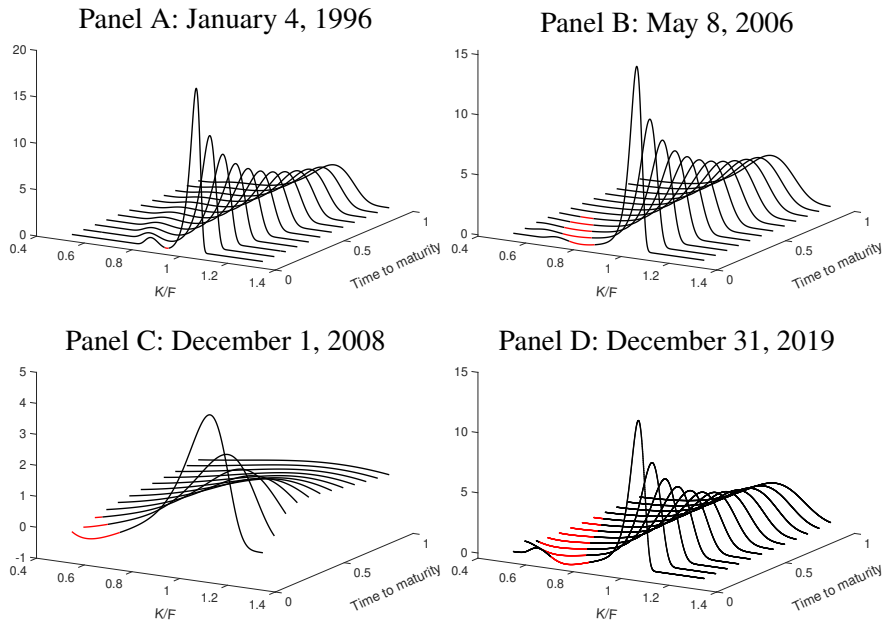
Recall that $\frac{\partial\varphi}{\partial z}\left(z\right)=-z\varphi\left(z\right)$. Therefore,

$$
\begin{aligned}
\Gamma =&\frac{\partial^2 c}{\partial S^2}=\frac{\partial F}{\partial S}\frac{\partial}{\partial F}\left(\frac{\partial c}{\partial S}\right)=\mathrm{e}^{-q\tau}\frac{\partial F}{\partial S}\frac{\partial}{\partial F}\left(\Phi\left(\delta_1\right)+\varphi\left(\delta_1\right)\frac{\partial\sigma}{\partial M}\right)\\
=&\mathrm{e}^{-q\tau}\frac{\partial F}{\partial S}\left(\varphi\left(\delta_1\right)\frac{\partial M}{\partial F}\frac{\partial\delta_1}{\partial M}-\delta_1\varphi\left(\delta_1\right)\frac{\partial M}{\partial F}\frac{\partial\delta_1}{\partial M}\frac{\partial\sigma}{\partial M}+\varphi\left(\delta_1\right)\frac{\partial M}{\partial F}\frac{\partial^2\sigma}{\partial M^2}\right)\\
=&\frac{\mathrm{e}^{-q\tau}\mathrm{e}^{(r-q)\tau}}{\sqrt{\tau}F_{0,\tau}}\varphi\left(\delta_1\right)\left(\frac{\partial\delta_1}{\partial M}-\delta_1\frac{\partial\delta_1}{\partial M}\frac{\partial\sigma}{\partial M}+\frac{\partial^2\sigma}{\partial M^2}\right)\\
=&\frac{\mathrm{e}^{-q\tau}}{\sqrt{\tau}S_0}\varphi\left(\delta_1\right)\left(\frac{\partial\delta_1}{\partial M}-\delta_1\frac{\partial\delta_1}{\partial M}\frac{\partial\sigma}{\partial M}+\frac{\partial^2\sigma}{\partial M^2}\right).
\end{aligned}
$$

A.7 Risk-neutral densities within CT and GG frameworks

This section illustrates the importance of the smoothness of the IV surface model as well as its asymptotic behaviour. For the same four days that were selected in Figure 2.1, Fig-

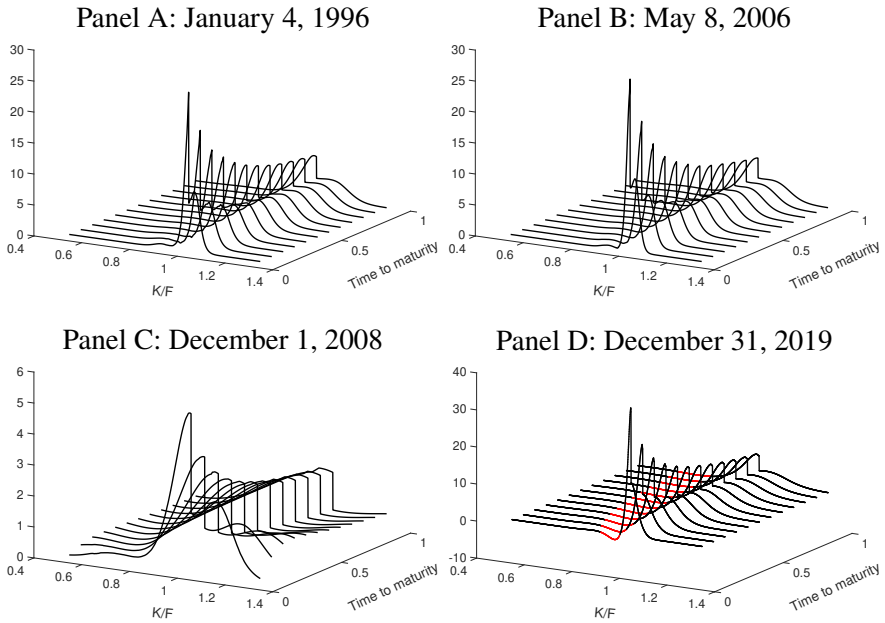
ures A.2 and A.3 present risk-neutral densities function obtained from the GG and CT benchmark models. We observe negative values and irregular tail behaviour.



Risk-neutral density functions $h_\tau(y) = \exp(y)g_\tau(\exp(y))$ of the log-prices derived from GG model call prices. January 4, 1996, is the first day in the sample. May 8, 2006, is a low volatility day. December 1, 2008, represents the peak of the 2008 financial crisis. December 31, 2019, is the last day of the sample. The red line highlights regions where the density is negative.

Figure A.2: Log-price risk-neutral densities implied by the GG model

As explained in Section 2.3, the Chalamandaris and Tsekrekos, 2011 moneyness measure consists of a linear transformation of the Black-Scholes Δ. To compute the risk-neutral density from the Chalamandaris and Tsekrekos, 2011 implied volatility surface, the moneyness measure is transformed back to the displayed moneyness measure K/F. To transform back the Chalamandaris and Tsekrekos, 2011 moneyness measure to K/F, the option implied volatility is required. A two-dimensional interpolation method using the observed options implied volatility with their respective characteristics is implemented to obtain implied volatilities with the desired characteristics (strike price and time-to-maturity). To extrapolate from the surface, the options with the largest (smallest) moneyness are considered.



Risk-neutral density functions $h_\tau(y) = \exp(y)g_\tau(\exp(y))$ of the log-prices derived from CT model call prices. January 4, 1996, is the first day in the sample. May 8, 2006, is a low volatility day. December 1, 2008, represents the peak of the 2008 financial crisis. December 31, 2019, is the last day of the sample. The red line highlights regions where the density is negative.

Figure A.3: Log-price risk-neutral densities implied by CT model

A.8 Abnormal IV surface on October 9, 2006

The observed implied volatilities from October 9, 2006, are provided in Figure A.4.

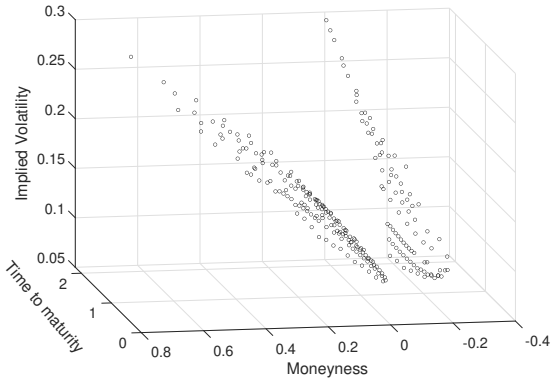


Figure A.4: IV surface on October 9, 2006

Chapter B

Appendices of *Joint dynamics for the underlying asset and its implied volatility surface: A new methodology for option risk management*

B.1 Model components' contribution to performance

The JIVR model, described in Section 3.4, encompasses multiple features such as (i) GARCH-type stochastic volatilities, (ii) non-Gaussian innovations following NIG distributions, (iii) the leverage effect captured in the NGARCH's asymmetric variance responses as in Duan, 1995, (iv) a two-component stochastic volatility process for the underlying asset log-returns and the long-term level factor β_1, and (v) a Gaussian copula to capture the dependence structure between the innovations.

To study the contribution of each model feature, we compare the performance of a sequence of nested sub-models with incremental complexity obtained by repeatedly adding a model feature to the previous element of the sequence.

There are six sub-models: (1) *BS*, the Black-Scholes model, (2) *Gaussian GARCH*, (3) *Gaussian NGARCH*, (4) *NIG NGARCH*, (5) the *Indep. JIVR* model where the Gaussian copula is set to the identity matrix and (6) the JIVR model. The specifications of the first four models are described in Table B.1. The fifth model (*Indep. JIVR*) is described by Equations (3.3)-(3.5), but the innovations across the underlying return and IV factor

components are assumed to be independent. Finally, the ***JIVR model*** from Section 3.4 includes the copula to capture the dependence between the various components.

	S&P 500 log-returns R_t		Volatility and surface factor coefficients β_i	
	$R_t = \left(r_{t,1/\Delta} - q_t + \lambda h_{t+1,R}\right)\Delta$ $-\psi(\sqrt{h_{t+1,R}\Delta}) + \sqrt{h_{t+1,R}\Delta}\epsilon_{t+1,R}$		$\beta_i = \alpha_i + \sum_{j=1}^{5}\theta_{i,j}\beta_{t,j} + \sqrt{h_{t+1,i}\Delta}\epsilon_{t+1,i}$	
BS	$h_{t+1,R} = h_R$	$\epsilon_{t,R} \sim \mathcal{N}(0,1)$	$h_{t+1,i} = h_i$	$\epsilon_{t,R} \sim \mathcal{N}(0,1)$
Gaussian GARCH	$h_{t+1,R} = \sigma_R^2 + \kappa_R\left(h_{t,R} - \sigma_R^2\right)$ $+a_R h_{t,R}\left(\epsilon_{t,R}^2 - 1\right)$	$\epsilon_{t,R} \sim \mathcal{N}(0,1)$	$h_{t+1,i} = \sigma_i^2 + \kappa_i\left(h_{t,i} - \sigma_i^2\right)$ $+a_i h_{t,i}\left(\epsilon_{t,i}^2 - 1\right)$	$\epsilon_{t,R} \sim \mathcal{N}(0,1)$
Gaussian NGARCH	$h_{t+1,R} = \sigma_R^2 + \kappa_R\left(h_{t,R} - \sigma_R^2\right)$ $+a_R h_{t,R}\left(\epsilon_{t,R}^2 - 1 - 2\gamma_R\epsilon_{t,R}\right)$	$\epsilon_{t,R} \sim \mathcal{N}(0,1)$	$h_{t+1,i} = \sigma_i^2 + \kappa_i\left(h_{t,i} - \sigma_i^2\right)$ $+a_i h_{t,i}\left(\epsilon_{t,i}^2 - 1 - 2\gamma_R\epsilon_{t,i}\right)$	$\epsilon_{t,R} \sim \mathcal{N}(0,1)$
NIG NGARCH	$h_{t+1,R} = \sigma_R^2 + \kappa_R\left(h_{t,R} - \sigma_R^2\right)$ $+a_R h_{t,R}\left(\epsilon_{t,R}^2 - 1 - 2\gamma_R\epsilon_{t,R}\right)$	$\epsilon_{t,R} \sim$ NIG$*$	$h_{t+1,i} = \sigma_i^2 + \kappa_i\left(h_{t,i} - \sigma_i^2\right)$ $+a_i h_{t,i}\left(\epsilon_{t,i}^2 - 1 - 2\gamma_R\epsilon_{t,i}\right)$	$\epsilon_{t,i} \sim$ NIG$*$

The first row displays the specification for the S&P 500 log-returns and the factor coefficients. The subsequent rows indicate the volatility process and the distribution of the innovations for their respective model. The NIG* refers to the standardized NIG distribution described in Appendix B.5.

Table B.1: Nested sub-models

Parameters from all sub-models are estimated through maximum likelihood, as described in Section 3.5. The log-likelihood performance metric is computed out-of-sample based on the expanding window method described in Section 3.6. The model performance is measured out-of-sample to ensure that overfitting does not unduly give an advantage to more complex sub-models. To compare the log-likelihoods across sub-models, the out-of-sample log-likelihood and the corresponding ALR metric defined in Equation (3.10) are considered.

Panel A of Table B.2 exhibits the out-of-sample log-likelihoods for each model. Results are presented for all constituents (i.e. the S&P 500 log-returns and the factor coefficients) individually, and the last row, *Joint*, presents the aggregated log-likelihood for the joint model. Panel B displays the ALR metrics, which are computed based on the log-likelihood of any model with respect to the previous one in the series of nested sub-models. The ALR is described in detail in Section 3.6.2. Again, results are presented for separate constituents

and in aggregate.

First, the results clearly highlight the importance of including a GARCH-type volatility for all five factor coefficients and for the S&P 500 log-returns. The ALRs are all large and far above 1, indicating that the superiority of the performance of the Gaussian GARCH model over that of the BS is unequivocal. The addition of an asymmetric volatility response through the replacement of GARCH with NGARCH processes results in improvements mainly for the S&P 500 log-returns and for the time-to-maturity slope. For other factors, such a modification does not lead to much (if any) improvement in performance. Integrating NIG-distributed innovations positively impacts the fitting performance for the S&P 500 log-returns and for all five factor coefficients. The second-to-last column (*Indep. JIVR*) displays the relative performance of the full specification described in Equations (3.3)-(3.5) over that of the NIG-NGARCH model. With an ALR of 1.03 for the S&P 500 log-returns and 1.02 for the long-term level factor, the results highlight a large improvement stemming from the inclusion of the two-component variance process. Lastly, the last column, which depicts the impact of including the Gaussian copula, exhibits an ALR of 2.29. This highlights the importance of taking the dependence structure into consideration.

	BS	Gaussian	Gaussian	NIG	JIVR Model	
	AR(1)	GARCH	NGARCH	NGARCH	Indep.	Copula
Panel A: Log-lik.						
S&P 500 log-returns	10,124	11,365	11,466	11,600	11,709	
Long-term level	13,788	15,617	15,620	15,785	15,861	
TmT Slope	7,115	8,681	8,843	9,304	9,319	
Moneyness Slope	15,392	15,674	15,672	15,709	15,709	
Smile attenuation	14,922	15,251	15,250	15,352	15,352	
Smirk	14,441	14,861	14,860	14,995	14,995	
Joint	75,785	81,452	81,713	82,748	82,947	85,873
Panel B: ALR						
S&P 500 log-returns		1.42	1.03	1.04	1.03	
Long-term level		1.68	1.00	1.05	1.02	
TmT Slope		1.56	1.05	1.14	1.01	
Moneyness Slope		1.08	1.00	1.01	1.00	
Smile attenuation		1.10	1.00	1.03	1.00	
Smirk		1.13	1.00	1.04	1.00	
Joint		4.99	1.08	1.34	1.06	2.29

The table exhibits log-likelihoods (Panel A) and the ALR (Panel B) for the aggregated out-of-sample period (2007-2020). The specification of the first four models (*BS*, *Gaussian GARCH*, *Gaussian NGARCH*, *NIG NGARCH*) are presented in Table B.1 and the specification of last two models (JIVR Indep. and JIVR copula) correspond to the specification exhibited in Section 3.4, i.e. Equation (3.3)-(3.5). The last row (Joint) displays the log-likelihood for the joint models assuming independence between the log-returns and the five factors, except for the last column (*JIVR model copula*) where the Gaussian copula captures the dependence. Panel B exhibits the ALR metrics, which use the log-likelihood of the corresponding row and column of Panel A as the numerator input, and the log-likelihood of the corresponding row but the previous column of Panel A as the denominator input.

Table B.2: Out-of-sample model performance

B.2 Cramér-von Mises test

The Cramér-von Mises p-value is computed through a two-step process. The first step involves generating a cumulative distribution function for the Cramér-von Mises statistic using a bootstrapping scheme.

More precisely, numerous sets of random variables following a NIG distribution with the parameters exhibited in Table 3.4 are simulated. Parameters of the NIG distribution are fitted over each simulated set. The Cramér-von Mises statistic is computed for all simulated sets using the fitted parameters. This method generates a large number of simulated Cramér-von Mises statistics, which forms the Cramér-von Mises test distribution. The second step consists in estimating the Cramér-von Mises statistic on the residuals. The p-value of the Cramér-von Mises statistic is obtained using the simulated Cramér-von Mises statistic's distribution, as computed in Step 1.

Algorithm 2 Cramér-von Mises test

for $n = 1$ to N **do**
  for $m = 1$ to M **do**
    draw $x_{m,n} \sim NIG(\zeta_{\text{NIG}},\, \phi_{\text{NIG}})$
  end for
  $[\tilde{\zeta}^{n*}_{\text{NIG}}, \tilde{\phi}^{n*}_{\text{NIG}}] = \arg\max_{\tilde{\zeta}^{n}_{\text{NIG}}, \tilde{\phi}^{n}_{\text{NIG}}} \sum_{j=1}^{M} \log\left(f_{NIG}(x_{j,n};\, \tilde{\zeta}^{n}_{\text{NIG}},\, \tilde{\phi}^{n}_{\text{NIG}}\right)$
  $CV(n) = \frac{1}{12M} + \sum_{j=1}^{M}\left(\frac{2j-1}{2M} - F_{NIG}(x_{j,n};\, \tilde{\zeta}^{n*}_{\text{NIG}},\, \tilde{\phi}^{n*}_{\text{NIG}})\right)^2$
end for
$CVS = \frac{1}{12M} + \sum_{j=1}^{M}\left(\frac{2j-1}{2M} - F_{NIG}(r_j;\, \zeta_{\text{NIG}},\, \phi_{\text{NIG}})\right)^2$
p-value $= \frac{\left(\sum_{j=1}^{N} \mathbb{1}_{\{CV(j)>CVS\}}\right)}{N}$;

In Algorithm 2, F_{NIG} refers to the cumulative distribution function of the NIG distribution, f_{NIG} is the density distribution function of the NIG distribution, and r is the vector of residuals.

B.3 VaR coverage tests

In this section, the methodology for the VaR coverage test is presented for the specific case where $\alpha < 0.5$. The principle to compute the test for $\alpha > 0.5$ is identical, *mutatis mutandis*. The VaR coverage test verifies if the proportion of ex-post VaR breaches are close to the model's α percentile level. The notation VaR_t^α represents the VaR estimate for a specific day t and confidence level α.

The test is performed following the backtesting procedure called the *hit sequence*, which is described in detail in Kupiec et al., 1995. The *hit sequence* of VaR breaches is defined as follows:

$$I_{t+d} = \begin{cases} 1 & \frac{V_{t+d}-V_t}{V_t} < \text{VaR}_{t+d}^\alpha, \\ 0 & \text{otherwise} \end{cases}$$

where V_t is the value of the strategy at time t. If the VaR calculation methodology is well specified, the frequency of VaR breaches (i.e. elements I_{t+d} equal to 1) should be close to α.

More precisely, the *hit sequence* should be composed of independent and identically distributed Bernoulli random variables. The null hypothesis of the test is thus $H_0 : \sum_{t=d+1}^{N} I_t \sim \text{Binomial}(\alpha, N-d)$. The test is performed as follows:

$$\begin{aligned} m_1 &= \sum_{t=d+1}^{N} I_t, \quad m_0 = N - d - \sum_{t=d+1}^{N} I_t, \\ L_1 &= m_0 \log(1-\alpha) + m_1 \log \alpha \\ L_2 &= m_0 \log\left(1 - \frac{m_1}{N-1}\right) + m_1 \log\left(\frac{m_1}{N-1}\right). \end{aligned}$$

According to the likelihood ratio test, $-2(L_1 - L_2) \sim \chi_1^2$ distribution where χ_1^2 is a Chi-squared distribution with one degree of freedom. The p-value of the test is computed as $1 - CDF_{\chi_1^2}(-2(L_1 - L_2))$.

B.4 Diebold & Mariano (1995) test

In Section 3.6.2, we use the Diebold and Mariano, 1995 test to compare the predictive performance of the JIVR model with that of the direct model from a statistical standpoint. In our particular case, the considered predictive performance measure is the log-likelihood. Let $\mathcal{L}_t^{(1)}$ and $\mathcal{L}_t^{(2)}$ be log-likelihoods at time t of the JIVR model and the direct model, respectively.

We refer to the time-series of $d_t = \mathcal{L}_t^{(1)} - \mathcal{L}_t^{(2)}$ as the loss differential. The sample autocovariance γ_k at lag k is defined as

$$\gamma_k = \frac{1}{N}\sum_{t=k+1}^{N}(d_t - \bar{d})(d_{t-k} - \bar{d}), \qquad \bar{d} = \frac{1}{N}\sum_{t=1}^{N} d_t.$$

The Diebold and Mariano, 1995 statistic is computed as $DM = \frac{\bar{d}}{\sqrt{\left(\gamma_0 + 2\sum_{k=1}^{h-1}\gamma_k\right)/N}}$ where $h = \lfloor N^{1/3} + 1 \rfloor$ and $DM \sim \mathcal{N}(0,1)$. The p-value of the test is computed as $1 - CDF_{\mathcal{N}}(DM)$.

B.5 Standardized Normal Inverse Gaussian probability density function

The probability density function of the standardized NIG distribution is defined as:

$$f(x) = \frac{B_1\left(\sqrt{\frac{\phi^6}{\phi^2+\zeta^2} + (\phi^2+\zeta^2)\left(x + \frac{\phi^2\zeta}{\phi^2+\zeta^2}\right)^2}\right)}{\pi\sqrt{\frac{1}{\phi^2+\zeta^2} + \frac{\phi^2+\zeta^2}{\phi^6}\left(x + \frac{\phi^2\zeta}{\phi^2+\zeta^2}\right)^2}} e^{\left(\frac{\phi^4}{\phi^2+\zeta^2} + \zeta\left(x + \frac{\phi^2\zeta}{\phi^2+\zeta^2}\right)\right)}, \quad x \in \mathbb{R},$$

where B_1 denotes the modified Bessel function of the second kind with index 1, which is described in Barndorff-Nielsen et al., 2001.[1]

[1]This density is obtained by replacing β^{NIG} and γ^{NIG} with ζ and ϕ, respectively, in the common $(\alpha^{NIG}, \beta^{NIG}, \delta^{NIG}, \mu^{NIG})$-specification of the NIG density and by imposing a null mean and unit variance to express δ^{NIG}, μ^{NIG} in terms of $\alpha^{NIG}, \beta^{NIG}$ (or alternatively $\gamma^{NIG} = \sqrt{(\alpha^{NIG})^2 - (\beta^{NIG})^2}$).

Chapter C

Appendices of *Foreseeing the worst: Forecasting electricity DART spikes*

C.1 Weather forecast simulation and interpolation

C.1.1 Temperature forecast interpolation

The temperature forecast data consist of daily weather forecasts as of 18:00 with horizons ranging from 30 to 54 hours in three-hour increments, i.e., three-hour forecast periods corresponding respectively to 00:00, 3:00, 6:00, ..., 21:00). Hourly spike forecasts performed in Section 4.3 require hourly temperature forecasts as input. To handle the missing temperature forecast for hours falling between the three-hour increments, a simple linear interpolation scheme is implemented. More precisely, interpolated temperature forecasts $\widehat{\text{TF}}$ are computed as follows:

$$\widehat{\text{TF}}_{3t+i} = \frac{(3-i)\text{TF}_{3t} + i\text{TF}_{3(t+1)}}{3}, \quad \text{where } i = 1, 2,$$

where TF is the observed temperature forecast.

C.1.2 Synthetic temperature forecasts before October 2017

The historical temperature forecast dataset obtained is only available as of October 7, 2017. To fill for missing observations in the dataset that starts in 2015, simulated forecasts are

generated by injecting noise in realized value. Temperature forecasts are described as a combination of the realized temperatures and error terms:

$$\mathrm{TF}_t = \mathrm{RT}_t + \epsilon_t,$$

where TF is the temperature forecast, RT is the realized temperature, and ϵ is the forecast error. To simulate temperature forecasts before October 7, 2017, a noise component ϵ_t is added to the realized temperature RT_t. Because the forecast error might be influenced by a multitude of seasonal factors, the noise component is sampled using a simple bootstrapping approach to circumvent this potential complexity.

The first step consists in storing the available forecast errors ϵ_t from October 7, 2017, to December 31, 2021. The second step consists in simulating the temperature forecasts from January 1, 2015, to October 6, 2017. For each hour of the sample, this is achieved by randomly sampling a forecast error among all post-October 6, 2017, observations sharing the same hour and date. For instance, to sample an error for hour 6:00 of May 8, 2017, we would randomly pick the forecast error among these from 6:00 on May 8 of either 2018, 2019, 2020, or 2021.

C.2 Predictive models

C.2.1 Logistic regression

Due to a large number of categorical features, the dummy variables are aggregated into buckets when applying the logistic regression. The categorical features month and hour encompass, respectively, 12 and 24 categories. Due to the similarity between multiple categories and to reduce the dimensionality of the feature vector, the categories are regrouped into bins: [January, February], [March, April, May], [June, July, August, September], [October, November, December] and [23:00 to 5:00], [6:00 to 10:00], [11:00 to 13:00], [14:00 to 16:00], [17:00 to 19:00], [20:00 to 22:00].

Another transformation is applied to the *load/grid* feature. The relation between the target variable (spikes) and the *load/grid* feature is non-linear. To capture the non-linearity

with the logistic regression model, a feature corresponding to the squared value of *load/grid* is added to the feature set.

C.2.2 Model estimation and hyperparameter tuning

The random forests, gradient-boosting trees, and DNNs all encompass a set of hyperparameters. The choice of hyperparameters is largely related to the model's performance. Unfortunately, hyperparameters cannot be optimized with the regular model parameters during the training step.

Several search methods, such as random search, grid search, or Bayesian optimization, can be used to search over many sets of hyperparameters. Furthermore, the search process must be paired with a performance evaluation technique to discriminate between the sets of hyperparameters tested. The current study implements a grid search algorithm paired with a k-fold cross-validation process. The grid search method takes as input a grid of hyperparameters. The algorithm trains the model and computes the performance for each possible combination in the grid. The k-fold cross-validation is a method that makes it possible to compute each set of hyperparameters' performance objectively. The k-fold cross-validation starts by splitting the dataset into k folds. The algorithm then lists the k possible combinations of the $k-1$ folds while keeping the remaining fold for a performance review. The model is then trained for each of the k combinations, and the performance of the model is assessed over the remaining fold. The overall model performance is computed by aggregating the testing fold results for all k combinations. The hyperparameter set with the greatest performance is selected. It is important to note that the size of the grids and the number of folds impact the numerical load. In this study, five-fold cross-validation is applied.

C.3 Model confidence set approach

In Section 4.3 and 4.4, the model confidence set approach of Hansen et al., 2011 is implemented to identify the best performing model(s) using the log-likelihood and the trading P&L as the performance measures. More precisely, the Hansen et al., 2011 method selects

the best-performing model(s) through an iterative process based on a two-step approach, with steps respectively being called the equivalence test and the elimination step. The equivalence test verifies whether all models in the model confidence set generate performances that are statistically indistinguishable. If such null hypothesis from the equivalence test is rejected, the elimination step identifies which model is to be removed from the confidence set.

Since the number of considered models is low compared to the number of out-of-sample observations, Hansen et al., 2011 indicates that the statistic of the equivalence test can be computed as follows. The forecast performance at time t for each model k in the model confidence set is represented by $L_{t,k}$. Such quantities are collected in the matrix $L = [L_{t,k}]$ where t refers to the row and k to the column. L_t refers to row t of L. The total number of considered models is K. In the current study, X_t is set to $X_t = L_t M$, where the M is a $K \times K-1$ matrix such that $M_{i,j} = 1$ when $i = j$, $M_{i,j} = -1$ when $i+1 = j$ and $M_{i,j} = 0$ otherwise. Under the null hypothesis, X_t has a zero mean.

The test statistic is

$$T = n\bar{X}'\hat{\Sigma}^{-1}\bar{X},$$

where $\bar{X}$ is the arithmetic average of $X_1, \ldots, X_n$, $\hat{\Sigma}$ is a consistent estimator of the covariance matrix of $\bar{X}$, and n is the number of observations. Under the null hypothesis, $T \rightarrow \chi^2(q)$, where $q = \text{rank}(\Sigma)$. In this study, the confidence level considered is 5%. If the equivalence test is rejected, the model with the largest standardized excess loss is removed from the model confidence set. The standardized excess loss for model j is computed as follows:

$$t_k = \frac{\bar{p}_k}{\sqrt{\widehat{\text{var}}(\bar{p}_k)}},$$

where $p_{t,k} = L_{t,k} - \frac{1}{K}\sum_{i=1}^{K} L_{t,i}$ and $\bar{p}_k = \frac{1}{n}\sum_{t=1}^{n} p_{t,k}$. The process iterates through steps 1 and 2 until either (i) the equivalence test is not rejected or (ii) only one model remains inside the model confidence set.

C.4 Assessment of the CTHI feature

A predictor used by the NYISO to forecast the load is the cumulative temperature and humidity index (CTHI). As explained in NYISO, 2021, the CTHI index is a weather metric that integrates information about temperature and humidity in the last three days and accounts for the "heat buildup within building structures during a heatwave." More precisely, the temperature and humidity index (THI) is a measure of bodily discomfort experienced during warm weather. As described in NYISO, 2021, it is constructed by combining the temperature and relative humidity:

$$\text{THI}_t = 0.6\, T_t + 0.4\, \text{WB}_t$$

where T_t represents the hour-t temperature and WB_t is the corresponding wet-bulb temperature.[1] The THI_d^*, for day d, is defined as the maximum observed hourly THI over the last 24 hours, i.e., the 24-hour period ending at the prediction time 18:00. The day-d CTHI, denoted CTHI_d, is a weighted average of the THI_d^* over the past three days:

$$\text{CTHI}_d = 0.7\, \text{THI}_d^* + 0.2\, \text{THI}_{d-1}^* + 0.1\, \text{THI}_{d-2}^*.$$

Table C.1 reports the performance metrics (AUC and average log-likelihood) across models and thresholds when the CTHI feature is added to the feature set described in Section 3.2.

Comparing the performance metrics reported in Table C.1 with those from Table 4.3 outlines the general slight underperformance across models and thresholds when the CTHI feature is included in the feature set.

[1] The hour-t wet-bulb temperature WB_t is defined as the temperature read from a thermometer submerged in water. Since this information is not commonly found in historical weather datasets, an estimate of the WB temperature is considered. As explained in Stull, 2011, the wet-bulb temperature is estimated by combining the dry-bulb temperature and the relative humidity as follows:

$$\begin{aligned}\text{WB}_t \;=\;& T_t \arctan\left(0.151977\sqrt{\text{RH}_t + 8.313659}\right) + \arctan(T_t + \text{RH}_t) \\ & - \arctan(\text{RH}_t - 1.676331) + (0.00391838\,(\text{RH}_t)^{1.5}) \arctan(2.3101\,\text{RH}_t) - 4.686035\end{aligned}$$

where RH_t is the hour-t relative humidity in %. The dry-bulb temperature is the temperature directly drawn from the dataset.

		Logistic Regression			Random Forest			Gradient Boosting Trees			DNN		
	γ^-	-30	-45	-60	-30	-45	-60	-30	-45	-60	-30	-45	-60
A	AUC												
In-sample	2015-2017	0.720	0.737	0.747	0.760	0.784	0.809	0.778	0.805	0.815	0.733	0.753	0.766
	2015-2018	0.707	0.730	0.745	0.747	0.771	0.792	0.751	0.774	0.805	0.735	0.713	0.730
	2015-2019	0.715	0.743	0.758	0.754	0.777	0.799	0.756	0.788	0.807	0.700	0.728	0.746
	2015-2020	0.716	0.743	0.759	0.752	0.778	0.803	0.805	0.783	0.806	0.719	0.746	0.756
Out-of-sample	2018	0.661	0.700	0.724	0.675	0.707	0.728	0.684	0.709	0.735	0.647	0.688	0.710
	2019	0.717	0.773	0.794	0.750	0.784	0.816	0.734	0.780	0.798	0.707	0.773	0.800
	2020	0.733	0.756	0.780	0.735	0.748	0.760	0.734	0.753	0.771	0.710	0.728	0.753
	2021	0.700	0.730	0.751	0.706	0.735	0.734	0.703	0.736	0.749	0.678	0.705	0.719
	Agg.	0.707	0.743	0.763	0.719	0.749	0.763	0.717	0.749	0.767	0.691	0.727	0.744
B	Average log-likelihood												
In-sample	2015-2017	-0.209	-0.154	-0.121	-0.201	-0.146	-0.113	-0.197	-0.143	-0.113	-0.207	-0.152	-0.119
	2015-2018	-0.213	-0.154	-0.120	-0.205	-0.148	-0.114	-0.205	-0.148	-0.113	-0.207	-0.156	-0.122
	2015-2019	-0.202	-0.145	-0.112	-0.196	-0.139	-0.106	-0.195	-0.138	-0.106	-0.204	-0.146	-0.113
	2015-2020	-0.196	-0.140	-0.107	-0.191	-0.136	-0.102	-0.180	-0.135	-0.102	-0.196	-0.141	-0.108
Out-of-sample	2018	-0.225	-0.157	-0.120	-0.223	-0.157	-0.121	-0.222	-0.156	-0.120	-0.229	-0.158	-0.122
	2019	-0.162	-0.109	-0.078	-0.164	-0.112	-0.080	-0.161	-0.109	-0.079	-0.164	-0.109	-0.078
	2020	-0.171	-0.122	-0.086	-0.169	-0.122	-0.086	-0.169	-0.121	-0.084	-0.174	-0.124	-0.088
	2021	-0.281	-0.197	-0.141	-0.279	-0.196	-0.144	-0.279	-0.195	-0.141	-0.286	-0.203	-0.146
	Agg.	-0.206	-0.143	-0.105	-0.205	-0.143	-0.105	**-0.203**	**-0.142**	**-0.104**	-0.208	-0.146	-0.106

The four models are the logistic regression, the random forest, gradient boosting trees, and the deep neural network (DNN). Panel A's performance metric is the area under the curve (AUC), while Panel B is the average log-likelihood. The models generate out-of-sample predictions for 2018 to 2021. The models are trained on the previous years' observations for each out-of-sample forecast. For example, to generate out-of-sample forecasts for 2019, the models are trained on the observations from 2015 to 2018. The last row of each panel displays the performance metric computed over the aggregated (Agg.) out-of-sample years.

Table C.1: In-sample and out-of-sample performance metrics when CTHI is added to the feature set

C.5 Stacked Classifier

A common approach in machine learning is to consider ensemble models, which combine the predictions of multiple models. In this study, four statistical learning algorithms are trained. As shown in Figure 4.8, the predicted probabilities of each model are not perfectly interchangeable and display differences from one another. Table C.2 exhibits the performance of the combined models.[2] Results show that combining the predictions from the various models does not improve the performance of the out-of-sample predictions.

		AUC			Average log-likelihood		
	γ^-	-30	-45	-60	-30	-45	-60
In-sample	2015-2017	0.760	0.771	0.796	-0.198	-0.144	-0.110
	2015-2018	0.756	0.768	0.791	-0.202	-0.147	-0.115
	2015-2019	0.751	0.782	0.801	-0.194	-0.137	-0.106
	2015-2020	0.752	0.793	0.798	-0.189	-0.133	-0.102
Out-of-sample	2018	0.659	0.673	0.650	-0.231	-0.166	-0.133
	2019	0.751	0.776	0.811	-0.160	-0.111	-0.081
	2020	0.727	0.692	0.688	-0.182	-0.135	-0.111
	2021	0.705	0.740	0.727	-0.292	-0.210	-0.181
	Aggregated	0.715	0.726	0.718	-0.213	-0.153	-0.124

The four models embedded in the ensemble model are: logistic regression, random forest, gradient boosting trees, and a deep neural network (DNN). The models are combined by considering each model's output as the input of a logistic regression. The logistic regression is then trained over an in-sample set to predict spikes, and the performance is computed over the out-of-sample set. Panel A's performance metric is the area under the curve (AUC), while Panel B is the average log-likelihood. The logistic regression generates out-of-sample predictions from 2018 to 2021. The models are trained on the previous years' observations for each out-of-sample forecast. For example, to generate out-of-sample forecasts for 2019, the models are trained on the observations from 2015 to 2018.

Table C.2: Ensemble model predictions

[2]The models are combined by considering each model's output as the input of a logistic regression model. The logistic regression is then trained over an in-sample set to predict spikes, and the performance is computed over the out-of-sample set.

C.6 Revised feature set based on Section 4.3.3 results

In Section 4.3.3, the feature importance assessment reveals that three features, namely *weekend/holidays*, *past day-ahead load forecast* and *past day-ahead price forecast*, contribute little to the predictions, not to mention sometimes being harmful in terms of predictive performance. To reduce overfitting issues and improve the out-of-sample model performance, a common practice consists in removing these features from the feature set. However, since some of the results presented in Section 4.3.3 are based on the out-of-sample dataset, removing these features in Section 4.4 would be regarded as integrating future information into the models and could potentially artificially improve the results. Nevertheless, it is still interesting to test whether the revised feature set would improve the results. Thus, Table C.3 reports the performance metrics over the in-sample and out-of-sample sets with the revised feature set excluding those three features. Results indicate that with such revised feature set, the aggregated performance is generally qualitatively similar (albeit often slightly higher) for the various models and thresholds.

		Logistic Regression			Random Forest			Gradient Boosting Trees			DNN		
	γ^-	-30	-45	-60	-30	-45	-60	-30	-45	-60	-30	-45	-60
A	AUC												
In-sample	2015-2017	0.711	0.729	0.741	0.747	0.776	0.799	0.754	0.771	0.780	0.720	0.738	0.749
	2015-2018	0.702	0.726	0.741	0.736	0.764	0.784	0.775	0.808	0.826	0.713	0.729	0.755
	2015-2019	0.709	0.737	0.753	0.741	0.772	0.793	0.773	0.773	0.790	0.716	0.745	0.764
	2015-2020	0.712	0.739	0.755	0.742	0.773	0.796	0.772	0.805	0.826	0.730	0.746	0.759
Out-of-sample	2018	0.669	0.709	0.729	0.682	0.719	0.734	0.688	0.712	0.725	0.666	0.707	0.732
	2019	0.714	0.768	0.791	0.751	0.784	0.822	0.744	0.786	0.821	0.746	0.772	0.793
	2020	0.743	0.764	0.789	0.744	0.755	0.765	0.742	0.767	0.774	0.723	0.733	0.741
	2021	0.705	0.734	0.756	0.706	0.732	0.745	0.699	0.734	0.761	0.668	0.719	0.741
	Agg.	0.711	0.747	0.766	0.724	0.752	0.769	0.691	0.752	0.771	0.697	0.736	0.757
B	Log-likelihood												
In-sample	2015-2017	-0.210	-0.154	-0.121	-0.204	-0.148	-0.115	-0.203	-0.149	-0.117	-0.209	-0.154	-0.121
	2015-2018	-0.213	-0.154	-0.121	-0.208	-0.149	-0.116	-0.200	-0.143	-0.111	-0.212	-0.154	-0.119
	2015-2019	-0.203	-0.145	-0.112	-0.198	-0.141	-0.108	-0.207	-0.140	-0.108	-0.202	-0.144	-0.111
	2015-2020	-0.197	-0.141	-0.107	-0.193	-0.137	-0.103	-0.186	-0.132	-0.100	-0.194	-0.140	-0.107
Out-of-sample	2018	-0.224	-0.155	-0.120	-0.222	-0.155	-0.121	-0.221	-0.155	-0.120	-0.226	-0.156	-0.120
	2019	-0.163	-0.109	-0.079	-0.162	-0.111	-0.079	-0.162	-0.110	-0.079	-0.160	-0.110	-0.079
	2020	-0.170	-0.121	-0.085	-0.168	-0.121	-0.086	-0.188	-0.118	-0.084	-0.172	-0.123	-0.087
	2021	-0.278	-0.195	-0.140	-0.278	-0.196	-0.142	-0.279	-0.195	-0.140	-0.290	-0.199	-0.143
	Agg.	-0.206	-0.143	-0.104	-0.204	-0.143	-0.105	-0.209	-0.142	-0.104	-0.208	-0.145	-0.105

The four models–logistic regression, random forest, gradient boosting trees, and deep neural network (DNN)–are trained over a revised feature set. The revised feature set is composed of the *heating degree days* (HDD), *cooling degree days* CDD, *hour* indicators, *month* indicators, and *past spikes*. Conversely, *weekend/holidays*, *past day-ahead load forecast*, and *past day-ahead price forecast* features are excluded from the feature set. Panel A's performance metric is the area under the curve (AUC), while Panel B is the log-likelihood divided by the number of observations. The models generate out-of-sample predictions for 2018 to 2021. The models are trained on the previous years' observations for each out-of-sample forecast. For example, to generate out-of-sample forecasts for 2019, the models are trained on the observations from 2015 to 2018. The last row of each panel displays the performance metric computed over the aggregated (Agg.) out-of-sample years.

Table C.3: Performance with the revised feature set